Policy Certificates: Towards Accountable Reinforcement Learning

Christoph Dann
Carnegie Mellon University

Lihong Li
Google Brain

Wei Wei
Google Cloud AI Research

Emma Brunskill
Stanford University

Abstract

The performance of a reinforcement learning algorithm can vary drastically during learning because of exploration. Existing algorithms provide little information about their current policy's quality before executing it, and thus have limited use in high-stakes applications like healthcare. In this paper, we address such a lack of *accountability* by proposing that algorithms output *policy certificates*, which upper bound the sub-optimality in the next episode, allowing humans to intervene when the certified quality is not satisfactory. We further present a new learning framework (IPOC) for finite-sample analysis with policy certificates, and develop two IPOC algorithms that enjoy guarantees for the quality of both their policies and certificates.

1 Introduction

There is increasing excitement around applications of machine learning (ML), but also growing awareness and concern. Recent research on FAT (fairness, accountability and transparency) ML aims to address these concerns but most work focuses on supervised learning settings and only few works exist on reinforcement learning or sequential decision making in general (Jabbari et al., 2016; Joseph et al., 2016; Kannan et al., 2017; Raghavan et al., 2018).

One challenge when applying reinforcement learning (RL) in practice is that, unlike in supervised learning, the performance of an RL algorithm is typically not monotonically increasing with more data due to the trial-and-error nature of RL that necessitates exploration. Even sharp drops in policy performance during learning are possible, for example, when the agent starts to explore a new part of the state space. Such unpredictable performance fluctuation has limited the use of RL in high-stakes applications like healthcare, and calls for more accountable algorithms that can quantify and reveal their performance during learning.

In this work, we propose that an RL algorithm outputs *policy certificates*, a form of confidence interval, in episodic reinforcement learning. Policy certificates are upper bounds on how far from optimal the return (expected sum of rewards) of an algorithm in the next episode can be. They allow one to monitor the policy's performance and intervene if necessary, thus improving accountability of the algorithm. Formally, we propose a theoretical framework called IPOC that not only guarantees that certificates are valid performance bounds but also that both, the algorithm's policy and certificates, improve with more data.

There are two relevant lines of research on RL with guaranteed performance for episodic reinforcement learning. The first area is on frameworks for guaranteeing the performance of a RL algorithm across many episodes, as it learns. Such frameworks, like regret (Jaksch et al., 2010), PAC (probably approximately correct, Kakade, 2003; Strehl et al., 2009) and Uniform-PAC (Dann et al., 2017) all provide apriori bounds about the cumulative performance of the algorithm, such as bounding the total number of times an algorithm may execute a policy that is not near optimal. However, these frameworks do not provide bounds for any individual episode. In contrast, the second main related area for providing guarantees focuses on estimating and guaranteeing the performance of a particular RL policy, given some prior data (e.g., Thomas et al., 2015b; Jiang and Li, 2016; Thomas and Brunskill, 2016). Such work typically provides limited or no guarantees for algorithms that are learning and updating their policies across episodes. In this paper, we unite both lines of work by providing performance guarantees *online* for a reinforcement learning algorithm in *individual episodes* and across all episodes. In fact, we show that bounds in our new IPOC framework imply strong guarantees in existing regret and PAC frameworks.

We consider policy certificates in two settings, finite episodic Markov decision processes (MDPs) and, more generally, finite MDPs with episodic side information (context) (Abbasi-Yadkori and Neu, 2014; Hallak et al., 2015; Modi et al., 2018). The latter is of particular interest in practice. For example, in a drug treatment optimization task where each patient is one episode, context is the background information of the patient which influences the treatment outcome. While one expects the algorithm to learn a good policy quickly for frequent contexts, the performance for unusual patients may be significantly more variable due to the limited prior experience of the algorithm. Policy certificates allow humans to detect when the current policy is good for the current patient and intervene if a certified performance is deemed inadequate. For example, for this health monitoring application, a human expert could intervene to either directly specify the policy for that episode, or in the context of automated customer service, the service could be provided at reduced cost to the customer.

Existing algorithms based on the optimism-in-the-face-of-uncertainty (OFU) principle (e.g., Auer et al., 2009) are natural to extend to learning with policy certificates. We demonstrate this by extending the `UBEV` algorithm (Dann et al., 2017) for episodic MDPs with finite state and action spaces, and show that with high probability it outputs certificates greater than ϵ at most $\tilde{O}(S^2AH^3/\epsilon^2)$ times for all ϵ. For problems with side information, we propose an algorithm that learns with policy certificates in episodic MDPs with adversarial linear side information (Abbasi-Yadkori and Neu, 2014; Modi et al., 2018) of dimension d, and bound the rate at which the cumulative sum of certificates can grow up to log terms by $\tilde{O}\left(H^2d\sqrt{S^3AT}\right)$.

2 Setting and Notation

In this work, we consider episodic RL problems where the agent interacts with the environment in episodes of a certain length. While the framework for policy certificates applies to a wide range of problems, we focus on finite Markov decision processes (MDP) with linear side information (Modi et al., 2018; Hallak et al., 2015; Abbasi-Yadkori and Neu, 2014) for concreteness. This setting includes tabular MDPs as a special case but is more general and can model variations in the environment across episodes, e.g., because different episodes correspond to treating different patients in a healthcare application. Unlike the tabular special case, function approximation is necessary for efficient learning.

Finite MDPs with linear side information. The agent interacts in episode k by observing a state $s_{k,t} \in \mathcal{S}$, taking action $a_{k,t} \in \mathcal{A}$ and observing the next state $s_{k,t+1}$ as well as a scalar reward $r_{k,t} \in [0,1]$. This interaction loop continues for H time steps $t \in [H] := \{1, 2, \dots H\}$, before a new episode starts. We assume that state- and action-space are of finite sizes S and A, respectively, as in the widely considered tabular MDPs (Osband and Van Roy, 2014; Dann and Brunskill, 2015; Azar et al., 2017; Jin et al., 2018). But here, the agent essentially interacts with a family of infinitely many tabular MDPs that is parameterized by linear contexts. At the beginning of episode k, two contexts, $x_k^{(r)} \in \mathbb{R}^{d^{(r)}}$ and $x_k^{(p)} \in \mathbb{R}^{d^{(p)}}$, are observed and the agent interacts in this episode with a tabular MDP, whose dynamics and reward function depend on the contexts in a linear fashion. Specifically, it is assumed that the rewards are sampled from $P_R(s,a)$ with means $r_k(s,a) = (x_k^{(r)})^\top \theta_{s,a}^{(r)}$ and transition probabilities are $P_k(s'|s,a) = (x_k^{(p)})^\top \theta_{s',s,a}^{(p)}$ where $\theta_{s,a}^{(r)} \in \mathbb{R}^{d^{(r)}}$ and $\theta_{s',s,a}^{(p)} \in \mathbb{R}^{d^{(p)}}$ are unknown parameter vectors for each $s, s' \in \mathcal{S}, a \in \mathcal{A}$. As a regularity condition, we assume bounded parameters, i.e., $\|\theta_{s,a}^{(r)}\|_2 \leq \xi_{\theta^{(r)}}$ and $\|\theta_{s',s,a}^{(p)}\|_2 \leq \xi_{\theta^{(p)}}$ as well as bounded contexts $\|x_k^{(r)}\|_2 \leq \xi_{x^{(r)}}$ and $\|x_k^{(p)}\|_2 \leq \xi_{x^{(p)}}$. We allow $x_k^{(r)}$ and $x_k^{(p)}$ to be different, and use x_k to denote $(x_k^{(r)}, x_k^{(p)})$ in the following. To further simplify notation, we assume w.l.o.g. that there is a fixed start state. Note that there is no assumption of the distribution of contexts; our framework and algorithms can handle adversarially chosen contexts.

Return and optimality gap. The quality of a policy π in any episode k is evaluated by the *total expected reward* or *return*: $\rho_k(\pi) := \mathbb{E}\left[\sum_{i=t}^H r_{k,h} \middle| a_{k,1:H} \sim \pi\right]$, where this notation means that all actions in the episode are taken as prescribed by π. We focus here on deterministic time-dependent policies $\pi : \mathcal{S} \times [H] \to \mathcal{A}$ and note that optimal policy and return $\rho_k^\star = \max_\pi \rho_k(\pi)$ depend on the context of the episode. The difference of achieved and optimal return is called *optimality gap* $\Delta_k = \rho_k^\star - \rho_k(\pi_k)$ for each episode k where π_k is the algorithm's policy in that episode.

Additional notation. We denote by $\Delta_{\max} = H$ the largest optimality gap possible and $Q_{k,h}^\pi(s,a) = \mathbb{E}[\sum_{t=h}^H r_{k,t} | a_{k,h} = a, a_{k,h+1:H} \sim \pi]$ and $V_{k,h}^\pi(s) = Q_{k,h}^\pi(s, \pi(s,h))$ are the Q- and value function of π in episode k. Optimal versions are marked by superscript $\star$ and subscripts k, h are omitted when unambiguous. We often treat $P(s,a)$ as linear operator, that is, $P(s,a)f = \sum_{s' \in \mathcal{S}} P(s'|s,a)f(s')$ for any $f : \mathcal{S} \to \mathbb{R}$.

3 Existing Learning Frameworks

During execution, the optimality gaps Δ_k are hidden, the algorithm only observes the sum of rewards which is a sample of $\rho_k(\pi_k)$. This causes risk as one does not know when the algorithm is playing a good policy and when a potentially bad policy. One might hope that performance guarantees for algorithms mitigate this risk but no existing theoretical framework gives guarantees for individual episodes during learning:

- *Mistake-style PAC bounds* (Strehl et al., 2006, 2009; Szita and Szepesvári, 2010; Lattimore and Hutter, 2012; Dann and Brunskill, 2015) bound the number of ϵ-mistakes, that is, the size of the superlevel set $\{k \in \mathbb{N} : \Delta_k > \epsilon\}$ with high probability, but do not tell us when mistakes can happen. The same is true for the recently proposed stronger Uniform-PAC bounds (Dann et al., 2017) which hold for all $\epsilon > 0$ jointly.
- *Supervised-learning style PAC bounds* (Kearns and Singh, 2002; Jiang et al., 2017; Dann et al., 2018) guarantee that the algorithm outputs an ϵ-optimal policy for a given ϵ, that is, they ensure that $\Delta_k \leq \epsilon$ for k greater than the bound. They do however require to know ϵ ahead of time and do not give any guarantee about Δ_k during learning (when k is smaller than the bound).
- *Regret bounds* (Osband et al., 2013, 2016; Azar et al., 2017; Jin et al., 2018) control the cumulative sum of optimality gaps $\sum_{k=1}^{T} \Delta_k$ (regret) which does not yield any nontrivial guarantee for individual Δ_k because it does not tell which optimality gaps are small.

Not knowing Δ_k during learning makes it difficult to stop an algorithm at some point and extract a good policy. For example, the common way to extract a good policy for algorithms with regret bound $R(T)$ is to pick one of the T policies executed so far at random (Jin et al., 2018). This only yields a good policy that has $\Delta_k \leq R(T)/T$ with probability $1/T$ in general. As a result, one requires $\Omega(1/\delta)$ episodes for a good policy with probability at least $1-\delta$ which is much larger than the $\tilde{O}(\ln 1/\delta)$ of algorithms with supervised-learning style PAC bounds. Note that the KWIK framework (Li et al., 2008) does guarantee the quality of individual predictions but is for supervised learning; its use in RL leads to mistake-style PAC bounds (see Section 7).

4 The IPOC Framework

We introduce a new learning framework that mitigates the limitations of prior guarantees highlighted above. This framework forces the algorithm to output its current policy π_k as well as a certificate $\epsilon_k \in \mathbb{R}_+$ before each episode k. This certificate ϵ_k informs the user how sub-optimal the policy can be for the current context, i.e., $\epsilon_k \geq \Delta_k$ and allows one to intervene if needed. For example, in automated customer services, one might reduce the service price in episode k if certificate ϵ_k is above a certain threshold, since the quality of the provided service cannot be guaranteed. When there is no context, a certificate upper bounds the suboptimality of the current policy in any episode which makes algorithms anytime interruptable (Zilberstein and Russell, 1996): one is guaranteed to always know a policy with improving performance. Our learning framework is formalized as follows:

Definition 1 (Individual Policy Certificates (IPOC) Bounds)**.** *An algorithm satisfies an individual policy certificate (IPOC) bound F if for a given $\delta \in (0, 1)$ it outputs a* certificate ϵ_k *and the current policy π_k before each episode k (after observing the contexts) so that with probability at least $1 - \delta$*

1. *all certificates are upper bounds on the sub-optimality of policy π_k played in episode k, i.e., $\forall\ k \in \mathbb{N} : \quad \epsilon_k \geq \rho^\star_k - \rho_k(\pi_k)$; and either*

2a. *for all number of episodes T the cumulative sum of certificates is bounded $\sum_{k=1}^{T} \epsilon_k \leq F(W, T, \delta)$ (Cumulative Version), or*

2b. *for any threshold ϵ, the number of times certificates can exceed the threshold is bounded as $\sum_{k=1}^{\infty} \mathbf{1}\{\epsilon_k > \epsilon\} \leq F(W, \epsilon, \delta)$ (Mistake Version).*

Here, W can be (known or unknown) properties of the environment. If conditions 1 and 2a hold, we say the algorithm has a cumulative IPOC bound and if conditions 1 and 2b hold, we say the algorithm has a mistake IPOC bound.

Condition 1 alone would be trivial to satisfy with $\epsilon_k = \Delta_{\max}$, but condition 2 prohibits this by controlling the size of ϵ_k. Condition 2a bounds the cumulative sum of certificates (similar to regret bounds), and condition 2b bounds the size of the superlevel sets of ϵ_k (similar to PAC bounds). We allow both alternatives as condition 2b is stronger but one sometimes can only prove condition 2a (see Sec. 5.2.1). An IPOC bound controls simultaneously the quality of certificates (how big $\epsilon_k - \rho^\star_k + \rho_k(\pi_k)$ is) as well as the optimality gaps $\rho^\star_k - \rho(\pi_k)$ themselves and hence an IPOC bound not only guarantees that the algorithm improves its policy but also becomes better at telling us how well the policy performs. As such it is stronger than existing frameworks. Besides this benefit, IPOC ensures that the algorithm is anytime interruptable, i.e., it can be used to find better and better policies that have small Δ_k with high probability $1 - \delta$. That means IPOC bounds imply supervised learning style PAC bounds for all ϵ jointly. These claims are formalized in the following statements:

Proposition 2. *Assume an algorithm has a cumulative IPOC bound $F(W, T, \delta)$.*

1. *Then it has a regret bound of same order, i.e., with probability at least $1-\delta$, for all T the regret $R(T) := \sum_{k=1}^{T} \Delta_k$ is bounded by $F(W, T, \delta)$.*

2. *If F has the form $\sum_{p=0}^{N} (C_p(W, \delta)T)^{\frac{p}{p+1}}$ for appropriate functions C_p, then with probability at least $1-\delta$ for any ϵ, it outputs a certificate $\epsilon_k \le \epsilon$ within*
$$\sum_{p=0}^{N} \frac{C_p(W,\delta)^p (N+1)^{p+1}}{\epsilon^{p+1}} \tag{1}$$
episodes. For settings without context, this implies that the algorithm outputs an ϵ-optimal policy within that number of episodes (supervised learning-style PAC bound).

Proposition 3. *If an algorithm has a mistake IPOC bound $F(W, \epsilon, \delta)$, then*

1. *it has a uniform PAC bound $F(W, \epsilon, \delta)$, i.e., with probability at least $1-\delta$, the number of episodes with $\Delta_k \ge \epsilon$ is at most $F(W, \epsilon, \delta)$ for all $\epsilon > 0$;*

2. *with probability at least $1-\delta$ for all ϵ, it outputs a certificate $\epsilon_k \le \epsilon$ within $F(W, \epsilon, \delta) + 1$ episodes. For settings without context, that means the algorithm outputs an ϵ-optimal policy within that many episodes (supervised learning-style PAC).*

3. *if F has the form $\sum_{p=1}^{N} \frac{C_p(W,\delta)}{\epsilon^p} \left(\ln \frac{\tilde{C}(W,\delta)}{\epsilon}\right)^{np}$ with $C_p(W, \delta) \ge 1$ it also has a cumulative IPOC bound of order*
$$\tilde{O}\left(\sum_{p=1}^{N} C_p(W,\delta)^{1/p} T^{\frac{p-1}{p}} \operatorname{polylog}(\Delta_{\max}, \tilde{C}(W,\delta), T)\right).$$

Note that the functional form in part 2 of Prop. 2 includes all common polynomial bounds like $O(\sqrt{T})$ or $O(T^{2/3})$ with appropriate factors and similarly for part 3 of Prop. 3 which covers for example $\tilde{O}(1/\epsilon^2)$.

