# Policy Certificates: Towards Accountable Reinforcement Learning

**Christoph Dann** [1] **Lihong Li** [2] **Wei Wei** [2] **Emma Brunskill** [3]

## Abstract

The performance of a reinforcement learning algorithm can vary drastically during learning because of exploration. Existing algorithms provide little information about the quality of their current policy before executing it, and thus have limited use in high-stakes applications like healthcare. We address this lack of *accountability* by proposing that algorithms output *policy certificates*. These certificates bound the sub-optimality and return of the policy in the next episode, allowing humans to intervene when the certified quality is not satisfactory. We further introduce two new algorithms with certificates and present a new framework for theoretical analysis that guarantees the quality of their policies and certificates. For tabular MDPs, we show that computing certificates can even improve the sample-efficiency of optimism-based exploration. As a result, one of our algorithms is the first to achieve minimax-optimal PAC bounds up to lower-order terms, and this algorithm also matches (and in some settings slightly improves upon) existing minimax regret bounds.

## 1. Introduction

There is increasing excitement around applications of machine learning, but also growing awareness and concerns about fairness, accountability and transparency. Recent research aims to address these concerns but most work focuses on supervised learning and only few results (Jabbari et al., 2016; Joseph et al., 2016; Kannan et al., 2017; Raghavan et al., 2018) exist on reinforcement learning (RL).

One challenge when applying 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 common, e.g., 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 online during learning.

To address this lack of accountability, we propose that RL algorithms output *policy certificates* in episodic RL. Policy certificates consist of (1) a confidence interval of the algorithm's expected sum of rewards (return) in the next episode (policy return certificates) and (2) a bound on how far from the optimal return the performance can be (policy optimality certificates). Certificates make the policy's performance more transparent and accountable, and allow designers to intervene if necessary. For example, in medical applications, one would need to intervene unless the policy achieves a certain minimum treatment outcome; in financial applications, policy optimality certificates can be used to assess the potential loss when learning a trading strategy. In addition to accountability, we also want RL algorithms to be sample-efficient and quickly achieve good performance. To formally quantify accountability and sample-efficiency of an algorithm, we introduce a new framework for theoretical analysis called IPOC. IPOC bounds guarantee that certificates indeed bound the algorithm's expected performance in an episode, and prescribe the rate at which the algorithm's policy and certificates improve with more data. IPOC is stronger than other frameworks like regret (Jaksch et al., 2010), PAC (Kakade, 2003) and Uniform-PAC (Dann et al., 2017), that only guarantee the cumulative performance of the algorithm, but do not provide bounds for *individual* episodes during learning. IPOC also provides stronger bounds and more nuanced guarantees on per episode performance than KWIK (Li et al., 2008).

A natural way to create accountable and sample-efficient RL algorithms is to combine existing sample-efficient algorithms with off-policy policy evaluation approaches to estimate the return (expected sum of rewards) of the algorithm's policy before each episode. Existing policy evaluation approaches estimate the return of a fixed policy from a batch of data (e.g., Thomas et al., 2015b; Jiang & Li, 2016; Thomas & Brunskill, 2016). They provide little to no guarantees when the policy is not fixed but computed from

[1]Carnegie Mellon University [2]Google Research [3]Stanford University. Correspondence to: Christoph Dann <cdann@cdann.net>.

*Proceedings of the $36^{th}$ International Conference on Machine Learning*, Long Beach, California, PMLR 97, 2019. 

that same batch of data, as is here the case. They also do not reason about the return of the unknown optimal policy which is necessary for providing policy optimality certificates. We found that by focusing on optimism-in-the-face-of-uncertainty (OFU) based RL algorithms for updating the policy and model-based policy evaluation techniques for estimating the policy returns, we can create sample-efficient algorithms that compute policy certificates on both the current policy's return and its difference to the optimal return. The main insight is that OFU algorithms compute an upper confidence bound on the optimal return from an empirical model when updating the policy. Model-based policy evaluation can leverage the same empirical model to compute a confidence interval on the policy return, even when the policy depends on the data. We illustrate this approach with new algorithms for two different episodic settings.