Algorithm 1: `ORLC` (**O**ptimistic **R**einforcement **L**earning with **C**ertificates)

Input: failure tolerance $\delta \in (0, 1]$

$n(s,a) = l(s,a) = m(s', s, a) = 0 \quad \forall s, s' \in \mathcal{S}, a \in \mathcal{A}$;
$\tilde{V}_{H+1}(s) \leftarrow 0; \quad \underset{\sim}{V}_{H+1}(s) \leftarrow 0 \quad \forall s \in \mathcal{S}$
$V_h^{\max} \leftarrow (H - h + 1) \qquad \forall h \in [H]$;
$\phi(n) := 1 \wedge \sqrt{\frac{2 \ln\ln(\max\{e,n\}) + \ln(18SAH/\delta)}{n}}$;
for $k = 1, 2, 3, \ldots$ **do**
  `// Optimistic planning and policy evaluation`
  for $h = H$ **to** 1 ***and*** $s \in \mathcal{S}$ **do**
    for $a \in \mathcal{A}$ **do**
      $\tilde{\psi}_h(s,a) \leftarrow V_h^{\max} \phi(n(s,a))$;
      $\underset{\sim}{\psi}_h(s,a) \leftarrow 2V_h^{\max}\sqrt{S}\phi(n(s,a))$;
      $\hat{r}(s,a) \leftarrow \frac{l(s,a)}{n(s,a)}; \quad \hat{P}(s,a) \leftarrow \frac{m(\cdot,s,a)}{n(s,a)}$;
      $\tilde{Q}_h(s,a) \leftarrow \hat{r}(s,a) + \hat{P}(s,a)\tilde{V}_{h+1} + \tilde{\psi}_h(s,a)$;
      $\underset{\sim}{Q}_h(s,a) \leftarrow \hat{r}(s,a) + \hat{P}(s,a)\underset{\sim}{V}_{h+1} - \underset{\sim}{\psi}_h(s,a)$;
      `// clip values`
      $\tilde{Q}_h(s,a) \leftarrow 0 \vee \tilde{Q}_h(s,a) \wedge V_h^{\max}$;
      $\underset{\sim}{Q}_h(s,a) \leftarrow 0 \vee \underset{\sim}{Q}_h(s,a) \wedge V_h^{\max}$;
    $\pi_k(s,t) \leftarrow \operatorname{argmax}_a \tilde{Q}_h(s,a)$;
    $\tilde{V}_h(s) \leftarrow \tilde{Q}(s, \pi_k(s,t)); \quad \underset{\sim}{V}_h(s) \leftarrow \underset{\sim}{Q}(s, \pi_k(s,t))$;
  `/* Execute policy for one episode */`
  $s_{k,1} \sim P_0$;
  $\epsilon_k \leftarrow \tilde{V}_1(s_{k,1}) - \underset{\sim}{V}_1(s_{k,1})$;
  output policy π_k **with certificate** ϵ_k;
  for $h = 1$ **to** H **do**
    $a_{k,h} \leftarrow \pi_k(s_{k,h}, h)$;
    $r_{k,h} \sim P_R(s_{k,h}, a_{k,h}); \quad s_{k,h+1} \sim P(s_{k,h}, a_{k,h})$;
    `// Update statistics`
    $n(s_{k,h}, a_{k,h}) \leftarrow n(s_{k,h}, a_{k,h}) + 1$;
    $m(s_{k,h+1}, s_{k,h}, a_{k,h}) \leftarrow m(s_{k,h+1}, s_{k,h}, a_{k,h}) + 1$;
    $l(s_{k,h}, a_{k,h}) \leftarrow l(s_{k,h}, a_{k,h}) + r_{k,h}$;

5 Algorithms with Policy Certificates

As shown above, IPOC is stricter than other learning frameworks. Existing algorithms based on the OFU principle (Auer et al., 2009) need extensions to satisfy IPOC bounds. OFU algorithms can be interpreted as maintaining a set of models defined by confidence sets of the individual components and picking the policy optimistically from that set of models. As a byproduct, this yields an upper confidence bound on the optimal value function and therefore optimal return $\rho_k^\star$. We augment this by computing a lower confidence bound on value function of the optimistic policy $V_k^{\pi_k}$ recursively using the same confidence set of models. This yields a lower confidence bound on $\rho_k(\pi_k)$ which is sufficient to compute ϵ_k. We demonstrate this approach by extending two similar OFU algorithms, one for tabular MDPs with no side information, and the other for the more general case with side information. While the algorithms have similar structure we consider them separately because we can prove stronger IPOC guarantees for the first (see Section 5.2.1).

5.1 Policy Certificates in Tabular MDPs

We present an extension of the `UBEV` algorithm by Dann et al. (2017) called `ORLC` (optimistic reinforcement learning with certificates) and shown in Algorithm 1. Algorithm 1 essentially combines the policy selection approach of `UBEV` with high-confidence model-based policy evaluation of the current policy. Before each episode, Algorithm 1 computes $\tilde{Q}$, an optimistic estimate of $Q^\star$, as well as $\underset{\sim}{Q}$, a pessimistic estimate of Q^{π_k}, by dynamic programming on the empirical model $(\hat{P}, \hat{r})$ and confidence intervals $\tilde{\psi}$ and $\underset{\sim}{\psi}$ for $\tilde{Q}$ and $\underset{\sim}{Q}$, respectively. Note that the width of the lower confidence bounds $\underset{\sim}{\psi}$ is by a factor $2\sqrt{S}$ larger than $\tilde{\psi}$, as the estimation target Q^{π_k} of $\underset{\sim}{Q}$ is a random quantity due to the dependency on π_k as opposed to $Q^\star$ (see discussion below).

We show the following IPOC bound for this algorithm:

Theorem 4 (Mistake IPOC Bound of Algorithm 1)**.** *For any given $\delta \in (0, 1)$, Algorithm 1 satisfies in any tabular MDP with S states, A actions and horizon H, the following Mistake IPOC bound: For all $\epsilon > 0$, the number of episodes where the algorithm outputs a certificate $\epsilon_k > \epsilon$ is*

$$\tilde{O}\left(\frac{S^2 A H^3}{\epsilon^2} \ln \frac{1}{\delta}\right). \tag{2}$$

By Proposition 3, this implies a PAC bound of same order as well as a $\tilde{O}(S\sqrt{AH^3 T \ln 1/\delta})$ regret bound. The PAC lower bound (Dann and Brunskill, 2015) for this setting is $\Omega\left(\frac{SAH^2}{\epsilon^2} \ln \frac{1}{\delta + c}\right)$, which implies an IPOC mistake lower bound of the same order by Proposition 3. We conjecture that Algorithm 1 satisfies a Uniform-PAC bound of order $O\left(\left(\frac{SAH^3}{\epsilon^2} + \frac{S^2 A H^3}{\epsilon}\right) \ln \frac{1}{\delta}\right)$, which is by a factor H lower than the Uniform-PAC bound of `UBEV` due to our assumption of time-independent dynamics. Using techniques by Azar et al. (2017), this bound can be reduced to match the lower PAC bound. However, as we sketch below, existing techniques cannot be directly applied to the lower confidence bounds in Algorithm 1. It is therefore an open question whether our IPOC bound in Theorem 4 is improvable, or whether the IPOC lower bound is strictly larger than the PAC lower bound. Interestingly, in the related active learning setting, such a discrepancy between achieved and certifiable performance is known to exist (Balcan et al., 2010).

5.1.1 Proof Sketch of the IPOC Bound

To show Theorem 4, we need to verify condition 1 and 2b of Definition 1. Condition 2b can be shown in similar way to existing Uniform-PAC (Dann et al., 2017) bounds but with optimality gaps Δ_k being replaced by certificates ϵ_k. For condition 1 it is sufficient to show $\tilde{Q}_{k,h} \geq Q_h^\star$ and $\underset{\sim}{Q}_{k,h} \leq Q_h^{\pi_k}$ holds in all episodes k. We use additional subscripts k to indicate the value of variables before sampling in episode k. Proving optimism, $\tilde{Q}_{k,h} \geq Q_h^\star$, is standard in analyses of OFU algorithms. Hence, we focus here on showing $\underset{\sim}{Q}_{k,h} \leq Q_h^{\pi_k}$. When there is no value clipping one can use the following common decomposition (Azar et al., 2017; Jin et al., 2018; Dann et al., 2017). All terms are functions of s, a which we omit in the following for readability.

$$\begin{aligned} Q_h^{\pi_k} - \underset{\sim}{Q}_{k,h} &= r - \hat{r}_k + \underset{\sim}{\psi}_{k,h} + PV_{h+1}^{\pi_k} - \hat{P}_k \underset{\sim}{V}_{h+1} && (3)\\ &\geq r - \hat{r}_k + \phi(n_k) && (4)\\ &\quad + 2V_{h+1}^{\max}\sqrt{S}\phi(n_k) + PV_{h+1}^{\pi_k} - \hat{P}_k \underset{\sim}{V}_{h+1}. && (5) \end{aligned}$$

Here, we expanded $\underset{\sim}{\psi}_{k,h}$ and used $2V_h^{\max}\sqrt{S} \geq 1 + 2V_{h+1}^{\max}\sqrt{S}$. We want to show that this value difference is non-negative. Using standard martingale concentration, one can show $r - \hat{r}_k + \phi(n_k) \geq 0$. It remains to show that $\hat{P}_k \underset{\sim}{V}_{h+1} - PV_{h+1}^{\pi_k} \leq 2V_{h+1}^{\max}\sqrt{S}\psi(n_k)$. We decompose the left-hand side as

$$\hat{P}_k \underset{\sim}{V}_{h+1} - PV_{h+1}^{\pi_k} = \hat{P}_k(\underset{\sim}{V}_{h+1} - V_{h+1}^{\pi_k}) + (\hat{P}_k - P)V_{h+1}^{\pi_k}.$$

Using an inductive assumption that $\underset{\sim}{V}_{h+1} \leq V_{h+1}^{\pi_k}$, the first term cannot be positive. Note that when showing optimism ($\tilde{Q}_{k,h} \geq Q_h^\star$) the second term is $(\hat{P}_k - P)V_{h+1}^\star$ which is a martingale that can be bounded directly by $V_{h+1}^{\max}\phi(n_k)$. Unfortunately, the second term here $(\hat{P}_k - P)V_{h+1}^{\pi_k}$ is not a martingale as $\hat{P}_k$ and V^{π_k} both depend on the samples. For that reason, we have to resort to Hölder's inequality to decompose

$$(\hat{P}_k - P)V_{h+1}^{\pi_k} \leq \|\hat{P}_k - P\|_1 V_{h+1}^{\max} \tag{6}$$

and apply concentration bounds on the ℓ_1 distance of empirical distributions to get the upper bound $2V_{h+1}^{\max}\sqrt{S}\phi(n_k)$. This is why the lower confidence bound width $\underset{\sim}{\psi}$ are by a factor $2\sqrt{S}$ larger than the upper confidence bound widths $\tilde{\psi}$. Eventually, this yields a $O(S^2)$ IPOC bound compared to the conjectured Uniform-PAC bound with $O(S)$ dependency in the $1/\epsilon^2$ term. Similarly, the difference in H-dependency of our IPOC and conjectured Uniform-PAC bounds origins from leveraging Bernstein's inequality for the upper confidence bound widths. That requires bounding how much larger the empirical variance estimate of $\tilde{Q}$-value of next state can be compared to using target $Q^\star$ values. While this is possible by exploiting that $\tilde{Q}$ is monotonically decreasing with k (Azar et al., 2017, Equation 5), this technique cannot be applied to the lower confidence widths as $\underset{\sim}{Q}$ is not monotone in k.

5.2 Policy Certificates in MDPs With Linear Side Information

After considering the tabular MDP setting, we now present an algorithm for the more general setting with side information, which for example allows us to take

Algorithm 2: `ORLC-SI` (**O**ptimistic **R**einforcement **L**earning with **C**ertificates and **S**ide **I**nformation)

Input: failure prob. $\delta \in (0, 1]$, regularizer $\lambda > 0$

$\forall s, s' \in \mathcal{S}, a \in \mathcal{A}, h \in [H]$:
  $N^{(p)}_{s,a} \leftarrow \lambda I_{d^{(p)} \times d^{(p)}}$; $\quad N^{(r)}_{s,a} \leftarrow \lambda I_{d^{(r)} \times d^{(r)}}$;
  $M^{(r)}_{s,a} \leftarrow \vec{0}_{d^{(r)}}$; $\quad M^{(p)}_{s',s,a} \leftarrow \vec{0}_{d^{(p)}}$;
  $\tilde{V}_{H+1} \leftarrow \vec{0}_S \quad \underset{\sim}{V}_{H+1} \leftarrow \vec{0}_S \quad V^{\max}_h \leftarrow (H - h + 1)$;
  $\xi_{\theta^{(r)}} \leftarrow \sqrt{d}$; $\xi_{\theta^{(p)}} \leftarrow \sqrt{d}$; $\quad \delta' \leftarrow \frac{\delta}{S(SA+A+H)}$;
  $\phi(N, x, \xi) := \left[\sqrt{\lambda}\xi + \sqrt{\frac{1}{2}\ln\frac{1}{\delta'} + \frac{1}{4}\ln\frac{\det N}{\det(\lambda I)}}\right] \|x\|_{N^{-1}}$;

for $k = 1, 2, 3, \ldots$ **do**
  Observe current contexts $x^{(r)}_k$ and $x^{(p)}_k$;
  `/* estimate model with least squares */`
  for $s, s' \in \mathcal{S}, a \in \mathcal{A}$ **do**
    $\hat{\theta}^{(r)}_{s,a} \leftarrow (N^{(r)}_{s,a})^{-1} M^{(r)}_{s,a}$;
    $\hat{r}(s, a) \leftarrow 0 \vee (x^{(r)}_k)^\top \hat{\theta}^{(r)}_{s,a} \wedge 1$;
    $\hat{\theta}^{(p)}_{s',s,a} \leftarrow (N^{(p)}_{s,a})^{-1} M^{(p)}_{s',s,a}$;
    $\hat{P}(s'|s, a) \leftarrow 0 \vee (x^{(p)}_k)^\top \hat{\theta}^{(p)}_{s',s,a} \wedge 1$;
  `/* optimistic planning */`
  for $h = H$ **to** 1 ***and*** $s \in \mathcal{S}$ **do**
    for $a \in \mathcal{A}$ **do**
      $\tilde{\psi}_h(s, a) \leftarrow \|\tilde{V}_{h+1}\|_1 \phi(N^{(p)}_{s,a}, x^{(p)}_k, \xi_{\theta^{(p)}}) + \phi(N^{(r)}_{s,a}, x^{(r)}_k, \xi_{\theta^{(r)}})$;
      $\underset{\sim}{\psi}_h(s, a) \leftarrow \|\underset{\sim}{V}_{h+1}\|_1 \phi(N^{(p)}_{s,a}, x^{(p)}_k, \xi_{\theta^{(p)}}) + \phi(N^{(r)}_{s,a}, x^{(r)}_k, \xi_{\theta^{(r)}})$;
      $\tilde{Q}_h(s, a) \leftarrow \hat{r}(s, a) + \hat{P}(s, a)\tilde{V}_{h+1} + \tilde{\psi}_h(s, a)$;
      $\underset{\sim}{Q}_h(s, a) \leftarrow \hat{r}(s, a) + \hat{P}(s, a)\underset{\sim}{V}_{h+1} - \underset{\sim}{\psi}_h(s, a)$;
      `// clip values`
      $\tilde{Q}_h(s, a) \leftarrow 0 \vee \tilde{Q}_h(s, a) \wedge V^{\max}_h$;
      $\underset{\sim}{Q}_h(s, a) \leftarrow 0 \vee \underset{\sim}{Q}_h(s, a) \wedge V^{\max}_h$;
    $\pi_k(s, h) \leftarrow \operatorname{argmax}_a \tilde{Q}_h(s, a)$;
    $\tilde{V}_h(s) \leftarrow \tilde{Q}_h(s, \pi_k(s, t))$;
    $\underset{\sim}{V}_h(s) \leftarrow \underset{\sim}{Q}_h(s, \pi_k(s, t))$;
  `/* Execute policy for one episode */`
  $s_{k,1} \sim P_0$;
  $\epsilon_k \leftarrow \tilde{V}_1(s_{k,1}) - \underset{\sim}{V}_1(s_{k,1})$;
  output policy π_k **with certificate** ϵ_k;
  for $h = 1$ **to** H **do**
    $a_{k,h} \leftarrow \pi_k(s_{k,h}, h)$;
    $r_{k,h} \sim P_{R,k}(s_{k,h}, a_{k,h})$; $s_{k,h+1} \sim P_k(s_{k,h}, a_{k,h})$;
    `// Update statistics`
    $N^{(p)}_{s_{k,h},a_{k,h}} \leftarrow N^{(p)}_{s_{k,h},a_{k,h}} + x^{(p)}_k (x^{(p)}_k)^\top$;
    $N^{(r)}_{s_{k,h},a_{k,h}} \leftarrow N^{(r)}_{s_{k,h},a_{k,h}} + x^{(r)}_k (x^{(r)}_k)^\top$;
    $M^{(p)}_{s_{k,h+1},s_{k,h},a_{k,h}} \leftarrow M^{(p)}_{s_{k,h+1},s_{k,h},a_{k,h}} + x^{(p)}_k$;
    $M^{(r)}_{s_{k,h},a_{k,h}} \leftarrow M^{(r)}_{s_{k,h},a_{k,h}} + r_{k,h} x^{(p)}_k$;

background information about a customer into account and generalize across different customers.

Algorithm 2 gives an extension, called `ORLC-SI`, of the OFU algorithm by Abbasi-Yadkori and Neu (2014). Similar to tabular Algorithm 1, Algorithm 2 computes $\tilde{Q}$ as an upper bound on the optimal Q-function $Q^\star_k$ and $\underset{\sim}{Q}$ as a lower bound on the Q-function $Q^{\pi_k}_k$ of the current policy using dynamic programming with the empirical model $(\hat{P}, \hat{r})$ as well as confidence bound widths $\tilde{\psi}$ and $\underset{\sim}{\psi}$. Unlike in the tabular case, the empirical model is now computed as least-squares estimates of the model parameters evaluated at the current contexts. Specifically, the empirical transition probability $\hat{P}(s'|s, a)$ is $(x^{(p)}_k)^\top \hat{\theta}_{s',s,a}$ where $\hat{\theta}_{s',s,a}$ is the least squares estimate of model parameter $\theta_{s',s,a}$. Since transition probabilities are normalized, this estimate is then clipped to $[0, 1]$. Note that this empirical model is estimated separately for each (s', s, a)-triple, but does generalize across different contexts. The confidence widths $\tilde{\psi}$ and $\underset{\sim}{\psi}$ are derived using ellipsoid confidence intervals on model parameters (Abbasi-Yadkori and Neu, 2014). We show the following IPOC bound:

Theorem 5 (Cumulative IPOC Bound for Alg. 2)**.** *For any $\delta \in (0, 1)$ and regularization parameter $\lambda > 0$, Algorithm 2 satisfies the following cumulative IPOC bound in any MDP with S states, A actions, contexts with dimensions $d^{(r)}$ and $d^{(p)}$ as well as bounded parameters $\xi_{\theta^{(r)}} \leq \sqrt{d^{(r)}}$ and $\xi_{\theta^{(p)}} \leq \sqrt{d^{(p)}}$. With probability at least $1 - \delta$ all certificates are upper bounds on the optimality gaps and their total sum after T episodes is bounded for all T by*

$$\tilde{O}\left(\sqrt{S^3AH^4T}\lambda(d^{(p)} + d^{(r)}) \log \frac{\xi^2_{x^{(p)}} + \xi^2_{x^{(r)}}}{\lambda\delta}\right). \tag{7}$$

By Proposition 2, this IPOC bound implies a regret bound of the same order which improves on the $\tilde{O}(\sqrt{d^2S^4AH^5T \log 1/\delta})$ regret bound of Abbasi-Yadkori and Neu (2014) with $d = d^{(p)} + d^{(r)}$ by a factor of $\sqrt{SAH}$. While they make a different modelling assumption (generalized linear instead of linear), we believe at least our better S dependency is due to using improved least-squares estimators for the transition dynamics [1] and can likely be transferred to their setting. The mistake-type PAC bound by Modi et al. (2018) is not directly comparable because our cumulative IPOC does not imply a mistake-type PAC bound.[2] Nonetheless, loosely translating our result to a PAC-like bound yields $\tilde{O}\left(\frac{d^2S^3AH^5}{\epsilon^2} \ln \frac{1}{\delta}\right)$ which is

[1] They estimate $\theta_{s',s,a}$ only from samples where the transition $s, a \to s'$ was observed instead of all occurrences of s, a (no matter whether s' was the next state).