Perhaps surprisingly, we show that in tabular Markov decision processes (MDPs) it can be beneficial to explicitly leverage the combination of OFU-based policy optimization and model-based policy evaluation to improve either component. Specifically, computing the certificates can directly improve the underlying OFU approach and knowing that the policy converges to the optimal policy at a certain rate improves the accuracy of policy return certificates. As a result, the guarantees for our new algorithm improve state-of-the-art regret and PAC bounds in problems with large horizons and are minimax-optimal up to lower-order terms.

The second setting we consider are finite MDPs with linear side information (context) (Abbasi-Yadkori & Neu, 2014; Hallak et al., 2015; Modi et al., 2018), which 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.

To summarize, We make the following main contributions:

1. We introduce policy certificates and the IPOC framework for evaluating RL algorithms with certificates. Similar to existing frameworks like PAC, it provides formal requirements to be satisfied by the algorithm, here requiring the algorithm to be an efficient learner and to quantify its performance online through policy certificates.
2. We provide a new RL algorithm for finite, episodic MDPs that satisfies this definition, and show that it has stronger, minimax regret and PAC guarantees than prior work. Formally, our sample complexity bound is $\tilde{O}(SAH^2/\epsilon^2 + S^2AH^3/\epsilon)$ vs. prior $\tilde{O}(SAH^4/\epsilon^2 + S^2AH^3/\epsilon)$ (Dann et al., 2017), and our regret bound $\tilde{O}(\sqrt{SAH^2T} + S^2AH^3)$ improves prior work (Azar et al., 2017) since it has minimax rate up to log-terms in the dominant term even for long horizons $H > SA$.
3. We introduce a new RL algorithm for finite, episodic MDPs with linear side information that has a cumulative IPOC bound, which is tighter than past results (Abbasi-Yadkori & Neu, 2014) by a factor of $\sqrt{SAH}$.

## 2. Setting and Notation

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 more breadly, we focus on finite MDPs with linear side information (Modi et al., 2018; Hallak et al., 2015; Abbasi-Yadkori & 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.

**Tabular MDPs** The agent interacts with the MDP in episodes indexed by $k$. Each episode is a sequence $(s_{k,1}, a_{k,1}, r_{k,1}, \ldots, s_{k,H}, a_{k,H}, r_{k,H})$ of $H$ states $s_{k,h} \in \mathcal{S}$, actions $a_{k,h} \in \mathcal{A}$ and scalar rewards $r_{k,h} \in [0,1]$. For notational simplicity, we assume that the initial state $s_{k,1}$ is deterministic. The actions are taken as prescribed by the agent's policy $\pi_k$ and we here focus on deterministic time-dependent policies, i.e., $a_{k,h} = \pi_k(s_{k,h}, h)$ for all time steps $h \in [H] := \{1, 2, \ldots H\}$. The successor states and rewards are sampled from the MDP as $s_{k,h+1} \sim P(s_{k,h}, a_{k,h})$ and $r_{k,h} \sim P_R(s_{k,h}, a_{k,h})$. In tabular MDPs the size of the state space $S = |\mathcal{S}|$ and action space $A = |\mathcal{A}|$ are finite.

**Finite MDPs with linear side information.** We assume that state- and action-space are finite as in tabular MDPs, 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.,

$\|\theta^{(r)}_{s,a}\|_2 \leq \xi_{\theta^{(r)}}$ and $\|\theta^{(p)}_{s',s,a}\|_2 \leq \xi_{\theta^{(p)}}$ as well as bounded contexts $\|x^{(r)}_k\|_2 \leq \xi_{x^{(r)}}$ and $\|x^{(p)}_k\|_2 \leq \xi_{x^{(p)}}$. We allow $x^{(r)}_k$ and $x^{(p)}_k$ to be different, and use $x_k$ to denote $(x^{(r)}_k, x^{(p)}_k)$ in the following. Note that our framework and algorithms can handle adversarially chosen contexts.

**Return and optimality gap.** The quality of a policy $\pi$ in any episode $k$ is evaluated by the *total expected reward* or *return*: $\rho_k(\pi) := \mathbb{E}\left[\sum_{h=1}^{H} r_{k,h} \middle| a_{k,1:H} \sim \pi\right]$, where this notation means that all actions in the episode are taken as prescribed by a policy $\pi$. Optimal policy and return $\rho^\star_k = \max_\pi \rho_k(\pi)$ may depend on the episode's contexts. The difference of achieved and optimal return is called *optimality gap* $\Delta_k = \rho^\star_k - \rho_k(\pi_k)$ for each episode $k$ where $\pi_k$ is the algorithm's policy in that episode.

**Additional notation.** We denote the largest possible optimality gap by $\Delta_{\max} = H$, and the value functions of $\pi$ in episode $k$ by $Q^{\pi_k}_h(s,a) = \mathbb{E}[\sum_{t=h}^{H} r_{k,t} | a_{k,h} = a, a_{k,h+1:H} \sim \pi]$ and $V^{\pi_k}_h(s) = Q^{\pi_k}_h(s, \pi(s,h))$. Optimal versions are marked by superscript $\star$ and subscripts are omitted when unambiguous. We treat $P(s,a)$ as a linear operator, that is, $P(s,a)f = \sum_{s' \in \mathcal{S}} P(s'|s,a) f(s')$ for any $f : \mathcal{S} \to \mathbb{R}$. We also use $\sigma_q(f) = \sqrt{q(f - qf)^2}$ for the standard deviation of $f$ with respect to a state distribution $q$ and $V^{\max}_h = (H - h + 1)$ for all $h \in [H]$. We also use the common short hand notation $a \vee b = \max\{a, b\}$ and $a \wedge b = \min\{a, b\}$ as well as $\tilde{O}(f) = O(f \cdot \mathrm{poly}(\log(f)))$.

## 3. The IPOC Framework

During execution, the optimality gaps $\Delta_k$ are hidden and 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 whether the algorithm is playing a good or potentially bad policy. We introduce a new learning framework that mitigates this limitation. This framework forces the algorithm to output its current policy $\pi_k$ as well as certificates $\epsilon_k \in \mathbb{R}_+$ and $\mathcal{I}_k \subseteq \mathbb{R}$ before each episode $k$. The *return certificate* $\mathcal{I}_k$ is a confidence interval on the return of the policy, while the *optimality certificate* $\epsilon_k$ informs the user how sub-optimal the policy can be for the current context, i.e., $\epsilon_k \geq \Delta_k$. Certificates allow one to intervene if needed. For example, in automated customer services, one might reduce the service price in episode $k$ if certificate $\epsilon_k$ is above a certain threshold, since the quality of the provided service cannot be guaranteed. When there is no context, an optimality certificate upper bounds the sub-optimality of the current policy in any episode which makes algorithms anytime interruptable (Zilberstein & 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 the current policy $\pi_k$, a* return certificate $\mathcal{I}_k \subseteq \mathbb{R}$ *and an* optimality certificate $\epsilon_k$ *with $\epsilon_k \geq |\mathcal{I}_k|$ before each episode $k$ (after observing the contexts) so that with probability at least $1 - \delta$:*

1. *all return certificates contain the return policy $\pi_k$ played in episode $k$ and all optimality certificates are upper bounds on the sub-optimality of $\pi_k$, i.e., $\forall\, k \in \mathbb{N}$ : $\epsilon_k \geq \Delta_k$ and $\rho_k(\pi_k) \in \mathcal{I}_k$ ; 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 $\epsilon$, 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 $\epsilon_k = \Delta_{\max}$ and $\mathcal{I}_k = [0, \Delta_{\max}]$, but condition 2 prohibits this by controlling the size of $\epsilon_k$ (and therefore the size of $|\mathcal{I}_k| \leq \epsilon_k$). Condition 2a bounds the cumulative sum of optimality certificates (similar to regret bounds), and condition 2b bounds the size of the superlevel sets of $\epsilon_k$ (similar to PAC bounds). We allow both alternatives as condition 2b is stronger but one sometimes can only prove condition 2a (see Appendix D[1]). An IPOC bound controls simultaneously the quality of certificates (how big $\epsilon_k - \Delta_k$ and $|\mathcal{I}_k|$ are) as well as the optimality gaps $\Delta_k$ 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. Note that the condition $\epsilon_k \geq |\mathcal{I}_k|$ in Definition 1 is natural as any upper bound on $\rho^\star_k$ is also an upper bound on $\rho_k(\pi_k)$ and is made for notational convenience.