[2] Similar to regret and PAC bounds (Dann et al., 2017), an algorithm with a sublinear cumulative IPOC bound can still output a certificate larger than a given threshold $\epsilon_k \geq \epsilon$ infinitely often as long as it does so sufficiently less frequently (see Section 5.2.1).

much smaller than their $\tilde{O}\left(\frac{d^2 SAH^4}{\epsilon^5}\max\{d^2, S^2\}\ln\frac{1}{\delta}\right)$ bound for sufficiently small ϵ.

The confidence bounds in Algorithm 2 are more general but looser compared to the confidence bounds specialized to the tabular case in Algorithm 1, in particular the upper confidence bounds. Instantiating the cumulative IPOC bound for Algorithm 2 from Theorem 5 for tabular MDPs (where $x_k^{(r)} = x_k^{(p)} = 1 \in \mathbb{R}$ for all k) yields $\tilde{O}(\sqrt{S^3AH^4T}\ln 1/\delta)$ which is worse than the $\tilde{O}(\sqrt{S^2AH^3T}\ln 1/\delta)$ cumulative IPOC bound for Algorithm 1 implied by Theorem 4.

5.2.1 Mistake IPOC Bound for Algorithm 2?

By Proposition 3, a mistake IPOC bound is stronger than the cumulative version we proved for Algorithm 2. One might wonder whether Algorithm 2 also satisfies this stronger bound, but this is not the case:

Proposition 6. *For any $\epsilon < 1$, there is an MDP with linear side information such that Algorithm 2 outputs certificates $\epsilon_k \geq \epsilon$ infinitely often with probability* 1.

Proof Sketch. Consider a two-armed bandit where the two-dimensional context is identical to the deterministic reward for both actions. The context alternates between $x_A := \begin{bmatrix}(1+\epsilon)/2\\(1-\epsilon)/2\end{bmatrix}$ and $x_B := \begin{bmatrix}(1-\epsilon)/2\\(1+\epsilon)/2\end{bmatrix}$. That means in odd-numbered episodes, the agent receives reward $\frac{1+\epsilon}{2}$ for action 1 and reward $\frac{1-\epsilon}{2}$ for action 2 (bandit A) and conversely in even-numbered episodes (bandit B). Let $n_{A,i}$ and $n_{B,i}$ be the current number of times action i was played in each bandit and $N_i = \operatorname{diag}(n_{A,i}+\lambda, n_{B,i}+\lambda)$ the covariance matrix. One can show that the optimistic Q-value of action 2 in bandit A is lower bounded as

$$\tilde{Q}(2) \geq \sqrt{\ln\det N_2}\|x_A\|_{N_2^{-1}} \wedge 1 \tag{8}$$

$$= \sqrt{\frac{\ln(\lambda + n_{A,2}) + \ln(\lambda + n_{B,2})}{n_{A,2}}} \wedge 1. \tag{9}$$

Assume now the agent stops playing action 2 in bandit A and playing action 1 in bandit B at some point. Then the denominator in Eq (9) stays constant but the numerator grows unboundedly as $n_{B,2} \to \infty$. That implies that $\tilde{Q}(2) \to 1$ but the optimistic Q-value for the other action $\tilde{Q}(1) \to \frac{1+\epsilon}{2} \leq 1$ approaches the true reward. Eventually $\tilde{Q}(2) > \tilde{Q}(1)$ and the agent will play the ϵ-suboptimal action 2 in bandit A again. Hence, Algorithm 2 has to output infinitely many $\epsilon_k \geq \epsilon$. $\square$

The construction in the proof illustrates that the non-decreasing nature of the ellipsoid confidence intervals cause this negative result (due to the $\ln\det(N)$ term in $\phi(N, x, \xi)$ in Line 6 of Alg 2). This does not rule out alternative algorithms with mistake IPOC bound for this setting, but they would likely require entirely different parameter estimators and confidence bounds.

6 Simulation Experiments

Certificates need to upper bound the optimality gap in each episode, even for the worst case up to a small failure probability, and Algorithms 1 and 2 are not optimized for empirical performance. As such, their certificates may be conservative, and potentially significantly overestimate the unobserved optimality gaps without further empirical tuning. Yet, one may wonder whether the certificates output by Algorithms 1 and 2 are simply a monotonically decreasing sequence, or whether they can indicate the actual performance variation during learning. In this section, we present the results of a small simulation study, which demonstrates that the certificates do inform us about when the algorithms execute a bad policy. For brevity, we focus on the more general Algorithm 2 in tasks with side information. Details are available in Appendix D.

We first apply Algorithm 2 to randomly generated contextual bandit problems ($S = H = 1$) with $d^{(r)} = 10$ dimensional context and 40 actions.[3] Certificates and optimality gaps have a correlation of 0.88 which confirms that certificates are informative about the policy's return. If one for example needs to intervene when the policy is more than 0.2 from optimal (e.g., by reducing the price for that customer), then in more than 42% of the cases where the certificate is above 0.2, the policy is worse than 0.2 suboptimal.

In practice, the distribution of contexts can change rapidly. For example, in a call center dialogue system, there can be a sudden increase of customers calling due to a certain regional outage. Such abrupt shifts in contexts are prevalent, and can cause a drop in performance. We demonstrate that certificates can identify such performance drops. We consider a simulated MDP with 10 states, 40 actions and horizon 5 where rewards depend on a 10-dimensional context and let the distribution of contexts change after 2M episodes. As seen in Figure 1 (left), this causes a spike in optimality gap as well as certificates. Our algorithm reliably detects this sudden decrease of performance. In fact, the scatter plot in Figure 1 (right) shows that certificates are highly correlated with optimality gaps.

We would like to emphasize that our focus in this paper is to provide a theoretical framework and not opti-

[3] We actually use a slightly more complicated version of Algorithm 2 with better empirical performance but the same theoretical guarantees. All details are in Appendix D. For clarity, we presented the simpler Algorithm 2 in the main paper.

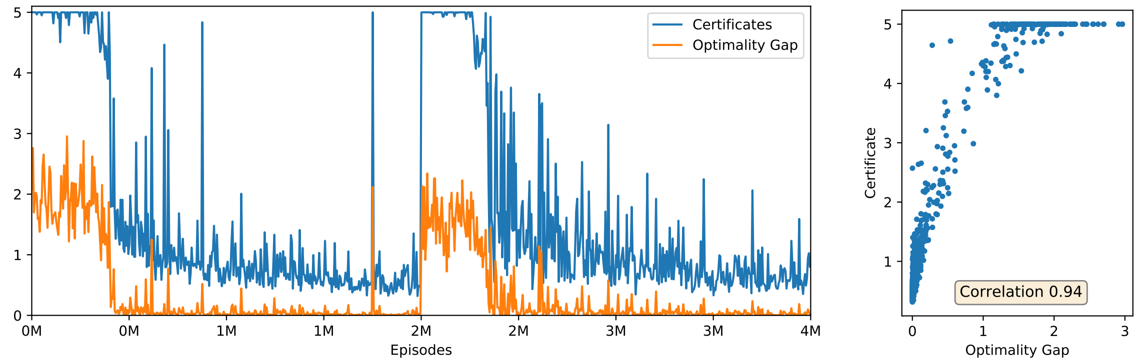

Figure 1: Results of Algorithm 2 for 4M episodes on an MDP with context distribution shift after 2M; Left: certificates and true (unobserved) optimality gap in temporal order (episodes sub-sampled for better visualization); Right: Scatter plot of all certificates and optimality gaps.

mize empirical performance. Nonetheless, these experiments indicates that even with no empirical tuning, certificates can be a useful indicator for the optimality gap of our algorithms. We expect that the empirical quality can be significantly improved in future work.

7 Related Work

The connection of IPOC to existing RL frameworks such as PAC and regret is shown formally in Section 4. In addition, IPOC is similar to the KWIK framework (Li et al., 2008), in that the algorithm is required to declare how well it will perform. However, KWIK is a framework for supervised learning algorithms, which are then used as building blocks to create PAC RL algorithms. In contrast, IPOC is a framework specifically for RL methods. Also, KWIK only requires to declare whether the output will perform better than a single threshold that is pre-specified in the input.

Our algorithms essentially compute confidence regions as in OFU algorithms, and then use that information in model-based policy evaluation to obtain policy certificates. There is a large body of works on off-policy policy evaluation (e.g., Jiang and Li, 2016; Thomas and Brunskill, 2016; Mahmood et al., 2017) including a few that provide non-asymptotic confidence intervals (e.g., Thomas et al., 2015b,a; Sajed et al., 2018). However, these methods focus on the batch setting where a batch of episodes sampled from known policies is given. Many approaches rely on importance weights that require stochastic data-collecting policies but most sample-efficient algorithms deploy deterministic policies. One could treat previous episodes to be collected by one stochastic data-dependent policy but that introduces bias in the importance-weighting estimators that is not accounted for in the analyses. In contrast to these batch offline approaches, our work focuses on providing guarantees for online RL in the form of policy certificates as well as proposing a theoretical framework for controlling the quality of these certificates and the learning speed of the algorithm.

There are also approaches on safe exploration (Kakade and Langford, 2002; Pirotta et al., 2013; Thomas et al., 2015a) that guarantee monotonically increasing policy performance by operating in a batch loop. Our work is orthogonal, as we do not aim to change exploration but rather expose its impact on performance to the users and give them the choice to intervene.

8 Conclusion and Future Work

We have introduced policy certificates to improve accountability in reinforcement learning by enabling users to intervene if the guaranteed performance is deemed inadequate. Bounds in our new theoretical framework IPOC ensure that certificates indeed upper bound the suboptimality in each episode and prescribe the rate at which certificates and policy improve. By extending two optimism-based algorithms, we have not only demonstrated RL with policy certificates but also our IPOC guarantees. This initial work on more accountable RL through online certificates opens up several exciting avenues for future work:

The high correlation of policy certificates and optimality gaps in our experiments motivates a more empirical study of RL with policy certificates, including the design of practical algorithms with well-calibrated policy certificates in more challenging settings (e.g. deep RL).

Policy certificates enable intervention and our theory already captures interventions that do not hinder execution of the algorithm (e.g. reducing the price for the current customer). As future work, it would be interesting to quantify how other interventions that provide alternative feedback (like expert demonstrations) affect learning speed.

Appropriate interventions may depend on *why* the algorithm chooses a potentially bad policy. The algorithm could for example explicitly explore as opposed to just not being able to exploit. To further improve accountability and interpretability, we could distinguish these cases by comparing certificates of the optimism-based policy and the policy that is optimal in the empirical model.

References

Yasin Abbasi-Yadkori and Gergely Neu. Online learning in mdps with side information. *arXiv preprint arXiv:1406.6812*, 2014.

Peter Auer, Thomas Jaksch, and Ronald Ortner. Near-optimal regret bounds for reinforcement learning. In *Advances in Neural Information Processing Systems*, 2009.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, pages 263–272, 2017.

Maria-Florina Balcan, Steve Hanneke, and Jennifer Wortman Vaughan. The true sample complexity of active learning. *Machine learning*, 80(2-3): 111–139, 2010.

Christoph Dann and Emma Brunskill. Sample complexity of episodic fixed-horizon reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 2818–2826, 2015.

Christoph Dann, Tor Lattimore, and Emma Brunskill. Unifying pac and regret: Uniform pac bounds for episodic reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 5713–5723, 2017.

Christoph Dann, Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E Schapire. On oracle-efficient pac reinforcement learning with rich observations. *arXiv preprint arXiv:1803.00606*, 2018.

Assaf Hallak, Dotan Di Castro, and Shie Mannor. Contextual Markov decision processes. *arXiv:1502.02259*, 2015.

Steven R. Howard, Aaditya Ramdas, Jon Mc Auliffe, and Jasjeet Sekhon. Uniform, nonparametric, non-asymptotic confidence sequences. *arXiv preprint arXiv:1810.08240*, 2018.

Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, and Aaron Roth. Fair learning in markovian environments. *arXiv preprint arXiv:1611.03071*, 2016.

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(Apr): 1563–1600, 2010.

Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning-Volume 48*, pages 652–661. JMLR. org, 2016.

Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E Schapire. Contextual decision processes with low bellman rank are pac-learnable. In *International Conference on Machine Learning*, pages 1704–1713, 2017.

Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is q-learning provably efficient? *arXiv preprint arXiv:1807.03765*, 2018.

Matthew Joseph, Michael Kearns, Jamie H Morgenstern, and Aaron Roth. Fairness in learning: Classic and contextual bandits. In *Advances in Neural Information Processing Systems*, pages 325–333, 2016.

Sham Kakade. *On the sample complexity of reinforcement learning*. PhD thesis, University College London, 2003.

Sham M. Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *International Conference on Machine Learning*, 2002.

Sampath Kannan, Michael Kearns, Jamie Morgenstern, Mallesh Pai, Aaron Roth, Rakesh Vohra, and Zhiwei Steven Wu. Fairness incentives for myopic agents. In *Proceedings of the 2017 ACM Conference on Economics and Computation*, pages 369–386. ACM, 2017.

Michael Kearns and Satinder Singh. Near-optimal reinforcement learning in polynomial time. *Machine Learning*, 2002.

Tor Lattimore and Csaba Czepesvari. *Bandit Algorithms*. Cambridge University Press, 2018.

Tor Lattimore and Marcus Hutter. Pac bounds for discounted mdps. In *International Conference on Algorithmic Learning Theory*, pages 320–334. Springer, 2012.

Lihong Li, Michael L Littman, and Thomas J Walsh. Knows what it knows: a framework for self-aware learning. In *Proceedings of the 25th international conference on Machine learning*, pages 568–575. ACM, 2008.

Ashique Rupam Mahmood, Huizhen Yu, and Richard S Sutton. Multi-step off-policy learning without importance sampling ratios. *arXiv preprint arXiv:1702.03006*, 2017.

Aditya Modi, Nan Jiang, Satinder Singh, and Ambuj Tewari. Markov decision processes with continuous side information. In *Algorithmic Learning Theory*, pages 597–618, 2018.

Ian Osband and Benjamin Van Roy. Model-based reinforcement learning and the eluder dimension. In *Advances in Neural Information Processing Systems*, 2014.

Ian Osband, Daniel Russo, and Benjamin Van Roy. (more) efficient reinforcement learning via posterior sampling. In *Advances in Neural Information Processing Systems*, pages 3003–3011, 2013.

Ian Osband, Benjamin Van Roy, and Zheng Wen. Generalization and exploration via randomized value functions. In *International Conference on Machine Learning*, pages 2377–2386, 2016.

Matteo Pirotta, Marcello Restelli, Alessio Pecorino, and Daniele Calandriello. Safe policy iteration. In *PInternational Conference on Machine learning*, pages 307–315, 2013.

Manish Raghavan, Aleksandrs Slivkins, Jennifer Wortman Vaughan, and Zhiwei Steven Wu. The externalities of exploration and how data diversity helps exploitation. *arXiv preprint arXiv:1806.00543*, 2018.

Touqir Sajed, Wesley Chung, and Martha White. High-confidence error estimates for learned value functions. *arXiv preprint arXiv:1808.09127*, 2018.

Alexander L. Strehl, Lihong Li, Eric Wiewiora, John Langford, and Michael L. Littman. PAC model-free reinforcement learning. In *International Conference on Machine Learning*, 2006.

Alexander L. Strehl, Lihong Li, and Michael L. Littman. Reinforcement learning in finite MDPs: PAC analysis. *Journal of Machine Learning Research*, 2009.

István Szita and Csaba Szepesvári. Model-based reinforcement learning with nearly tight exploration complexity bounds. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 1031–1038, 2010.

Philip Thomas and Emma Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. In *International Conference on Machine Learning*, pages 2139–2148, 2016.

Philip Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. High confidence policy improvement. In *International Conference on Machine Learning*, pages 2380–2388, 2015a.

Philip S Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. High-confidence off-policy evaluation. 2015b.

Shlomo Zilberstein and Stuart Russell. Optimal composition of real-time systems. *Artificial Intelligence*, 82(1-2):181–213, 1996.

Appendices

Contents

A Proofs of relationship of IPOC bounds

A.1 Proof of Proposition 2

Proof of Proposition 2. We prove each part separately:

Part 1: With probability at least $1-\delta$, for all T, the regret is bounded as

$$\sum_{k=1}^{T} \Delta_k \leq \sum_{k=1}^{T} \epsilon_k \leq F(W, T, \delta) \tag{10}$$

where the first inequality follows from condition 1 and the second from condition 2a. Hence, the algorithm satisfied a high-probability regret bound $F(W, T, \delta)$ uniformly for all T.

Part 2: By assumption, the cumulative sum of certificates is bounded by $F(W, T, \delta) = \sum_{p=0}^{N} (C_p(W, \delta)T)^{\frac{p}{p+1}}$. Since the minimum is always smaller than the average, the smallest certificates output in the first T episodes is at most

$$\min_{k \in [T]} \epsilon_k \leq \frac{\sum_{k=1}^{T} \epsilon_k}{T} \leq \frac{F(W, T, \delta)}{T} = \sum_{p=0}^{N} C_p(W, \delta)^{\frac{p}{p+1}} T^{-\frac{1}{p+1}}. \tag{11}$$

For $T \geq \frac{C_p(W,\delta)^p (N+1)^{p+1}}{\epsilon^{p+1}}$ we can bound

$$C_p(W, \delta)^{\frac{p}{p+1}} T^{-\frac{1}{p+1}} \leq C_p(W, \delta)^{\frac{p}{p+1}} \left(\frac{C_p(W, \delta)^p (N+1)^{p+1}}{\epsilon^{p+1}} \right)^{-\frac{1}{p+1}} \leq \frac{\epsilon}{N}. \tag{12}$$

As a result, for $T \geq \sum_{p=0}^{N} \frac{C_p(W,\delta)^p (N+1)^{p+1}}{\epsilon^{p+1}} \geq \max_{p \in [N] \cup \{0\}} \frac{C_p(W,\delta)^p (N+1)^{p+1}}{\epsilon^{p+1}}$, we can ensure that $\frac{F(W,T,\delta)}{T} \leq \epsilon$, which completes the proof. □

A.2 Proof of Proposition 3

Proof of Proposition 3. We prove each part separately:

Part 1:

By Definition 1 and the assumption, we have that with probability at least $1-\delta$ for all $\epsilon > 0$, it holds

$$\sum_{k}^{\infty} \mathbf{1}\{\Delta_k > \epsilon\} \leq \sum_{k}^{\infty} \mathbf{1}\{\epsilon_k > \epsilon\} \leq F(W, \epsilon, \delta), \tag{13}$$

where the first inequality follows from condition 1 of IPOC and the second from condition 2b. This proves that the algorithm also satisfies a Uniform-PAC bound as defined by Dann et al. (2017).

Part 2: Since by definition of IPOC, with probability at least $1-\delta$ for all $\epsilon > 0$, the algorithm can output a certificate $\epsilon_k > \epsilon$ at most $F(W, \epsilon, \delta)$ times. By the pigeon hole principle, the algorithm has to output at least one certificate $\epsilon_k \leq \epsilon$ in the first $F(W, \epsilon, \delta) + 1$ episodes.

Part 3: This part of the proof is based on the proof of Theorem A.1 in Dann et al. (2017). For convenience, we omit the dependency of $\tilde{C}$ and C_p on W and δ in the following. We assume

$$F(W, \epsilon, \delta) = \sum_{p=1}^{N} \frac{C_p}{\epsilon^p} \left(\ln \frac{\bar{C}}{\epsilon} \right)^{np} = \sum_{p=1}^{N} C_p g(\epsilon)^p \tag{14}$$

where $\bar{C}$ is chosen so that for all $p \in [N]$ holds $\bar{C}^p \geq \Delta_{\max} C_p$ as well as $\bar{C} \geq \tilde{C}$. We also defined $g(\epsilon) := \frac{1}{\epsilon} \left(\ln \frac{\bar{C}}{\epsilon} \right)^n$.
Consider now the cumulative sum of certificates after T episodes. We distinguish two cases:
Case 1: $T \leq \max_{p \in [N]} \frac{e^p}{\bar{C}^p} N C_p$**.** Note that $e = \exp(1)$ here. We use the fact that all certificates are at most $\Delta_{\max}$ and bound

$$\sum_{k=1}^{T} \epsilon_k \leq \Delta_{\max} T \leq \max_{p \in [N]} \frac{e^p}{\bar{C}^p} N C_p \Delta_{\max} \leq N e^N \tag{15}$$

where the final inequality leverages the assumption on $\bar{C}$.

Case 2: $T \geq \max_{p\in[N]} \frac{e^p}{\bar{C}^p} N C_p$**.** The mistake bound $F(W,\epsilon,\delta)$ is monotonically decreasing for $\epsilon \in (0, \Delta_{\max}]$. If T is large enough, we can therefore find an $\epsilon_{\min} \in (0, \Delta_{\max}]$ such that $F(W,\epsilon,\delta) \leq T$ for all $\epsilon \in (\epsilon_{\min}, \Delta_{\max}]$. The cumulative sum of certificates can then be bounded as follows

$$\sum_{k=1}^T \epsilon_k \leq T\epsilon_{\min} + \int_{\epsilon_{\min}}^{\Delta_{\max}} F(W,\epsilon,\delta) d\epsilon. \tag{16}$$

This bound assumes the worst case where the algorithm first outputs as many $\epsilon_k = \Delta_{\max}$ as allowed and subsequently smaller certificates as controlled by the mistake bound.

Before further simplifying this expression, we claim that

$$\epsilon_{\min} = \frac{\ln\left(\bar{C} \min_{p\in[N]} \left(\frac{T}{NC_p}\right)^{1/p}\right)^n}{\min_{p\in[N]} \left(\frac{T}{NC_p}\right)^{1/p}} \tag{17}$$

satisfies the desired property $F(W,\epsilon_{\min},\delta) \leq T$. To see this, it is sufficient to show that $g(\epsilon_{\min}) \leq \min_{p\in[N]} \left(\frac{T}{NC_p}\right)^{1/p}$, as it implies

$$\sum_{p=1}^N C_p g(\epsilon_{\min})^p = \sum_{p=1}^N C_p \min_{p\in[N]} \left(\frac{T}{NC_p}\right)^{p/p} \leq \sum_{p=1}^N \frac{T}{N}\frac{C_p}{C_p} = T. \tag{18}$$

To show the bound on $g(\epsilon_{\min})$, we verify that for any $x \geq \exp(1)/\bar{C}$

$$g\left(\frac{(\ln(\bar{C}x))^n}{x}\right) = x\frac{\ln\left(\frac{\bar{C}x}{\ln(x\bar{C})^n}\right)^n}{\ln(\bar{C}x)^n} = x\frac{1}{\ln(\bar{C}x)^n}(\ln(\bar{C}x) - n\ln(\ln(x\bar{C})))^n \leq x. \tag{19}$$

Since $\epsilon_{\min}$ has this form for $x = \min_{p\in[N]} \left(\frac{T}{(N)C_p}\right)^{1/p}$ and $\min_{p\in[N]} \left(\frac{T}{(N)C_p}\right)^{1/p} \geq \frac{e}{\bar{C}}$ by case assumption on T, $\epsilon_{\min}$ satisfies the desired property $F(W,\epsilon_{\min},\delta) \leq T$.

We now go back to Equation (16) and simplify it to

$$\begin{aligned}
\sum_{k=1}^T \epsilon_k \leq& T\epsilon_{\min} + \int_{\epsilon_{\min}}^{\Delta_{\max}} F(W,\epsilon,\delta) d\epsilon. && (20)\\
=& T\epsilon_{\min} + \sum_{p=1}^N C_p \int_{\epsilon_{\min}}^{\Delta_{\max}} g(\epsilon)^p d\epsilon && (21)\\
=& T\epsilon_{\min} + \sum_{p=1}^N C_p \int_{\epsilon_{\min}}^{\Delta_{\max}} \frac{1}{\epsilon^p} \ln\left(\frac{\bar{C}}{\epsilon}\right)^{np} d\epsilon && (22)\\
\leq& T\epsilon_{\min} + \sum_{p=1}^N C_p \ln\left(\frac{\bar{C}}{\epsilon_{\min}}\right)^{np} \int_{\epsilon_{\min}}^{\Delta_{\max}} \frac{1}{\epsilon^p} d\epsilon && (23)\\
=& T\epsilon_{\min} + C_1\left(\ln\frac{\bar{C}}{\epsilon_{\min}}\right)^n \ln\frac{\Delta_{\max}}{\epsilon_{\min}} + \sum_{p=2}^N \frac{C_p}{1-p}\left(\ln\frac{\bar{C}}{\epsilon_{\min}}\right)^{np}\left[\Delta_{\max}^{1-p} - \epsilon_{\min}^{1-p}\right]. && (24)
\end{aligned}$$

For each term in the final expression, we show that it is $\tilde{O}\left(\sum_{p=1}^N C_p^{1/p} T^{\frac{p-1}{p}} \operatorname{polylog}(\bar{C}T)\right)$. Starting with the first, we bound

$$
\begin{aligned}
T\epsilon_{\min} =& \frac{T \ln\left(\bar C \min_{p \in [N]} \left(\frac{T}{NC_p}\right)^{1/p}\right)^n}{\min_{p\in[N]}\left(\frac{T}{NC_p}\right)^{1/p}} = \ln\left(\bar C \min_{p \in [N]} \left(\frac{T}{NC_p}\right)^{1/p}\right)^n \max_{p \in [N]} \frac{T N^{1/p} C_p^{1/p}}{T^{1/p}} && (25)\\
\leq& \ln\left(\bar C \frac{T}{NC_1}\right)^n N \max_{p\in[N]} T^{\frac{p-1}{p}} C_p^{1/p} \leq \ln\left(\bar C T\right)^n N \max_{p\in[N]} T^{\frac{p-1}{p}} C_p^{1/p} && (26)\\
=& \tilde O\left(\sum_{p=1}^N C_p^{1/p} T^{\frac{p-1}{p}} \operatorname{polylog}(\bar C T)\right). && (27)
\end{aligned}
$$

For the second term, we start with bounding the inverse of $\epsilon_{\min}$ separately leveraging the case assumption on T:

$$
\begin{aligned}
\frac{1}{\epsilon_{\min}} =& \min_{p\in[N]} \left(\frac{T}{NC_p}\right)^{1/p} \frac{1}{\ln\left(\bar C \min_{p\in[N]}\left(\frac{T}{NC_p}\right)^{1/p}\right)^n} \leq \min_{p\in[N]} \left(\frac{T}{NC_p}\right)^{1/p} \frac{1}{\ln\left(\bar C \min_{p\in[N]}\left(\frac{e^p}{\bar C^p}\right)^{1/p}\right)^n} && (28)\\
\leq& \min_{p\in[N]}\left(\frac{T}{NC_p}\right)^{1/p} \leq T. && (29)
\end{aligned}
$$

The second term of Equation (24) can now be upper bounded by:

$$
C_1 \left(\ln \frac{\bar C}{\epsilon_{\min}}\right)^n \ln \frac{\Delta_{\max}}{\epsilon_{\min}} \leq C_1 \ln(\bar C T)^n \ln(\Delta_{\max} T) \leq C_1 \ln(\bar C T)^{n+1} = \tilde O\left(\sum_{p=1}^N C_p^{1/p} T^{\frac{p-1}{p}} \operatorname{polylog}(\bar C T)\right) \qquad (30)
$$

where the last inequality leverages the definition of $\bar C$. Finally, consider the last term of Equation (24) for $p > 2$:

$$
\begin{aligned}
&\frac{C_p}{1-p}\left(\ln\frac{\bar C}{\epsilon_{\min}}\right)^{np}\left[\Delta_{\max}^{1-p} - \epsilon_{\min}^{1-p}\right] = \frac{C_p}{p-1}\left(\ln\frac{\bar C}{\epsilon_{\min}}\right)^{np}\left[\epsilon_{\min}^{1-p} - \Delta_{\max}^{1-p}\right] \leq \frac{C_p}{p-1}\ln(\bar C T)^{np}\epsilon_{\min}^{1-p} && (31)\\
=&\frac{C_p}{p-1}\ln(\bar C T)^{np}(\epsilon_{\min}^{-1})^{p-1} \leq \frac{C_p}{p-1}\ln(\bar C T)^{np}\left(\frac{T}{NC_p}\right)^{(p-1)/p} \leq \ln(\bar C T)^{np} C_p^{1/p} T^{(p-1)/p} && (32)\\
=&\tilde O\left(\sum_{p=1}^N C_p^{1/p} T^{\frac{p-1}{p}} \operatorname{polylog}(\bar C T)\right). && (33)
\end{aligned}
$$

Combining all bounds above we obtain that

$$
\sum_{k=1}^T \epsilon_k \leq \tilde O\left(\sum_{p=1}^N C_p^{1/p} T^{\frac{p-1}{p}} \operatorname{polylog}(\bar C T)\right) \leq \tilde O\left(\sum_{p=1}^N C_p^{1/p} T^{\frac{p-1}{p}} \operatorname{polylog}(\Delta_{\max}, \tilde C, T)\right). \qquad (34)
$$

$\square$

B Theoretical Analysis of Algorithm 1 for Tabular MDPs

First, we introduce several helpful definitions:

$$
\begin{aligned}
w_k(s,a) =& \mathbb{E}\left[\sum_{h=1}^H \mathbf{1}\{s_{k,h} = s, a_{k,h} = a\} \,\middle|\, a_{k,1:H} \sim \pi_k, s_{k,1} = s_{k,1}\right] && (35)\\
w_{\min} =& \frac{\epsilon c_\epsilon}{S(A \wedge H)H} && (36)\\
c_\epsilon =& \frac{1}{4e} && (37)\\
L_k =& \{(s,a) \in \mathcal{S} \times \mathcal{A} \,:\, w_k(s,a) \geq w_{\min}\} && (38)
\end{aligned}
$$

$$\begin{aligned}
\operatorname{llnp}(x) =& \ln(\ln(\max\{x, e\})) && (39)\\
\operatorname{rng}(x) =& \max(x) - \min(x) && (40)\\
\delta' =& \frac{1}{9}\delta && (41)\\
\phi(n) =& 1 \wedge \sqrt{\frac{1}{n}\left(2\operatorname{llnp}(n) + \ln\frac{3S^2AH}{\delta'}\right)}. && (42)
\end{aligned}$$

The failure event is defined as

$$F = \bigcup_{k=1}^{\infty}\left[F_k^N \cup F_k^P \cup F_k^V \cup F_k^{L1} \cup F_k^{L10} \cup F_k^R\right]. \tag{43}$$

where

$$\begin{aligned}
F_k^N =& \left\{\exists\ s, a : n_k(s,a) < \frac{1}{2}\sum_{i<k} w_i(s,a) - H\ln\frac{SAH}{\delta'}\right\} && (44)\\
F_k^V =& \left\{\exists s, a, t : |(\tilde{P}_k(s,a) - P(s,a))^\top V_{t+1}^\star| \geq \operatorname{rng}(V_{t+1}^\star)\phi(n_k(s,a))\right\} && (45)\\
F_k^P =& \left\{\exists\ s, s', a : |\tilde{P}_k(s'|s,a) - P(s'|s,a)| \geq \sqrt{2P(s'|s,a)}\phi(n_k(s,a)) + \phi(n_k(s,a))^2\right\} && (46)\\
F_k^{L1} =& \left\{\exists\ s, a : \|\tilde{P}_k(s,a) - P(s,a)\|_1 \geq 2\sqrt{S}\phi(n_k(s,a))\right\} && (47)\\
F_k^{L10} =& \left\{\|\tilde{P}_{k0} - P_0\|_1 \geq 2\sqrt{S}\phi(k-1)\right\} && (48)\\
F_k^R =& \left\{\exists\ s, a : |\tilde{r}_k(s,a) - r(s,a)| \geq \phi(n_k(s,a))\right\}. && (49)
\end{aligned}$$

Lemma 7. *The failure event has low probability, that is, $\mathbb{P}(F) \leq \delta$.*

Proof. Combining Lemmas 8, 9, 10 and 11 from below with a union bound, we get that $\mathbb{P}(F) \leq 9\delta' = \delta$. ☐

Lemma 8. *For any $\delta' > 0$, it holds that $\mathbb{P}\left(\bigcup_{k=1}^{\infty} F_k^V\right) \leq 2\delta'$ and $\mathbb{P}\left(\bigcup_{k=1}^{\infty} F_k^R\right) \leq 2\delta'$.*

Proof. Consider a fixed $(s,a,t) \in \mathcal{S} \times \mathcal{A} \times [H]$, and denote by $\mathcal{F}_k$ the sigma-field induced by the first $k-1$ episodes and the k-th episode up to $s_{k,t}$ and $a_{k,t}$ but not $s_{k,t+1}$. Define τ_i to be the index of the episode where (s,a) was observed at time t the ith time. Note that τ_i are stopping times with respect to $\mathcal{F}_i$. Define now the filtration $\mathcal{G}_i = \mathcal{F}_{\tau_i} = \{A \in \mathcal{F}_\infty : A \cap \{\tau_i \leq t\} \in \mathcal{F}_t\ \forall t \geq 0\}$ and $X_k = (V_{t+1}^\star(s_k') - P(s,a,t)^\top V_{t+1}^\star)\mathbb{I}\{\tau_k < \infty\}$ where s_i' is the value of s_{t+1} in episode τ_i (or arbitrary, if $\tau_i = \infty$).

By the Markov property of the MDP, we have that X_i is a martingale difference sequence with respect to the filtration $\mathcal{G}_i$. Furthermore, since $\mathbb{E}[X_i|\mathcal{G}_{i-1}] = 0$ and $|X_i| \in [0, \operatorname{rng}(V_{t+1}^\star)]$, X_i is conditionally $\operatorname{rng}(V_{t+1}^\star)/2$-sub-Gaussian due to Hoeffding's Lemma, i.e., satisfies $\mathbb{E}[\exp(\lambda X_i)|\mathcal{G}_{i-1}] \leq \exp(\lambda^2 \operatorname{rng}(V_{t+1}^\star)^2/2)$.

We can therefore apply Lemma F.1 by Dann et al. (2017) and conclude that

$$\mathbb{P}\left(\exists k : |(\hat{P}_k(s,a,t) - P(s,a,t))^\top V_{t+1}^\star| \geq \sqrt{\frac{\operatorname{rng}(V_{t+1}^\star)^2}{n_{tk}(s,a)}\left(2\operatorname{llnp}(n_{tk}(s,a)) + \ln\frac{3}{\delta'}\right)}\right) \leq 2\delta'. \tag{50}$$

Analogously,

$$\mathbb{P}\left(\exists k : |\hat{r}_k(s,a,t) - r(s,a,t)| \geq \sqrt{\frac{1}{n_{tk}(s,a)}\left(2\operatorname{llnp}(n_{tk}(s,a)) + \ln\frac{3}{\delta'}\right)}\right) \leq 2\delta'. \tag{51}$$

Applying the union bound over all $s \in \mathcal{S}, a \in \mathcal{A}$ and $t \in [H]$, we obtain the desired statement for F^V. In complete analogy using the same filtration, we can show the statement for F^R. ☐

Lemma 9. *For any $\delta' > 0$, it holds that $\mathbb{P}\left(\bigcup_{k=1}^{\infty} F_k^P\right) \leq 2\delta'$.*

Proof. Consider first a fixed $s', s \in \mathcal{S}$, $t \in [H]$ and $a \in \mathcal{A}$. Let K denote the number of times the triple s, a, t was encountered in total during the run of the algorithm. Define the random sequence X_i as follows. For $i \leq K$, let X_i be the indicator of whether s' was the next state when s, a, t was encountered the ith time and for $i > K$, let $X_i \sim \text{Bernoulli}(P(s'|s,a,t))$ be drawn i.i.d. By construction this is a sequence of i.i.d. Bernoulli random variables with mean $P(s'|s,a,t)$. Further the event

$$\bigcup_k \left\{ \left|\hat{P}_k(s'|s,a,t) - P(s'|s,a,t)\right| \geq \sqrt{\frac{2P(s'|s,a,t)}{n_{tk}(s,a)}\left(2\operatorname{llnp}(n(s,a,t)) + \ln\frac{3S^2AH}{\delta'}\right)} \right. \tag{52}$$

$$\left. + \frac{1}{n_{tk}(s,a)}\left(2\operatorname{llnp}(n_{tk}(s,a)) + \ln\frac{3S^2AH}{\delta'}\right) \right\} \tag{53}$$

is contained in the event

$$\bigcup_i \left\{ |\hat{\mu}_i - \mu| \geq \sqrt{\frac{2\mu}{i}\left(2\operatorname{llnp}(i) + \ln\frac{3S^2AH}{\delta'}\right)} + \frac{1}{i}\left(2\operatorname{llnp}(i) + \ln\frac{3S^2AH}{\delta'}\right) \right\} \tag{54}$$

where $\mu = P(s'|s,a,t)$ and $\hat{\mu}_i = i^{-1}\sum_{j=1}^i X_j$ and whose probability can be bounded by $2\delta'/S^2/A/H$ using Lemma F.2 by Dann et al. (2017). The statement now follows by applying the union bound. ☐

Lemma 10. *For any $\delta' > 0$, it holds that $\mathbb{P}\left(\bigcup_{k=1}^\infty F_k^{L1}\right) \leq \delta'$ and $\mathbb{P}\left(\bigcup_{k=1}^\infty F_k^{L10}\right) \leq \delta'$.*

Proof. Using the same argument as in the proof of Lemma 9 for F_k^{L1}, the statement follows from Lemma F.3 by Dann et al. (2017). ☐

Lemma 11. *For any δ', it holds that $\mathbb{P}\left(\bigcup_{k=1}^\infty F_k^N\right) \leq \delta'$.*

Proof. Consider a fixed $s \in \mathcal{S}, a \in \mathcal{A}, t \in [H]$. We define $\mathcal{F}_k$ to be the sigma-field induced by the first $k-1$ episodes and $s_{k,1}$. Let X_k as the indicator whether s, a was observed in episode k at time t. The probability $\mathbb{P}(s = s_{k,t}, a = a_{k,t}|s_{k,1})$ of whether $X_k = 1$ is $\mathcal{F}_k$-measurable and hence we can apply Lemma F.4 by Dann et al. (2017) with $W = \ln\frac{SAH}{\delta'}$ and obtain that $\mathbb{P}\left(\bigcup_k F_k^N\right) \leq \delta'$ after summing over all statements for $t \in [H]$ and applying the union bound. ☐

B.1 Admissibility of Certificates

We now show that the algorithm always gives a valid certificate in all episodes (outside the failure event). The following three lemmas prove the admissibility.