We would like to emphasize that we provide certificates on the return, the *expected* sum of rewards, in the next episode. Due to the stochasticity in the environment, one in general cannot hope to accurately predict the sum of rewards directly. Since return is the default optimization criteria in RL, certificates for it are a natural starting point and relevant in many scenarios. Nonetheless, certificates for other properties of the sum-of-reward distribution of a policy are an interesting direction for future work. For example, one might want certificates on properties that take into account the variability of the sum of rewards (e.g., conditional value at risk) in high-stakes applications which are often the objective in risk-sensitive RL.

[1] See full version of this paper at https://arxiv.org/abs/1811.03056.

### 3.1. Relation to Existing Frameworks

Unlike IPOC, existing frameworks for RL only guarantee sample-efficiency of the algorithm over multiple episodes and do not provide performance bounds for single episodes during learning. The common existing frameworks are:

- *Mistake-style PAC bounds* (Strehl et al., 2006; 2009; Szita & Szepesvári, 2010; Lattimore & Hutter, 2012; Dann & Brunskill, 2015) bound the number of $\epsilon$-mistakes, that is, the size of the set $\{k \in \mathbb{N} : \Delta_k > \epsilon\}$ with high probability, but do not tell us when mistakes happen. The same is true for the stronger Uniform-PAC bounds (Dann et al., 2017) which hold for all $\epsilon$ jointly.
- *Supervised-learning style PAC bounds* (Kearns & Singh, 2002; Jiang et al., 2017; Dann et al., 2018) ensure that the algorithm outputs an $\epsilon$-optimal policy for a given $\epsilon$, i.e., they ensure $\Delta_k \leq \epsilon$ for $k$ greater than the bound. Yet, they need to know $\epsilon$ ahead of time and tell us nothing about $\Delta_k$ during learning (for $k$ 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 $\Delta_k$ because it does not reveal which optimality gaps are small.

We show that mistake IPOC bounds are stronger than any of the above guarantees, i.e., they imply Uniform PAC, PAC, and regret bounds. Cumulative IPOC bounds are slightly weaker but still imply regret bounds. Both versions of IPOC also ensure that the algorithm is anytime interruptable, i.e., it can be used to find better and better policies that have small $\Delta_k$ with high probability $1-\delta$. That means IPOC bounds imply supervised-learning style PAC bounds for all $\epsilon$ jointly. These claims are formalized as follows:

**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 $\epsilon$, it outputs a certificate $\epsilon_k \leq \epsilon$ within*

$$\sum_{p=0}^{N} \frac{C_p(W, \delta)^p (N+1)^{p+1}}{\epsilon^{p+1}} \tag{1}$$

*episodes. Hence, for settings without context, the algorithm outputs an $\epsilon$-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 $\Delta_k \geq \epsilon$ is at most $F(W, \epsilon, \delta)$ for all $\epsilon > 0$;*
2. *with probability $\geq 1-\delta$ for all $\epsilon$, it outputs a certificate $\epsilon_k \leq \epsilon$ within $F(W, \epsilon, \delta) + 1$ episodes. For settings without context, that means the algorithm outputs an $\epsilon$-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 $C_p(W, \delta) \geq 1$ and constants $N, n \in \mathbb{N}$, it also has a cumulative IPOC bound of order*

$$\tilde{O}\left(\sum_{p=1}^{N} C_p(W, \delta)^{1/p} T^{\frac{p-1}{p}} \operatorname{polylog}(\Delta_{\max}, \tilde{C}(W, \delta), T)\right).$$

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

Our IPOC framework is similar to KWIK (Li et al., 2008), in that the algorithm is required to declare how well it will perform. Hower, KWIK only requires an algorithm to declare whether the output will perform better than a single pre-specified input threshold. Existing KWIK for RL methods only provide such a binary classification, and have less strong learning guarantees. In a sense IPOC is a generalization of KWIK.

## 4. Algorithms with Policy Certificates

A natural path to obtain RL algorithms with IPOC bounds is to combine existing provably efficient online RL algorithms with an off-policy policy evaluation method to compute a confidence interval on the online RL algorithm's policy for the current episode. This yields policy return certificates, but not necessarily policy optimality certificates – bounds on the difference of the optimal and current policy's return. Estimating the optimal return using off-policy evaluation algorithms in order to compute optimality certificates would require a significant computational burden, e.g. evaluating all (exponentially many) policies.