Lemma 12 (Lower bound admissible)**.** *In event F^c, for all episodes k, $h \in [H]$ and $s, a \in \mathcal{S} \times \mathcal{A}$*

$$Q_h^{\pi_k}(s,a) \geq \underset{\sim}{Q}_{k,h}(s,a). \tag{55}$$

Proof. Consider a fixed episode k. For $h = H+1$ the claim holds by definition. Assume the claim holds for $h+1$ and consider

$$Q_h^{\pi_k}(s,a) - \underset{\sim}{Q}_{k,h}(s,a) \tag{56}$$

$$= P(s,a)V_{h+1}^{\pi_k} + r(s,a) - \hat{r}_k(s,a) + \underset{\sim}{\psi}_{k,h}(s,a) - \hat{P}_k(s,a)\underset{\sim}{V}_{k,h+1} \tag{57}$$

$$= \underset{\sim}{\psi}_{k,h}(s,a) + P(s,a)(V_{h+1}^{\pi_k} - \underset{\sim}{V}_{k,h+1}) + (P(s,a) - \hat{P}_k(s,a))\underset{\sim}{V}_{k,h+1} + r(s,a) - \hat{r}_k(s,a) \tag{58}$$

using induction hypothesis and Hölder's inequality we bound

$$\geq \underset{\sim}{\psi}_{k,h}(s,a) + 0 - \|P(s,a) - \hat{P}_k(s,a)\|_1 \|\underset{\sim}{V}_{k,h+1}\|_\infty - |r(s,a) - \hat{r}_k(s,a)| \tag{59}$$

applying definition of F^c

$$\geq 2\sqrt{S}V_h^{\max}\phi(n_k(s,a)) - 2\sqrt{S}\phi(n_k(s,a))V_{h+1}^{\max} - \phi(n_k(s,a)) \geq 0\,. \tag{60}$$

☐

Lemma 13 (Upper bound admissible)**.** *In event F^c, for all episodes k, $h \in [H]$ and $s, a \in \mathcal{S} \times \mathcal{A}$*

$$Q^\star_h(s,a) \leq \tilde{Q}_{k,h}(s,a). \tag{61}$$

Proof. Consider a fixed episode k. For $h = H+1$ the claim holds by definition. Assume the claim holds for $h+1$ and consider $\tilde{Q}_{k,h}(s,a) - Q^\star_h(s,a)$. Since $Q^\star_h \leq V^{\max}_h$, this quantity is non-negative when $\tilde{Q}_{k,h}(s,a) = V^{\max}_h$. In the other case

$$\tilde{Q}_{k,h}(s,a) - Q^\star_h(s,a) \tag{62}$$
$$= \hat{r}_k(s,a) + \tilde{\psi}_{k,h}(s,a) + \hat{P}_k(s,a)\tilde{V}_{k,h+1} - P(s,a)V^\star_{h+1} - r(s,a) \tag{63}$$
$$= \hat{r}_k(s,a) - r(s,a) + \tilde{\psi}_{k,h}(s,a) + \hat{P}_k(s,a)(\tilde{V}_{k,h+1} - V^\star_{h+1}) + (\hat{P}_k(s,a) - P(s,a))V^\star_{h+1} \tag{64}$$

by induction hypothesis $\hat{P}_k(s,a)(\tilde{V}_{k,h+1} - V^\star_{h+1}) \geq 0$

$$\geq \hat{r}_k(s,a) - r(s,a) + \tilde{\psi}_{k,h}(s,a) + (\hat{P}_k(s,a) - P(s,a))V^\star_{h+1} \tag{65}$$
$$\geq -|\hat{r}_k(s,a) - r(s,a)| + \tilde{\psi}_{k,h}(s,a) - |(\hat{P}_k(s,a) - P(s,a))V^\star_{h+1}| \tag{66}$$

Applying the definition of $\tilde{\psi}_{k,h}$ and the failure event F^c

$$\geq \tilde{\psi}_{k,h}(s,a) - \phi(n_k(s,a)) - V^{\max}_{h+1}\phi(n_k(s,a)) \tag{67}$$
$$\geq \tilde{\psi}_{k,h}(s,a) - V^{\max}_h \phi(n_k(s,a)) = 0\,. \tag{68}$$

□

Lemma 14 (Optimality guarantees admissible)**.** *In event F^c (outside the failure event), for all episodes k, the certificate is valid, that is, $\Delta_k \leq \epsilon_k$.*

Proof. Since we assume that the initial state is deterministic, we have $\Delta_k = V^\star(s_{k,1}) - V^{\pi_k}(s_{k,1})$. It then follows using admissibility of upper and lower bounds (Lemmas 12 and 13) that

$$\Delta_k = V^\star(s_{k,1}) - V^{\pi_k}(s_{k,1} \leq \tilde{V}_{k,1}(s_{k,1}) - \underset{\sim}{V}_{k,1}(s_{k,1}) = \epsilon_k\,. \tag{69}$$

□

B.2 Bounding the Number of Large Certificates

We start by deriving an upper bound on each certificate in terms of the confidence bound widths.

Lemma 15 (Upper bound on certificates)**.** *In event F^c (outside the failure event), for all episodes k, the following bound on the optimality-guarantee holds*

$$\epsilon_k \leq e \sum_{s,a \in \mathcal{S} \times \mathcal{A}} w_k(s,a) \min\left\{H, 3\sqrt{S}H\phi(n_k(s,a)) + 3SH^2\phi(n_k(s,a))^2\right\}\,. \tag{70}$$

Proof. By definition of the upper and lower bound estimates

$$\tilde{Q}_{k,h}(s,a) - \underset{\sim}{Q}_{k,h}(s,a) \tag{71}$$
$$\leq \tilde{\psi}_{k,h}(s,a) + \underset{\sim}{\psi}_{k,h}(s,a) + \hat{P}_k(s,a)\tilde{V}_{k,h+1} - \hat{P}_k(s,a)\underset{\sim}{V}_{k,h+1} \tag{72}$$
$$= \tilde{\psi}_{k,h}(s,a) + \underset{\sim}{\psi}_{k,h}(s,a) + P(s,a)(\tilde{V}_{k,h+1} - \underset{\sim}{V}_{k,h+1}) + (\hat{P}_k - P)(s,a)(\tilde{V}_{k,h+1} - \underset{\sim}{V}_{k,h+1}). \tag{73}$$

We bound the final term with Lemma 20 and $f = \tilde{V}_{k,h+1} - \underset{\sim}{V}_{k,h+1} \in [0, H]$ and $C = H$ to get

$$\tilde{Q}_{k,h}(s,a) - \underset{\sim}{Q}_{k,h}(s,a) \tag{74}$$
$$\leq \tilde{\psi}_{k,h}(s,a) + \underset{\sim}{\psi}_{k,h}(s,a) + \left(1 + \frac{1}{H}\right) P(s,a)(\tilde{V}_{k,h+1} - \underset{\sim}{V}_{k,h+1}) + 3SH^2\phi(n_k(s,a))^2 \tag{75}$$
$$\leq 3H\sqrt{S}\phi(n_k(s,a)) + \left(1 + \frac{1}{H}\right) P^{\pi_k}_h(s,a)(\tilde{Q}_{k,h+1} - \underset{\sim}{Q}_{k,h+1}) + 3SH^2\phi(n_k(s,a))^2, \tag{76}$$

where we bounded $\tilde{\psi}_{k,h}(s,a) + \underset{\sim}{\psi}_{k,h}(s,a) \leq 3H\sqrt{S}\phi(n_k(s,a))$ in the final step. Here $P_h^{\pi_k}(s,a)f = \mathbb{E}[f(s_{k,h+1}, \pi(s_{k,h+1}, h+1))|s_{k,h} = s, a_{k,h} = a, \pi_k]$ denotes the composition of $P(s,a)$ and the policy action selection operator at time $h+1$. In addition to the bound above, by construction also $0 \leq \tilde{Q}_{k,h}(s,a) - \underset{\sim}{Q}_{k,h}(s,a) \leq V_h^{\max}$ holds at all times. Resolving this recursive bound yields

$$\epsilon_k = (\tilde{V}_{k,1} - \underset{\sim}{V}_{k,1})(s_{k,1}) = (\tilde{Q}_{k,1} - \underset{\sim}{Q}_{k,1})(s_{k,1}, \pi_k(s_{k,1}, 1)) \tag{77}$$

$$\leq \sum_{s,a \in \mathcal{S}\times\mathcal{A}} \sum_{h=1}^{H} \left(1 + \frac{1}{H}\right)^h \mathbb{P}(s_{k,h} = s, a_{k,h} = a | s_{k,1}, \pi_k)[V_h^{\max} \wedge (3\sqrt{S}H\phi(n_k(s,a)) + 3SH^2\phi(n_k(s,a))^2)] \tag{78}$$

$$\leq e \sum_{s,a \in \mathcal{S}\times\mathcal{A}} w_k(s,a) \min\left\{V_1^{\max}, 3\sqrt{S}H\phi(n_k(s,a)) + 3SH^2\phi(n_k(s,a))^2\right\}. \tag{79}$$

□

We now follow the proof structure of Dann et al. (2017) amd define *nice* episodes, in which all state-action pairs either have low probability of occurring or the sum of probability of occurring in the previous episodes is large enough so that outside the failure event we can guarantee that

$$n_k(s,a) \geq \frac{1}{4}\sum_{i<k} w_i(s,a). \tag{80}$$

This allows us then to bound the number of nice episodes with large certificates by the number of times terms of the form

$$\sum_{s,a \in L_k} w_k(s,a)\sqrt{\frac{\operatorname{llnp}(n_k(s,a)) + D}{n_k(s,a)}} \tag{81}$$

can exceed a chosen threshold (see Lemma 21 below).

Definition 16 (Nice Episodes)**.** *An episode k is* nice *if and only if for all $s \in \mathcal{S}$, $a \in \mathcal{A}$ the following two conditions hold:*

$$w_k(s,a) \leq w_{\min} \quad \vee \quad \frac{1}{4}\sum_{i<k} w_i(s,a) \geq H \ln\frac{SAH}{\delta'}. \tag{82}$$

We denote the set of indices of all nice episodes as $\mathcal{N} \subseteq \mathbb{N}$.

Lemma 17 (Properties of nice episodes)**.** *If an episode k is nice, i.e., $k \in \mathcal{N}$, then on F^c (outside the failure event) for all $s \in \mathcal{S}$, $a \in \mathcal{A}$ the following statement holds:*

$$w_k(s,a) \leq w_{\min} \quad \vee \quad n_k(s,a) \geq \frac{1}{4}\sum_{i<k} w_i(s,a). \tag{83}$$

Proof. Since we consider the event F^{N^c}, it holds for all s, a pairs with $w_k(s,a) > w_{\min}$

$$n_k(s,a) \geq \frac{1}{2}\sum_{i<k} w_i(s,a) - H\ln\frac{SAH}{\delta'} \geq \frac{1}{4}\sum_{i<k} w_i(s,a) \tag{84}$$

for $k \in N$. □

Lemma 18 (Number of episodes that are not nice)**.** *On the good event F^c, the number episodes that are not nice is at most*

$$\frac{4S^2A(A \wedge H)H^2}{c_\epsilon \epsilon} \ln\frac{SAH}{\delta'}. \tag{85}$$

Proof. If an episode k is not nice, then there is s, a with $w_k(s,a) > w_{\min}$ and

$$\sum_{i<k} w_i(s,a) < 4H \ln \frac{SAH}{\delta'}. \tag{86}$$

The sum on the left-hand side of this inequality increases by at least $w_{\min}$ after the episode while the right hand side stays constant, this situation can occur at most

$$\frac{4SAH}{w_{\min}} \ln \frac{SAH}{\delta'} = \frac{4S^2A(A \wedge H)H^2}{c_\epsilon \epsilon} \ln \frac{SAH}{\delta'} \tag{87}$$

times in total. □

B.3 Proof of IPOC bound of `ORLC`, Theorem 4

We are now equipped with all tools to complete the proof of Theorem 4:

Proof of Theorem 4. Consider event F^c which has probability at least $1-\delta$ due to Lemma 7. In this event, all optimality guarantees are admissible by Lemma 14. Further, using Lemma 15, the optimality guarantees are bounded as

$$\epsilon_k \leq e \sum_{s,a \in \mathcal{S}\times\mathcal{A}} w_k(s,a) \min\left\{H, 3\sqrt{S}H\phi(n_k(s,a)) + 3SH^2\phi(n_k(s,a))^2\right\}. \tag{88}$$

It remains to show that for any given threshold $\epsilon > 0$ this bound can exceed ϵ only the number of times prescribed by Equation 2. Recall the definition of L_k as the set of state-action pairs with significant probability of occurring, $L_k = \{(s,a) \in \mathcal{S}\times\mathcal{A} : w_k(s,a) \geq w_{\min}\}$, and split the sum as

$$\epsilon_k \leq e \sum_{s,a \notin L_k} w_k(s,a)H \tag{89}$$

$$+ e \sum_{s,a \in L_k} w_k(s,a) \left(3\sqrt{S}H\phi(n_k(s,a)) + 3SH^2\phi(n_k(s,a))^2\right) \tag{90}$$

and bound each of the three remaining terms individually. First, the definition of L_k was chosen such that

$$e \sum_{s,a \notin L_k} w_k(s,a)H \leq eHw_{\min}S(A \wedge H) = \frac{eHS(A \wedge H)\epsilon c_\epsilon}{HS(A \wedge H)} = c_\epsilon e \epsilon, \tag{91}$$

where we used the fact that the number of positive $w_k(s,a)$ is at most SA as well as SH per episode k. Second, we use Lemma 21 with $r=2, C=2, D=\frac{1}{2}\ln\frac{3S^2AH}{\delta'}$ and $\epsilon' = \frac{c_\epsilon \epsilon}{3\sqrt{S}H}$ to bound

$$\sum_{s,a \in L_k} w_k(s,a)\phi(n_k(s,a)) \leq \frac{c_\epsilon \epsilon}{3\sqrt{S}H} \tag{92}$$

on all but at most

$$O\left(\frac{S^2AH^3}{\epsilon^2} \operatorname{polylog}(S,A,H,1/\epsilon,\ln(1/\delta))\right) \tag{93}$$

nice episodes. Similarly, Lemma 21 with $r=1, C=2, D=\frac{1}{2}\ln\frac{3S^2AH}{\delta'}$ and $\epsilon' = \frac{c_\epsilon \epsilon}{3SH^2}$ to bound

$$\sum_{s,a \in L_k} w_k(s,a)\phi(n_k(s,a))^2 \leq \frac{c_\epsilon \epsilon}{6SH^2} \tag{94}$$

on all but at most

$$O\left(\frac{S^2AH^2}{\epsilon} \operatorname{polylog}(S,A,H,1/\epsilon,\ln(1/\delta))\right) \tag{95}$$

nice episodes. Further, Lemma 18 states that the number of episodes that are not nice is bounded by

$$O\left(\frac{S^2 A(A \wedge H)H^2}{\epsilon} \operatorname{polylog}(S, A, H, 1/\epsilon, \ln(1/\delta))\right). \tag{96}$$

Taking all these bounds together, we can bound $\epsilon \leq 4c_\epsilon \epsilon \leq \epsilon$ for all episodes k except at most

$$O\left(\left(\frac{S^2 AH^3}{\epsilon^2} + \frac{S^2 A(A \wedge H)H^2}{\epsilon}\right) \operatorname{polylog}(S, A, H, 1/\epsilon, \ln(1/\delta))\right) \tag{97}$$

which completes the proof. □

B.4 Technical Lemmas

Lemma 19. *Let $\tau \in (0, \hat{\tau}]$ and $D \geq 1$. Then for all $x \geq \check{x} = \frac{\ln(C/\tau)+D}{\tau}$ with $C = 16 \vee \hat{\tau} D^2$, the following inequality holds*

$$\frac{\operatorname{llnp}(x) + D}{x} \leq \tau. \tag{98}$$

Proof. Since by Lemma 23 the function $\frac{\operatorname{llnp}(x)+D}{x}$ is monotonically decreasing in x, we can bound

$$\frac{\operatorname{llnp}(x) + D}{x} \leq \frac{\operatorname{llnp}(\check{x}) + D}{\check{x}} = \frac{\operatorname{llnp}(\check{x}) + D}{\ln(C/\tau) + D}\tau. \tag{99}$$

It remains to show that $\ln(\check{x}) \vee 1 \leq \frac{C}{\tau}$. First note that $\frac{C}{\tau} \geq \frac{C}{\hat{\tau}} \geq 1$ because $C \geq 1$. Also, we can bound using $\ln(x) \leq 2\sqrt{x}$

$$\ln(\check{x}) = \ln\left(\frac{\ln(C/\tau) + D}{\tau}\right) \leq 2\sqrt{\frac{\ln(C/\tau) + D}{\tau}} \leq 2\sqrt{\frac{2\sqrt{C/\tau} + D}{\tau}} \tag{100}$$

$$\leq 4\left(\frac{C}{\tau^3}\right)^{1/4} = \frac{C}{\tau}\left(\frac{4}{\sqrt{C}}\frac{\tau^{1/4}}{C^{1/4}}\right) \leq \frac{C}{\tau}, \tag{101}$$

since $\sqrt{C} \geq 4$ and $C \geq \hat{\tau} D^2 \geq \hat{\tau}$. □

Lemma 20. *Let $f : \mathcal{S} \mapsto [0, \hat{f}]$ be a function on states. In event F^c (outside the failure event), for all episodes k, states $s \in \mathcal{S}$ and actions $a \in \mathcal{A}$, the following bound holds for any $C > 0$*

$$|(\hat{P}_k - P)(s, a)f| \leq (2C + 1)\hat{f} S\phi(n_k(s, a))^2 + \frac{1}{C}P(s, a)f. \tag{102}$$

Proof.

$$|(\hat{P}_k - P)(s, a)f| \leq \sum_{s' \in \mathcal{S}} |(\tilde{P}_k - P)(s'|s, a)|f \tag{103}$$

applying the concentration statements in the definition of F^c

$$\leq \sum_{s' \in \mathcal{S}} (f(s')\phi(n_k(s, a))^2 + \sqrt{2P(s'|s, a)}\phi(n_k(s, a))f(s')) \tag{104}$$

$$\leq S\hat{f}\phi(n_k(s, a))^2 + \phi(n_k(s, a)) \sum_{s' \in \mathcal{S}} \sqrt{2P(s'|s, a)}f(s') \tag{105}$$

$$= S\hat{f}\phi(n_k(s, a))^2 + \phi(n_k(s, a)) \sum_{s' \in \mathcal{S}} \sqrt{\frac{2}{P(s'|s, a)}}P(s'|s, a)f(s'). \tag{106}$$

Splitting the last sum based on whether $\sqrt{P(s'|s,a)}$ is smaller or larger than $\sqrt{2}C\phi(n_k(s,a))$

$$\leq S\hat{f}\phi(n_k(s,a))^2 + \frac{1}{C}P_h(s,a)f \tag{107}$$

$$+ \phi(n_k(s,a))\sum_{s'\in\mathcal{S}}\sqrt{\frac{2}{P(s'|s,a)}}P(s'|s,a)f(s')\mathbf{1}\{\sqrt{P(s'|s,a)} < \sqrt{2}C\phi(n_k(s,a))\} \tag{108}$$

$$\leq S\hat{f}\phi(n_k(s,a))^2 + \frac{1}{C}P(s,a)f + \phi(n_k(s,a))^2 2C\hat{f}S \tag{109}$$

$$\leq (2C+1)\hat{f}S\phi(n_k(s,a))^2 + \frac{1}{C}P(s,a)f\,. \tag{110}$$

□

Lemma 21 (Rate Lemma, Adaption of Lemma E.3 by Dann et al. (2017))**.** *Let $r \geq 1$ fix and $C > 0$ which can depend polynomially on the relevant quantities and $\epsilon' > 0$ and let $D \geq 1$ which can depend poly-logarithmically on the relevant quantities. Then*

$$\sum_{s,a\in L_k} w_k(s,a)\left(\frac{C(\operatorname{llnp}(n_k(s,a)) + D)}{n_k(s,a)}\right)^{1/r} \leq \epsilon' \tag{111}$$

on all but at most

$$\frac{6CASH^{r-1}}{\epsilon'^r}\operatorname{polylog}(S,A,H,\delta^{-1},\epsilon'^{-1}). \tag{112}$$

nice episodes.

Proof. Define

$$\Delta_k = \sum_{s,a\in L_k} w_k(s,a)\left(\frac{C(\operatorname{llnp}(n_k(s,a)) + D)}{n_k(s,a)}\right)^{1/r} \tag{113}$$

$$= \sum_{s,a\in L_k} w_k(s,a)^{1-\frac{1}{r}}\left(w_k(s,a)\frac{C(\operatorname{llnp}(n_k(s,a)) + D)}{n_k(s,a)}\right)^{1/r}. \tag{114}$$

We first bound using Hölder's inequality

$$\Delta_k \leq \left(\sum_{s,a\in L_k}\frac{CH^{r-1}w_k(s,a)(\operatorname{llnp}(n_k(s,a)) + D)}{n_k(s,a)}\right)^{\frac{1}{r}}. \tag{115}$$

Using the property in Lemma 17 of nice episodes as well as the fact that $w_k(s,a) \leq H$ and $\sum_{i<k} w_i(s,a) \geq 4H\ln\frac{SAH}{\delta'} \geq 4H\ln(2) \geq 2H$, we bound

$$n_{tk}(s,a) \geq \frac{1}{4}\sum_{i<k} w_{ti}(s,a) \geq \frac{1}{6}\sum_{i\leq k} w_{ti}(s,a). \tag{116}$$

The function $\frac{\operatorname{llnp}(x)+D}{x}$ is monotonically decreasing in $x \geq 0$ since $D \geq 1$ (see Lemma 23). This allows us to bound

$$\Delta_k^r \leq \sum_{s,a\in L_k}\frac{CH^{r-1}w_k(s,a)(\operatorname{llnp}(n_k(s,a)) + D)}{n_k(s,a)} \tag{117}$$

$$\leq 6CH^{r-1}\sum_{s,a\in L_k}\frac{w_k(s,a)\left(\operatorname{llnp}\left(\frac{1}{6}\sum_{i\leq k} w_i(s,a)\right) + D\right)}{\sum_{i\leq k} w_i(s,a)} \tag{118}$$

$$\leq 6CH^{r-1}\sum_{s,a\in L_k}\frac{w_k(s,a)\left(\operatorname{llnp}\left(\sum_{i\leq k} w_i(s,a)\right) + D\right)}{\sum_{i\leq k} w_i(s,a)}. \tag{119}$$

Assume now $\Delta_k > \epsilon'$. In this case the right-hand side of the inequality above is also larger than ϵ'^r and there is at least one (s,a) with $w_k(s,a) > w_{\min}$ and

$$\frac{6CSAH^{r-1}\left(\operatorname{llnp}\left(\sum_{i\le k} w_i(s,a)\right)+D\right)}{\sum_{i\le k} w_i(s,a)} > \epsilon'^r \tag{120}$$

$$\Leftrightarrow \frac{\operatorname{llnp}\left(\sum_{i\le k} w_i(s,a)\right)+D}{\sum_{i\le k} w_i(s,a)} > \frac{\epsilon'^r}{6CSAH^{r-1}}. \tag{121}$$

Let us denote $C' = \frac{6CASH^{r-1}}{\epsilon'^r}$. Since $\frac{\operatorname{llnp}(x)+D}{x}$ is monotonically decreasing and $x = C'^2 + 3C'D$ satisfies $\frac{\operatorname{llnp}(x)+D}{x} \le \frac{\sqrt{x}+D}{x} \le \frac{1}{C'}$, we know that if $\sum_{i\le k} w_i(s,a) \ge C'^2 + 3C'D$ then the above condition cannot be satisfied for s,a. Since each time the condition is satisfied, it holds that $w_k(s,a) > w_{\min}$ and so $\sum_{i\le k} w_i(s,a)$ increases by at least $w_{\min}$, it can happen at most

$$m \le \frac{SA(C'^2 + 3C'D)}{w_{\min}} \tag{122}$$

times that $\Delta_k > \epsilon'$. Define $K = \{k : \Delta_k > \epsilon'\} \cap N$ and we know that $|K| \le m$. Now we consider the sum

$$\sum_{k\in K} \Delta_k^r \le \sum_{k\in K} 6CH^{r-1} \sum_{s,a\in L_k} \frac{w_k(s,a)\left(\operatorname{llnp}\left(\sum_{i\le k} w_i(s,a)\right)+D\right)}{\sum_{i\le k} w_i(s,a)} \tag{123}$$

$$\le 6CH^{r-1}\left(\operatorname{llnp}\left(C'^2+3C'D\right)+D\right) \sum_{s,a\in L_k}\sum_{k\in K} \frac{w_k(s,a)}{\sum_{i\le k} w_i(s,a)\mathbb{I}\{w_i(s,a)\ge w_{\min}\}}. \tag{124}$$

For every (s,a), we consider the sequence of $w_i(s,a) \in [w_{\min}, H]$ with $i \in I = \{i \in \mathbb{N} : w_i(s,a) \ge w_{\min}\}$ and apply Lemma 22. This yields that

$$\sum_{k\in K} \frac{w_k(s,a)}{\sum_{i\le k} w_i(s,a)\mathbb{I}\{w_i(s,a)\ge w_{\min}\}} \le 1 + \ln(mH/w_{\min}) = \ln\left(\frac{Hme}{w_{\min}}\right) \tag{125}$$

and hence

$$\sum_{k\in K} \Delta_k^r \le 6CASH^{r-1} \ln\left(\frac{Hme}{w_{\min}}\right)\left(\operatorname{llnp}\left(C'^2+3C'D\right)+D\right). \tag{126}$$

Since each element in K has to contribute at least ϵ'^r to this bound, we can conclude that

$$\sum_{k\in N} \mathbb{I}\{\Delta_k \ge \epsilon'\} \le \sum_{k\in K} \mathbb{I}\{\Delta_k \ge \epsilon'\} \le |K| \le \frac{6CASH^{r-1}}{\epsilon'^r} \ln\left(\frac{Hme}{w_{\min}}\right)\left(\operatorname{llnp}\left(C'^2+3C'D\right)+D\right). \tag{127}$$

Since $\ln\left(\frac{me}{w_{\min}}\right)\left(\operatorname{llnp}\left(C'^2+3C'D\right)+D\right)$ is $\operatorname{polylog}(S,A,H,\delta^{-1},\epsilon'^{-1})$, the proof is complete. ☐

Lemma 22 (Lemma E.5 by Dann et al. (2017)). *Let a_i be a sequence taking values in $[a_{\min}, a_{\max}]$ with $a_{\max} \ge a_{\min} > 0$ and $m > 0$, then*

$$\sum_{k=1}^{m} \frac{a_k}{\sum_{i=1}^{k} a_i} \le \ln\left(\frac{mea_{\max}}{a_{\min}}\right). \tag{128}$$

Lemma 23 (Properties of llnp, Lemma E.6 by Dann et al. (2017)). *The following properties hold:*

1. llnp *is continuous and nondecreasing.*
2. $f(x) = \frac{\operatorname{llnp}(nx)+D}{x}$ *with $n \ge 0$ and $D \ge 1$ is monotonically decreasing on $\mathbb{R}_+$.*
3. $\operatorname{llnp}(xy) \le \operatorname{llnp}(x) + \operatorname{llnp}(y) + 1$ *for all $x, y \ge 0$.*

C Theoretical analysis of Algorithm 2 for finite episodic MDPs with side information

C.1 Failure event and bounding the failure probability

We define the following failure event

$$F = F^{(r)} \cup F^{(p)} \cup F^{O} \tag{129}$$

where

$$F^{(r)} = \left\{ \exists s, a \in \mathcal{S} \times \mathcal{A}, k \in \mathbb{N} \ : \ \|\hat{\theta}^{(r)}_{k,s,a} - \theta^{(r)}_{s,a}\|_{N^{(r)}_{k,s,a}} \geq \sqrt{\lambda}\|\theta^{(r)}_{s,a}\|_2 + \sqrt{\frac{1}{2}\log\frac{1}{\delta'} + \frac{1}{4}\log\frac{\det N^{(r)}_{k,s,a}}{\det \lambda I}} \right\}, \tag{130}$$

$$F^{(p)} = \Bigg\{ \exists s', s, a \in \mathcal{S} \times \mathcal{S} \times \mathcal{A}, k \in \mathbb{N} \ : \ \|\hat{\theta}^{(p)}_{k,s',s,a} - \theta^{(p)}_{s',s,a}\|_{N^{(p)}_{k,s,a}} \tag{131}$$

$$\geq \sqrt{\lambda}\|\theta^{(p)}_{s',s,a}\|_2 + \sqrt{\frac{1}{2}\log\frac{1}{\delta'} + \frac{1}{4}\log\frac{\det N^{(p)}_{k,s,a}}{\det \lambda I}} \Bigg\}, \tag{132}$$

$$F^{O} = \Bigg\{ \exists T \in \mathbb{N} \ : \ \sum_{k=1}^{T} \sum_{s,a \in \mathcal{S}\times\mathcal{A}} \sum_{h=1}^{H} [\mathbb{P}(s_{k,h} = s, a_{k,h} = a | s_{k,1}, \pi_k) - \mathbf{1}\{s = s_{k,h}, a = a_{k,h}\}] \tag{133}$$

$$\geq SH\sqrt{T \log\frac{6\log(2T)}{\delta'}} \Bigg\}. \tag{134}$$

Lemma 24. *With $\delta' \leq \frac{\delta}{SA + S^2A + SH}$, the failure probability $\mathbb{P}(F)$ is bounded by δ.*

Proof. Consider an arbitrary $s \in \mathcal{S}$, $a \in \mathcal{A}$ and define $\mathcal{F}_t$ where $t = Hk + h$ with $h \in [H]$ is the running time step index as follows: $\mathcal{F}_t$ is the sigma-field induced by all observations up to $s_{k,h}$ and $a_{k,h}$ including x_k but not $r_{k,h}$ and not $s_{k,h+1}$. Then $\eta_t = 2\mathbf{1}\{s_{k,h} = s, a_{k,h} = a\}((x^{(r)}_k)^\top \theta^{(r)}_{s,a} - r_{k,h})$ is a martingale difference sequence adapted to $\mathcal{F}_t$. Moreover, since η_t takes values in $[2(x^{(r)}_k)^\top \theta^{(r)}_{s,a} - 2, 2(x^{(r)}_k)^\top \theta^{(r)}_{s,a}]$ almost surely it is conditionally sub-Gaussian with parameter 1. We can then apply Theorem 20.2 in Lattimore and Czepesvari (2018) to get

$$2\|\hat{\theta}^{(r)}_{k,s,a} - \theta^{(r)}_{s,a}\|_{N^{(r)}_{k,s,a}} \leq \sqrt{\lambda}2\|\theta^{(r)}_{s,a}\|_2 + \sqrt{2\log\frac{1}{\delta'} + \log\frac{\det N^{(r)}_{k,s,a}}{\det \lambda I}} \tag{135}$$

for all $k \in \mathbb{N}$ with probability at least $1 - \delta'$. Similarly for any fixed $s' \in \mathcal{S}$, using $\eta_t = 2\mathbf{1}\{s_{k,h} = s, a_{k,h} = a\}((x^{(p)}_k)^\top \theta^{(p)}_{s',s,a} - \mathbf{1}\{s_{k,h+1} = s'\})$, it holds with probability at least $1 - \delta'$ that

$$\|\hat{\theta}^{(p)}_{k,s',s,a} - \theta^{(p)}_{s',s,a}\|_{N^{(p)}_{k,s,a}} \leq \sqrt{\lambda}\|\theta^{(p)}_{s',s,a}\|_2 + \sqrt{\frac{1}{2}\log\frac{1}{\delta'} + \frac{1}{4}\log\frac{\det N^{(p)}_{k,s,a}}{\det \lambda I}} \tag{136}$$

for all episodes k. Finally, for a fixed $s \in \mathcal{S}$ and $h \in [H]$ the sequence

$$\eta_k = \sum_{a \in \mathcal{A}} [\mathbb{P}(s_{k,h} = s, a_{k,h} = a | s_{k,1}, \pi_k) - \mathbf{1}\{s = s_{k,h}, a = a_{k,h}\}] \tag{137}$$

is a martingale difference sequence with respect to $\mathcal{G}_k$, defined as the sigma-field induced by all observations up to including episode $k-1$ and x_k and $s_{k,1}$. All but at most one action has zero probability of occurring (π_k is deterministic) and therefore $\eta_k \in [c, c+1]$ with probability 1 for some c that is measurable in $\mathcal{G}_k$. Hence, $S_t = \sum_{k=1}^{t} \eta_k$ satisfies Assumption 1 with $V_t = t/4$ and ψ_N and $\mathbb{E}L_0 = 1$ (Hoeffding I case in Table 2 of the appendix). This allows us to apply Theorem 2 by Howard et al. (2018) where we choose $h(k) = (1+k)^s\zeta(s)$ with $s = 1.4$ and $\eta = 2$, which gives us (see Eq. (8) and Eq. (9)) specifically) that with probability at least $1 - \delta'$ for all $T \in \mathbb{N}$

$$\sum_{k=1}^{T} \sum_{a \in \mathcal{A}} [\mathbb{P}(s_{k,h} = s, a_{k,h} = a | s_{k,1}, \pi_k) - \mathbf{1}\{s = s_{k,h}, a = a_{k,h}\}] = \sum_{k=1}^{T} \eta_k \leq \sqrt{T(\log\log(T/2) + \log(6/\delta'))}. \tag{138}$$

Setting $\delta' = \frac{\delta}{SA+S^2A+SH}$, all statements above hold for all s', s, a, h jointly using a union bound with probability at least $1-\delta$. This implies that $\mathbb{P}(F) \leq \delta$. $\square$

Using the bounds on the linear parameter estimates, the following lemma derives bounds on the empirical model.

Lemma 25 (Bounds on model parameters)**.** *Outside the failure event F, assuming $\|\theta^{(p)}_{s',s,a}\|_2 \leq \xi_{\theta^{(p)}}$ and $\|\theta^{(r)}_{s,a}\|_2 \leq \xi_{\theta^{(r)}}$ for all $s', s \in \mathcal{S}$ and $a \in \mathcal{A}$ we have*

$$|\hat{r}_k(s,a) - r_k(s,a)| \leq 1 \wedge \alpha_{k,s,a} \|x^{(r)}_k\|_{(N^{(r)}_{k,s,a})^{-1}} \tag{139}$$

$$|\hat{P}_k(s'|s,a) - P_k(s,a)\|_1 \leq 1 \wedge \gamma_{k,s,a} \|x^{(p)}_k\|_{(N^{(p)}_{k,s,a})^{-1}} \tag{140}$$

where

$$\alpha_{k,s,a} = \sqrt{\lambda}\xi_{\theta^{(r)}} + \sqrt{\frac{1}{2}\log\frac{1}{\delta'} + \frac{1}{4}\log\frac{\det N^{(r)}_{k,s,a}}{\det(\lambda I)}} \tag{141}$$

$$\gamma_{k,s,a} = \sqrt{\lambda}\xi_{\theta^{(p)}} + \sqrt{\frac{1}{2}\log\frac{1}{\delta'} + \frac{1}{4}\log\frac{\det N^{(p)}_{k,s,a}}{\det(\lambda I)}}\,. \tag{142}$$

Proof. Since $\hat{r}_k \in [0,1]$ and $r_k \in [0,1]$, we have

$$|\hat{r}_k(s,a) - r_k(s,a)| \leq 1 \wedge |(x^{(r)}_k)^\top \hat{\theta}^{(r)}_{k,s,a} - r_k(s,a)|. \tag{143}$$

The last term can be bounded as

$$|(x^{(r)}_k)^\top \hat{\theta}^{(r)}_{k,s,a} - r_k(s,a)| = |(x^{(r)}_k)^\top (\hat{\theta}^{(r)}_{k,s,a} - \theta^{(r)}_{s,a})| \leq \|x^{(r)}_k\|_{(N^{(r)}_{k,s,a})^{-1}} \|\hat{\theta}^{(r)}_{k,s,a} - \theta^{(r)}_{s,a}\|_{N^{(r)}_{k,s,a}} \tag{144}$$

$$\leq \|x^{(r)}_k\|_{(N^{(r)}_{k,s,a})^{-1}} \left[\sqrt{\lambda}\|\theta^{(r)}_{s,a}\|_2 + \sqrt{\frac{1}{2}\log\frac{1}{\delta'} + \frac{1}{4}\log\frac{\det(N^{(r)}_{k,s,a})}{\det(\lambda I)}}\right] \tag{145}$$

$$\leq \alpha_{k,s,a} \|x^{(r)}_k\|_{(N^{(r)}_{k,s,a})^{-1}} \tag{146}$$

where we first used Hölder's inequality, then the definition of $F^{(r)}$, and finally the assumption $\|\theta^{(r)}_{s,a}\|_2 \leq \xi_{\theta^{(r)}}$. This proves the first inequality. Consider now the second inequality, which we bound analogously as

$$|\hat{P}_k(s'|s,a) - P_k(s'|s,a)| \leq 1 \wedge |(x^{(p)}_k)^\top \hat{\theta}^{(p)}_{k,s',s,a} - P_k(s'|s,a)| \tag{147}$$

$$= 1 \wedge |(x^{(p)}_k)^\top (\hat{\theta}^{(p)}_{k,s',s,a} - \theta^{(p)}_{s',s,a})| \leq 1 \wedge \|x^{(p)}_k\|_{(N^{(p)}_{k,s,a})^{-1}} \|\hat{\theta}^{(p)}_{k,s',s,a} - \theta^{(p)}_{s',s,a}\|_{N^{(p)}_{k,s,a}} \tag{148}$$

$$\leq 1 \wedge \|x^{(p)}_k\|_{(N^{(p)}_{k,s,a})^{-1}} \left[\sqrt{\lambda}\|\theta_{s',s,a}\|_2 + \sqrt{\frac{1}{2}\log\frac{1}{\delta'} + \frac{1}{4}\log\frac{\det(N^{(p)}_{k,s,a})}{\det(\lambda I)}}\right] \tag{149}$$

$$\leq 1 \wedge \gamma_{k,s,a} \|x^{(p)}_k\|_{(N^{(p)}_{k,s,a})^{-1}}\,. \tag{150}$$

$\square$

C.2 Admissibility of guarantees

Lemma 26 (Upper bound admissible)**.** *Outside the failure event F, for all episodes k, $h \in [H]$ and $s, a \in \mathcal{S} \times \mathcal{A}$*

$$Q^\star_{k,h}(s,a) \leq \tilde{Q}_{k,h}(s,a). \tag{151}$$

Proof. Consider a fixed episode k. For $h = H+1$ the claim holds by definition. Assume the claim holds for $h+1$ and consider $\tilde{Q}_{k,h}(s,a) - Q^\star_{k,h}(s,a)$. Since $Q^\star_{k,h} \leq V^{\max}_h$, this quantity is non-negative when $\tilde{Q}_{k,h}(s,a) = V^{\max}_h$.

In the other case

$$\tilde{Q}_{k,h}(s,a) - Q^\star_{k,h}(s,a) \tag{152}$$
$$\geq \hat{r}_k(s,a) + \hat{P}_k(s,a)\tilde{V}_{k,h+1} + \psi_{k,h}(s,a) - P_k(s,a)V^\star_{k,h+1} - r_k(s,a) \tag{153}$$
$$= \hat{r}_k(s,a) - r_k(s,a) + \psi_{k,h}(s,a) + \hat{P}_k(s,a)(\tilde{V}_{k,h+1} - V^\star_{k,h+1}) \tag{154}$$
$$+ (\hat{P}_k(s,a) - P_k(s,a))V^\star_{k,h+1} \tag{155}$$

by induction hypothesis and $\hat{P}_k(s'|s,a) \geq 0$

$$\geq \hat{r}_k(s,a) - r_k(s,a) + \psi_{k,h}(s,a) + (\hat{P}_k(s,a) - P_k(s,a))V^\star_{h+1} \tag{156}$$
$$\geq -|\hat{r}_k(s,a) - r_k(s,a)| + \hat{\psi}_{kh}(s,a) - \sum_{s' \in \mathcal{S}} V^\star_{h+1}(s')|\hat{P}_k(s'|s,a) - P_k(s'|s,a)| \tag{157}$$

by induction hypothesis

$$\geq -|\hat{r}_k(s,a) - r_k(s,a)| + \hat{\psi}_{kh}(s,a) - \sum_{s' \in \mathcal{S}} \tilde{V}_{h+1}(s')|\hat{P}_k(s'|s,a) - P_k(s'|s,a)| \tag{158}$$

using Lemma 25

$$\geq \psi_{k,h}(s,a) - \alpha_{k,s,a}\|x_k^{(r)}\|_{(N^{(r)}_{k,s,a})^{-1}} - \|\tilde{V}_{h+1}\|_1 \gamma_{k,s,a}\|x_k^{(p)}\|_{(N^{(p)}_{k,s,a})^{-1}} = 0\,. \tag{159}$$

□

Using the same technique, we can prove the following result.

Lemma 27 (Lower bound admissible)**.** *Outside the failure event F, for all episodes k, $h \in [H]$ and $s,a \in \mathcal{S} \times \mathcal{A}$*

$$Q^{\pi_k}_{k,h}(s,a) \geq \underset{\sim}{Q}_{k,h}(s,a). \tag{160}$$

Proof. Consider a fixed episode k. For $h = H+1$ the claim holds by definition. Assume the claim holds for $h+1$ and consider $Q^{\pi_k}_{k,h}(s,a) - \underset{\sim}{Q}_{k,h}(s,a)$. Since $Q^{\pi_k}_{k,h} \geq 0$, this quantity is non-negative when $\underset{\sim}{Q}_{k,h}(s,a) = 0$. In the other case

$$Q^{\pi_k}_{k,h}(s,a) - \underset{\sim}{Q}_{k,h}(s,a) \tag{161}$$
$$= P_k(s,a)V^{\pi_k}_{k,h+1} + r_k(s,a) - \hat{r}_k(s,a) - \hat{P}_k(s,a)\underset{\sim}{V}_{k,h+1} + \psi_{k,h}(s,a) \tag{162}$$
$$= r_k(s,a) - \hat{r}_k(s,a) + \psi_{k,h}(s,a) + P_k(s,a)(V^{\pi_k}_{k,h+1} - \underset{\sim}{V}_{k,h+1}) + (P_k - \hat{P}_k)(s,a)\underset{\sim}{V}_{k,h+1} \tag{163}$$

by induction hypothesis and $P_k(s'|s,a) \geq 0$

$$\geq \psi_{k,h}(s,a) - |r_k(s,a) - \hat{r}_k(s,a)| - |(P_k(s,a) - \hat{P}_k(s,a))\underset{\sim}{V}_{k,h+1}| \tag{164}$$

using Lemma 25

$$\geq \psi_{k,h}(s,a) - \alpha_{k,s,a}\|x_k^{(r)}\|_{(N^{(r)}_{k,s,a})^{-1}} - \|\underset{\sim}{V}_{h+1}\|_1 \gamma_{k,s,a}\|x_k^{(p)}\|_{(N^{(p)}_{k,s,a})^{-1}} = 0\,. \tag{165}$$

□

C.3 Cumulative certificate bound

Lemma 28. *Outside the failure event F, the cumulative certificates after T episodes for all T are bounded by*

$$\sum_{k=1}^T \epsilon_k \leq \tilde{O}\left(\sqrt{S^3AH^2T}V_1^{\max}\lambda(\xi^2_{\theta^{(p)}} + \xi^2_{\theta^{(r)}} + d^{(p)} + d^{(r)})\log\frac{\xi^2_{x^{(p)}} + \xi^2_{x^{(r)}}}{\lambda\delta}\right). \tag{166}$$

Proof. Let $\psi_{k,h}(s,a) = \alpha_{k,s,a}\|x_k^{(r)}\|_{(N^{(r)}_{k,s,a})^{-1}} + V^{\max}_{h+1}S\gamma_{k,s,a}\|x_k^{(p)}\|_{(N^{(p)}_{k,s,a})^{-1}}$ We bound the difference between upper and lower Q-estimate as

$$\tilde{Q}_{k,h}(s,a) - \underset{\sim}{Q}_{k,h}(s,a) \tag{167}$$
$$\leq 2\psi_{k,h}(s,a) + \hat{P}_k(s,a)^\top(\tilde{V}_{k,h+1} - \underset{\sim}{V}_{k,h+1}) \tag{168}$$
$$= 2\psi_{k,h}(s,a) + (\hat{P}_k(s,a) - P_k(s,a))^\top(\tilde{V}_{k,h+1} - \underset{\sim}{V}_{k,h+1}) + P_k(s,a)(\tilde{V}_{k,h+1} - \underset{\sim}{V}_{k,h+1}) \tag{169}$$
$$\leq 2\psi_{k,h}(s,a) + V^{\max}_{h+1}\|\hat{P}_k(s,a) - P_k(s,a)\|_1 + P_k(s,a)(\tilde{V}_{k,h+1} - \underset{\sim}{V}_{k,h+1}) \tag{170}$$

and by construction we also can bound $\tilde{Q}_{k,h}(s,a) - \underset{\sim}{Q}_{k,h}(s,a) \leq V^{\max}_h$. Applying both bounds above recursively, we arrive at

$$\epsilon_k = (\tilde{V}_{k,1} - \underset{\sim}{V}_{k,1})(s_{k,1}) = (\tilde{Q}_{k,1} - \underset{\sim}{Q}_{k,1})(s_{k,1}, \pi_k(s_{k,1}, 1)) \tag{171}$$

$$\leq \sum_{s,a \in \mathcal{S}\times\mathcal{A}} \sum_{h=1}^{H} \mathbb{P}(s_h = s, a_h = a | s_{k,1}, \pi_k)[V^{\max}_h \wedge (2\psi_{k,h}(s,a) + V^{\max}_{h+1}\|\hat{P}_k(s,a) - P_k(s,a)\|_1)] \tag{172}$$

$$\leq \sum_{s,a \in \mathcal{S}\times\mathcal{A}} \sum_{h=1}^{H} \mathbb{P}(s_h = s, a_h = a | s_{k,1}, \pi_k)[V^{\max}_h \wedge (2\alpha_{k,s,a}\|x^{(r)}_k\|_{(N^{(r)}_{k,s,a})^{-1}} + 3V^{\max}_{h+1} S\gamma_{k,s,a}\|x^{(p)}_k\|_{(N^{(p)}_{k,s,a})^{-1}})] \tag{173}$$

where we used Lemma 25 in the last step. We are now ready to bound the cumulative certificates after T episodes as

$$\sum_{k=1}^{T} \epsilon_k$$

$$\leq \sum_{k=1}^{T} \sum_{s,a \in \mathcal{S}\times\mathcal{A}} \sum_{h=1}^{H} \mathbb{P}(s_{k,h} = s, a_{k,h} = a | s_{k,1}, \pi_k)[V^{\max}_h \wedge (2\alpha_{k,s,a}\|x^{(r)}_k\|_{(N^{(r)}_{k,s,a})^{-1}} + 3V^{\max}_{h+1} S\gamma_{k,s,a}\|x^{(p)}_k\|_{(N^{(p)}_{k,s,a})^{-1}})] \tag{174}$$

$$\leq \sum_{k=1}^{T} \sum_{h=1}^{H} [V^{\max}_h \wedge (2\alpha_{k,s_{k,h},a_{k,h}}\|x^{(r)}_k\|_{(N^{(r)}_{k,s_{k,h},a_{k,h}})^{-1}} + 3V^{\max}_{h+1} S\gamma_{k,s_{k,h},a_{k,h}}\|x^{(p)}_k\|_{(N^{(p)}_{k,s_{k,h},a_{k,h}})^{-1}})] \tag{175}$$

$$+ \sum_{k=1}^{T} \sum_{s,a \in \mathcal{S}\times\mathcal{A}} \sum_{h=1}^{H} [\mathbb{P}(s_{k,h} = s, a_{k,h} = a | s_{k,1}, \pi_k) - \mathbf{1}\{s = s_{k,h}, a = a_{k,h}\}] V^{\max}_h \tag{176}$$

applying definition of failure event F^O

$$\leq \sum_{k=1}^{T} \sum_{h=1}^{H} [V^{\max}_h \wedge (2\alpha_{k,s_{k,h},a_{k,h}}\|x^{(r)}_k\|_{(N^{(r)}_{k,s_{k,h},a_{k,h}})^{-1}} + 3V^{\max}_{h+1} S\gamma_{k,s_{k,h},a_{k,h}}\|x^{(p)}_k\|_{(N^{(p)}_{k,s_{k,h},a_{k,h}})^{-1}})] \tag{177}$$

$$+ V^{\max}_1 SH\sqrt{T \log \frac{6\log(2T)}{\delta'}} \tag{178}$$

splitting reward and transition terms

$$\leq \sum_{k=1}^{T} \sum_{h=1}^{H} [V^{\max}_h \wedge 2\alpha_{k,s_{k,h},a_{k,h}}\|x^{(r)}_k\|_{(N^{(r)}_{k,s_{k,h},a_{k,h}})^{-1}}] \tag{179}$$

$$+ \sum_{k=1}^{T} \sum_{h=1}^{H} [V^{\max}_h \wedge 3V^{\max}_{h+1} S\gamma_{k,s_{k,h},a_{k,h}}\|x^{(p)}_k\|_{(N^{(p)}_{k,s_{k,h},a_{k,h}})^{-1}}] \tag{180}$$

$$+ V^{\max}_1 SH\sqrt{T \log \frac{6\log(2T)}{\delta'}}. \tag{181}$$

Before bounding the first two terms further, we first derive the following useful inequality using AM-GM inequality which holds for any $s \in \mathcal{A}$ and $s \in \mathcal{S}$

$$\log \frac{\det N^{(r)}_{k,s,a}}{\det(\lambda I)} \leq \log \frac{\left(\frac{1}{d}^{(r)} \operatorname{tr} N^{(r)}_{k,s,a}\right)^{d^{(r)}}}{\lambda^{d^{(r)}}} = d^{(r)} \log \frac{\operatorname{tr} N^{(r)}_{k,s,a}}{d^{(r)}\lambda} \leq d^{(r)} \log \frac{d^{(r)}\lambda + \xi^2_{x^{(r)}}(k-1)H}{d^{(r)}\lambda} \tag{182}$$

where in the last inequality we used the fact that $N^{(r)}_{k,s,a}$ is the sum of λI and at most $H(k-1)$ outer products of feature vectors. Analogously, the following inequality holds for the covariance matrix of the transition features

$$\log \frac{\det N^{(p)}_{k,s,a}}{\det(\lambda I)} \leq d^{(p)} \log \frac{d^{(p)}\lambda + \xi^2_{x^{(p)}}(k-1)H}{d^{(p)}\lambda}. \tag{183}$$

This inequality allows us to upper-bound for $k \leq T$

$$\alpha_{k,s_{k,h},a_{k,h}} =\sqrt{\lambda}\xi_{\theta^{(r)}} + \sqrt{\frac{1}{2}\log\frac{1}{\delta'} + \frac{1}{4}\log\frac{\det N^{(r)}_{k,s_{k,h},a_{k,h}}}{\det(\lambda I)}} \tag{184}$$

$$\leq\sqrt{\lambda}\xi_{\theta^{(r)}} + \sqrt{\frac{1}{2}\log\frac{1}{\delta'} + \frac{1}{4}d^{(r)}\log\frac{d^{(r)}\lambda + \xi^2_{x^{(r)}}(k-1)H}{d^{(r)}\lambda}} \tag{185}$$

$$\leq\sqrt{\lambda}\xi_{\theta^{(r)}} + \sqrt{\frac{1}{2}d^{(r)}\log\frac{d\lambda + \xi^2_{x^{(r)}}HT}{d^{(r)}\lambda\delta'}} \tag{186}$$

using the fact that $\sqrt{a} + \sqrt{b} \leq 2\sqrt{a+b}$ for all $a, b \in \mathbb{R}_+$

$$\leq 2\sqrt{\lambda\xi^2_{\theta^{(r)}} + \frac{1}{2}d^{(r)}\log\frac{d\lambda + \xi^2_{x^{(r)}}HT}{d^{(r)}\lambda\delta'}} \tag{187}$$

$$\leq 2V_1^{\max}\sqrt{\frac{1}{4} + \lambda\xi^2_{\theta^{(r)}} + \frac{1}{2}d^{(r)}\log\frac{d^{(r)}\lambda + \xi^2_{x^{(r)}}HT}{d^{(r)}\lambda\delta'}} =: \alpha_T. \tag{188}$$

Note that the last inequality ensures $\alpha_T \geq V_1^{\max}$. We now use α_T to bound the term in Equation (179)

$$\sum_{k=1}^T\sum_{h=1}^H [V_h^{\max} \wedge 2\alpha_{k,s_{k,h},a_{k,h}}\|x_k^{(r)}\|_{(N^{(r)}_{k,s_{k,h},a_{k,h}})^{-1}}] \leq \sum_{k=1}^T\sum_{h=1}^H [V_h^{\max} \wedge 2\alpha_T\|x_k^{(r)}\|_{(N^{(r)}_{k,s_{k,h},a_{k,h}})^{-1}}] \tag{189}$$

$$\leq 2\alpha_T\sum_{k=1}^T\sum_{h=1}^H [1 \wedge \|x_k^{(r)}\|_{(N^{(r)}_{k,s_{k,h},a_{k,h}})^{-1}}] \tag{190}$$

using Cauchy-Schwarz inequality

$$\leq\sqrt{4\alpha_T^2 TH\sum_{k=1}^T\sum_{h=1}^H [1 \wedge \|x_k^{(r)}\|^2_{(N^{(r)}_{k,s_{k,h},a_{k,h}})^{-1}}]}\,. \tag{191}$$

Leveraging Lemma 30, we can bound the elliptical potential inside the square-root as

$$\sum_{k=1}^T\sum_{h=1}^H [1 \wedge \|x_k^{(r)}\|^2_{(N^{(r)}_{k,s_{k,h},a_{k,h}})^{-1}}] = \sum_{s,a\in\mathcal{S}\times\mathcal{A}}\sum_{h=1}^H\sum_{k=1}^T \mathbf{1}\{s = s_{k,h}, a = a_{k,h}\}[1 \wedge \|x_k^{(r)}\|^2_{(N^{(r)}_{k,s,a})^{-1}}] \tag{192}$$

$$\leq \sum_{s,a\in\mathcal{S}\times\mathcal{A}} H\sum_{k=1}^T [1 \wedge \|x_k^{(r)}\|^2_{(N^{(r)}_{k,s,a})^{-1}}] \leq \sum_{s,a\in\mathcal{S}\times\mathcal{A}} 2H\log\frac{\det N^{(r)}_{k,s,a}}{\det \lambda I} \tag{193}$$

applying Equation (182)

$$\leq 2SAHd^{(r)}\log\frac{d^{(r)}\lambda + \xi^2_{x^{(r)}}HT}{d^{(r)}\lambda} \tag{194}$$

and applying the definition of α_T

$$\leq 2SAH\frac{\alpha_T^2}{2(V_1^{\max})^2} \leq \frac{SAH\alpha_T^2}{(V_1^{\max})^2}\,. \tag{195}$$

We plug this bound back in (191) to get

$$\sum_{k=1}^T\sum_{h=1}^H [V_h^{\max} \wedge 2\alpha_{k,s_{k,h},a_{k,h}}\|x_k^{(r)}\|_{(N^{(r)}_{k,s_{k,h},a_{k,h}})^{-1}}]$$

$$\leq\sqrt{\frac{4\alpha_T^4 SAH^2T}{(V_1^{\max})^2}} = \sqrt{SAH^2T}\frac{2\alpha_T^2}{V_1^{\max}} \tag{196}$$

$$\leq\sqrt{SAH^2T}V_1^{\max}\left[2 + 8\lambda\xi^2_{\theta^{(r)}} + 4d^{(r)}\log\frac{d^{(r)}\lambda + \xi^2_{x^{(r)}}HT}{d^{(r)}\lambda\delta'}\right]. \tag{197}$$

After deriving this upper bound on the term in Equation (179), we bound the term in Equation (180) in similar fashion. We start with an upper bound on $S\gamma_{k,s_{k,h},a_{k,h}}$ which holds for $k \leq T$:

$$S\gamma_{k,s_{k,h},a_{k,h}} \leq \sqrt{1 + 4\lambda S^2 \xi^2_{\theta^{(p)}} + 2S^2 d^{(p)} \log \frac{d^{(p)}\lambda + \xi^2_{x^{(p)}} HT}{d^{(p)}\lambda\delta'}} =: \gamma_T \, , \tag{198}$$

which is by construction at least 1. We now use this definition to bound as above

$$\sum_{k=1}^{T}\sum_{h=1}^{H}[V_h^{\max} \wedge 3V_{h+1}^{\max} S\gamma_{k,s_{k,h},a_{k,h}} \|x_k^{(p)}\|_{(N^{(p)}_{k,s_{k,h},a_{k,h}})^{-1}}] \leq 3V_1^{\max}\gamma_T \sum_{k=1}^{T}\sum_{h=1}^{H}[1 \wedge \|x_k^{(p)}\|_{(N^{(p)}_{k,s_{k,h},a_{k,h}})^{-1}}] \tag{199}$$

$$\leq 3V_1^{\max}\gamma_T \sqrt{TH\sum_{k=1}^{T}\sum_{h=1}^{H}[1 \wedge \|x_k^{(p)}\|^2_{(N^{(p)}_{k,s_{k,h},a_{k,h}})^{-1}}]} \leq 3V_1^{\max}\gamma_T \sqrt{TH 2SAHd^{(p)} \log \frac{d^{(p)}\lambda + \xi^2_{x^{(p)}}HT}{d^{(p)}\lambda}} \tag{200}$$

$$\leq 3V_1^{\max}\gamma_T \sqrt{2SAH^2T\frac{\gamma_T^2}{2S^2}} \leq \sqrt{S^3AH^2T}V_1^{\max}\left[3 + 12\lambda\xi^2_{\theta^{(p)}} + 6d^{(p)}\log\frac{d^{(p)}\lambda + \xi^2_{x^{(p)}}HT}{d^{(p)}\lambda\delta'}\right]. \tag{201}$$

Combining (181), (197) and (201), the cumulative certificates after T episodes are bounded by

$$\sum_{k=1}^{T}\epsilon_k \leq \sqrt{S^3AH^2T}V_1^{\max}\left[14 + 12\lambda(\xi^2_{\theta^{(p)}} + \xi^2_{\theta^{(r)}}) + 6(d^{(p)} + d^{(r)})\log\frac{(d^{(p)} + d^{(r)})\lambda + (\xi^2_{x^{(p)}} + \xi^2_{x^{(r)}})HT}{(d^{(p)} \wedge d^{(r)})\lambda\delta'}\right] \tag{202}$$

$$+ V_1^{\max}SH\sqrt{T\log\frac{6\log(2T)}{\delta'}} \tag{203}$$

$$= \tilde{O}\left(\sqrt{S^3AH^2T}V_1^{\max}\lambda(\xi^2_{\theta^{(p)}} + \xi^2_{\theta^{(r)}} + d^{(p)} + d^{(r)})\log\frac{\xi^2_{x^{(p)}} + \xi^2_{x^{(r)}}}{\lambda\delta}\right). \tag{204}$$

□

C.4 Proof of Theorem 5

We are now ready to assemble the arguments above and prove the cumulative IPOC bound for Algorithm 2:

Proof. By Lemma 24, the failure event F has probability at most δ. Outside the failure event, for every episode k, the upper and lower Q-value estimates are valid upper bounds on the optimal Q-function and lower bounds on the Q-function of the current policy π_k, respectively (Lemmas 26 and 27). Further, Lemma 28 shows that the cumulative certificates grow at the desired rate

$$\tilde{O}\left(\sqrt{S^3AH^2T}V_1^{\max}\lambda(\xi^2_{\theta^{(p)}} + \xi^2_{\theta^{(r)}} + d^{(p)} + d^{(r)})\log\frac{\xi^2_{x^{(p)}} + \xi^2_{x^{(r)}}}{\lambda\delta}\right). \tag{205}$$

□

C.5 Technical Lemmas

We now state two existing technical lemmas used in our proof.

Lemma 29 (Elliptical confidence sets; Theorem 20.1 in Lattimore and Czepesvari (2018))**.** *Let $\lambda > 0$, $\theta \in \mathbb{R}^d$ and $(r_i)_{i\in\mathbb{N}}$ and $(x_i)_{i\in\mathbb{N}}$ random processes adapted to a filtration $\mathcal{F}_i$ so that $r_i - x_i^\top\theta$ are conditionally 1-sub-Gaussian. Then with probability at least $1-\delta$ for all $k \in \mathbb{N}$*

$$\|\theta - \tilde{\theta}_k\|_{N_k(\lambda)} \leq \sqrt{\lambda}\|\theta\|_2 + \sqrt{2\log\frac{1}{\delta} + \log\frac{\det(N_k(\lambda))}{\det(\lambda I)}} \tag{206}$$

where $N_k(\lambda) = \lambda I + \sum_{i=1}^{k} x_i x_i^\top$ is the covariance matrix and $\tilde{\theta}_k = N_k(\lambda)^{-1}\sum_{i=1}^{k} r_i x_i$ is the least-squares estimate.

Lemma 30 (Elliptical potential; Lemma 19.1 in Lattimore and Czepesvari (2018))**.** *Let $x_1, \dots, x_n \in \mathbb{R}^d$ with $L \geq \max_i \|x_i\|_2$ and $N_i = N_0 + \sum_{j=1}^i x_j x_j^\top$ with N_0 being psd. Then*

$$\sum_{i=1}^n 1 \wedge \|x_i\|_{N_{i-1}^{-1}} \leq 2 \log \frac{\det N_n}{\det N_0} \leq 2d \log \frac{\operatorname{tr}(N_0) + nL^2}{d \det(N_0)^{1/d}}. \tag{207}$$

D Experimental Setup

The first benchmark problem is a contextual bandit ($S = H = 1$) with $A = 40$ actions and $d^{(r)} = 10$ dimensional context. We sample the reward parameters independently for all $s = 1$, $a \in \mathcal{A}$ and $i \in [d^{(r)}]$ as

$$\theta^{(r)}_{i,s,a} = X_{i,s,a} Y_{i,s,a}, \qquad X_{i,s,a} \sim \text{Bernoulli}(0.9), \quad Y_{i,s,a} \sim \text{Unif}(0, 1). \tag{208}$$

The context in each episode is sampled from a Dirichlet distribution with parameter $\alpha \in \mathbb{R}^{d^{(r)}}$ where $\alpha_i = 0.7$ for $i \leq 7$ and $\alpha_i = 0.01$ for $i \geq 10$. This choice was made to simulate both frequent as well as a few rare context dimensions. The `ORLC-SI` algorithm was run for 8 million episodes and we changed context, certificate and policy only every 1000 episodes for faster experimentation. Figure 2 shows certificates and optimality gaps of a representative run. Note that we sub-sampled the number of episodes shown for clearer visualization.

The second benchmark problem is an MDP with side information. It has $S = 10$ states, $A = 40$ actions, horizon of $H = 5$, reward context dimension $d^{(r)} = 10$, and transition context dimension $d^{(p)} = 1$. The transition context $x^{(p)}_k$ is always constant 1. We sample the reward parameters independently for all $s \in \mathcal{S}$, $a \in \mathcal{A}$ and $i \in [d^{(r)}]$ as

$$\theta^{(r)}_{i,s,a} = X_{i,s,a} Y_{i,s,a}, \quad X_{i,s,a} \sim \text{Bernoulli}(0.5), \quad Y_{i,s,a} \sim \text{Unif}(0, 1). \tag{209}$$

and the transition kernel for each $s \in \mathcal{S}, a \in \mathcal{A}$ as

$$P(s, a) = \theta^{(p)}_{s,a} \sim \text{Dirichlet}(\alpha^{(p)}) \tag{210}$$

where $\alpha^{(p)} \in \mathbb{R}^S$ with $\alpha^{(p)}_i = 0.3$ for $i \in [S]$. The reward context is again sampled from a Dirichlet distribution with parameter $\alpha^{(r)} \in \mathbb{R}^{d^{(r)}}$ where $\alpha^{(r)}_i = 0.01$ for $i \leq 4$ in the first 2 million episodes and all other times $\alpha^{(r)}_i = 0.7$. This shift in context distribution after 2 million episodes simulates rare contexts becoming more frequent.

In both experiments, we use a slightly modified version of `ORLC-SI` listed in Algorithm 3 which computes the optimistic and pessimistic Q estimates $\tilde{Q}$ and $\underset{\sim}{Q}$ using subroutine ProbEstNorm in Algorithm 4. While the original version of `ORLC-SI` does not leverage that the true transition kernel $P_k(s, a)$ has total mass 1, this version adds this as a constraint (see Lemma 31) similar to Abbasi-Yadkori and Neu (2014). This change yielded improved estimates empirically in our simulation. Note that this does not harm the theoretical properties. One can show the same cumulative IPOC bound for Algorithm 3 by slightly modifying the proof for Algorithm 2.

Lemma 31. *Let $\hat{p} \in [0, 1]^d$, $\psi \geq 0$ and $v \in \mathbb{R}^d$ and define $\mathcal{P}_{\hat{p}} = \{p \in [0, 1]^d \ : \ \leq \hat{p} - \psi \mathbf{1}_d \leq p \leq \hat{p} + \psi \mathbf{1}_d \wedge \|p\|_1 = 1\}$. Then, as long as $\mathcal{P}_{\hat{p}} \neq \emptyset$, the value returned by Algorithm 4 satisfies*

$$\text{ProbEstNorm}(\hat{p}, \psi, v) = \max_{p \in \mathcal{P}_{\hat{p}}} p^\top v \tag{211}$$

$$- \text{ProbEstNorm}(\hat{p}, \psi, -v) = \min_{p \in \mathcal{P}_{\hat{p}}} p^\top v \tag{212}$$

and for any two $p, \tilde{p} \in \mathcal{P}_{\hat{p}}$ it holds that $|p^\top v - \tilde{p}^\top v| \leq \|v\|_1 \|p - \tilde{p}\|_\infty = 2\psi \|v\|_1$.

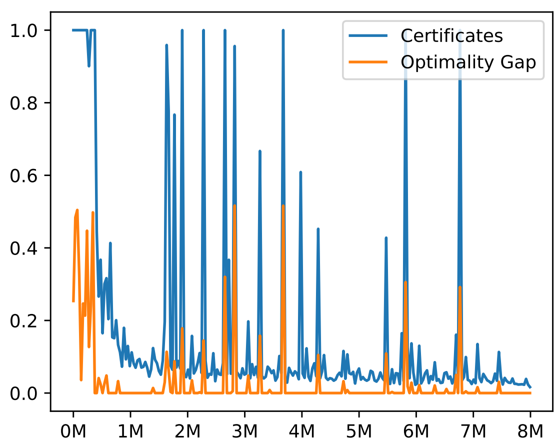

Figure 2: Results of `ORLC-SI` for 8M episodes on a linear contextual bandit problem; certificates are shown in blue and the true (unobserved) optimality gap in orange for increasing number of episodes.

Algorithm 3: `ORLC-SI` algorithm with probability mass constraints

Input: failure prob. $\delta \in (0, 1]$, regularizer $\lambda > 0$

$\forall s, s' \in \mathcal{S}, a \in \mathcal{A}, h \in [H]$:

$N^{(p)}_{s,a} \leftarrow \lambda I_{d^{(p)} \times d^{(p)}}$; $\quad N^{(r)}_{s,a} \leftarrow \lambda I_{d^{(r)} \times d^{(r)}}$;

$M^{(r)}_{s,a} \leftarrow \vec{0}_{d^{(r)}}$; $\quad M^{(p)}_{s',s,a} \leftarrow \vec{0}_{d^{(p)}}$;

$\tilde{V}_{H+1} \leftarrow \vec{0}_S \quad \underset{\sim}{V}_{H+1} \leftarrow \vec{0}_S \quad V^{\max}_h \leftarrow (H - h + 1)$;

$\xi_{\theta^{(r)}} \leftarrow \sqrt{d}$; $\xi_{\theta^{(p)}} \leftarrow \sqrt{d} \quad \delta' \leftarrow \frac{\delta}{S(SA+A+H)}$;

$\phi(N, x, \xi) := \left[\sqrt{\lambda}\xi + \sqrt{\frac{1}{2} \ln \frac{1}{\delta'} + \frac{1}{4} \ln \frac{\det N}{\det(\lambda I)}}\right] \|x\|_{N^{-1}}$;

for $k = 1, 2, 3, \dots$ **do**

 Observe current contexts $x^{(r)}_k$ and $x^{(p)}_k$;

 `/* estimate model with least squares */`

 for $s, s' \in \mathcal{S}, a \in \mathcal{A}$ **do**

  $\hat{\theta}^{(r)}_{s,a} \leftarrow (N^{(r)}_{s,a})^{-1} M^{(r)}_{s,a}$;

  $\hat{r}(s, a) \leftarrow 0 \vee (x^{(r)}_k)^\top \hat{\theta}^{(r)}_{s,a} \wedge 1$;

  $\hat{\theta}^{(p)}_{s',s,a} \leftarrow (N^{(p)}_{s,a})^{-1} M^{(p)}_{s',s,a}$;

  $\hat{P}(s'|s, a) \leftarrow 0 \vee (x^{(p)}_k)^\top \hat{\theta}^{(p)}_{s',s,a} \wedge 1$;

 `/* optimistic planning */`

 for $h = H$ **to** 1 ***and*** $s \in \mathcal{S}$ **do**

  for $a \in \mathcal{A}$ **do**

   $\tilde{\psi}_h(s, a) \leftarrow \phi(N^{(r)}_{s,a}, x^{(r)}_k, \xi_{\theta^{(r)}})$;

   $\underset{\sim}{\psi}_h(s, a) \leftarrow \phi(N^{(r)}_{s,a}, x^{(r)}_k, \xi_{\theta^{(r)}})$;

   $\tilde{Q}_h(s, a) \leftarrow \hat{r}(s, a) + \text{ProbEstNorm}(\hat{P}(s, a), \phi(N^{(p)}_{s,a}, x^{(p)}_k, \xi_{\theta^{(p)}}), \tilde{V}_{h+1}) + \tilde{\psi}_h(s, a)$;

   $\underset{\sim}{Q}_h(s, a) \leftarrow \hat{r}(s, a) - \text{ProbEstNorm}(\hat{P}(s, a), \phi(N^{(p)}_{s,a}, x^{(p)}_k, \xi_{\theta^{(p)}}), -\underset{\sim}{V}_{h+1}) - \underset{\sim}{\psi}_h(s, a)$;

   `// clip values`

   $\tilde{Q}_h(s, a) \leftarrow 0 \vee \tilde{Q}_h(s, a) \wedge V^{\max}_h$;

   $\underset{\sim}{Q}_h(s, a) \leftarrow 0 \vee \underset{\sim}{Q}_h(s, a) \wedge V^{\max}_h$;

  $\pi_k(s, h) \leftarrow \operatorname{argmax}_a \tilde{Q}_h(s, a)$;

  $\tilde{V}_h(s) \leftarrow \tilde{Q}_h(s, \pi_k(s, t))$;

  $\underset{\sim}{V}_h(s) \leftarrow \underset{\sim}{Q}_h(s, \pi_k(s, t))$;

 `/* Execute policy for one episode */`

 $s_{k,1} \sim P_0$;

 $\epsilon_k \leftarrow \tilde{V}_1(s_{k,1}) - \underset{\sim}{V}_1(s_{k,1})$;

 output policy π_k **with certificate** ϵ_k;

 for $h = 1$ **to** H **do**

  $a_{k,h} \leftarrow \pi_k(s_{k,h}, h)$;

  $r_{k,h} \sim P_R(s_{k,h}, a_{k,h})$; $\quad s_{k,h+1} \sim P(s_{k,h}, a_{k,h})$;

  `// Update statistics`

  $N^{(p)}_{s_{k,h},a_{k,h}} \leftarrow N^{(p)}_{s_{k,h},a_{k,h}} + x^{(p)}_k (x^{(p)}_k)^\top$;

  $N^{(r)}_{s_{k,h},a_{k,h}} \leftarrow N^{(r)}_{s_{k,h},a_{k,h}} + x^{(r)}_k (x^{(r)}_k)^\top$;

  $M^{(p)}_{s_{k,h+1},s_{k,h},a_{k,h}} \leftarrow M^{(p)}_{s_{k,h+1},s_{k,h},a_{k,h}} + x^{(p)}_k$;

  $M^{(p)}_{s_{k,h},a_{k,h}} \leftarrow M^{(p)}_{s_{k,h},a_{k,h}} + x^{(p)}_k$;

Algorithm 4: ProbEstNorm($\hat{p}$, ψ, v) function to compute normalized estimated expectation of v

Input : estimated probability vector $\hat{p} \in [0, 1]^S$
Input : confidence width $\psi \in \mathbb{R}_+$
Input : value vector $v \in \mathbb{R}^S$

```
Compute sorting σ of v so that v_{σ_i} ≥ v_{σ_j} for all i ≤ j;
p ← p̂ − ψ ∨ 0;
m ← p^⊤ 1;
r ← 0;
for i ∈ [S] do
    s ← m ∧ ((p̂_{σ_i} + ψ ∧ 1) − p_{σ_i});
    m ← m − s;
    r ← r + v_{σ_i}(p_{σ_i} + s);
```

Return: r