However optimism in the face of uncertainty (OFU) algorithms can be modified to provide both policy return certificates and optimality certificates without the need for a separate off-policy policy optimization step. Specifically, we here consider OFU algorithms that maintain an upper confidence bound (for a potentially changing confidence level) on the optimal value function $Q^\star_{k,h}$ and therefore optimal return $\rho^\star_k$. This bound is also an upper bound on the return of the current policy which is chosen to maximize this bound. Many OFU methods explicitly maintain a confidence set of the MDP model to compute the upper confidence bound on $Q^\star_{k,h}$. These same confidence sets of the model can be used to compute a lower bound on the value function of the current policy. In doing so, OFU algo-

rithms can be modified with little computational overhead to provide policy return and optimality certificates.

For these reasons, we focus on OFU methods, introducing two new algorithms with policy certificates, one for tabular MDPs and and one for the more general MDPs with linear side information setting. Both approaches have a similar structure, but leverage different confidence sets and model estimators. In the first case, we show that maintaining lower bounds on the current policy's value has significant benefits beyond enabling policy certificates: lower bounds help us to derive a tighter bound on our uncertainty over the range of future values. Thus we are able to provide the strongest, to our knowledge, PAC and regret bounds for tabular MDPs. It remains an intriguing but non-trivial question if we can create confidence sets that leverage explicit upper and lower bounds for the linear side information setting.

### 4.1. Tabular MDPs

We present the ORLC (optimistic RL with certificates) Algorithm shown in Algorithm 1 (see the appendix for a version with empirically tighter confidence bounds but same theoretical guarantees). It shares similar structure with recent OFU algorithms like UBEV (Dann et al., 2017) and UCBVI-BF (Azar et al., 2017) but has some significant differences highlighted in red. Before each episode $k$, Algorithm 1 computes an optimistic estimate $\tilde{Q}_{k,h}$ of $Q^\star_h$ in Line 10 by dynamic programming on the empirical model $(\hat{P}_k, \hat{r}_k)$ with confidence intervals $\psi_{k,h}$. Importantly, it also computes $\underline{Q}_{k,h}$, a pessimistic estimate of $Q^{\pi_k}_h$ in similar fashion in Line 11. The optimistic and pessimistic estimates $\underline{Q}_{k,h}, \tilde{Q}_{k,h}$ (resp. $\underline{V}_{k,h}, \tilde{V}_{k,h}$) allow us to compute the certificates $\epsilon_k$ and $\mathcal{I}_k$ and enables more sample-efficient learning. Specifically, Algorithm 1 uses a novel form of confidence intervals $\psi$ that explicitly depends on this difference. These confidence intervals are key for proving the following IPOC bound:

**Theorem 4** (Mistake IPOC Bound of Alg. 1)**.** *For any given $\delta \in (0,1)$, Alg. 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 Alg. 1 outputs a certificate $|\mathcal{I}_k| = \epsilon_k > \epsilon$ is*

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

By Proposition 3, this implies a Uniform-PAC bound of same order as well as the regret and PAC bounds listed in Table 1. This table also contains previous state of the art bounds of each type[2] as well as lower bounds. The IPOC lower bound follows from the PAC lower bound by Dann & Brunskill (2015) and Proposition 3. For $\epsilon$ small enough ($\leq O(1/(SH))$ specifically), our IPOC bound is minimax, i.e., the best achievable, up to log-factors. This is also true for the Uniform-PAC and PAC bounds implied by Theorem 4 as well as the implied regret bound when the number of episodes $T = \Omega(S^3AH^4)$ is large enough. ORLC is the first algorithm to achieve this minimax rate for PAC and Uniform-PAC. While UCBVI-BF achieves minimax regret for problems with small horizon, their bound is suboptimal when $H > SA$. The lower-order term in our regret bound $\tilde{O}(S^2AH^3)$ has a slightly worse dependency on $H$ than Azar et al. (2017) but we can trade-off a factor of $H$ for $A$ (see appendix) and believe that this term can be further reduced by a more involved analysis.

We defer details of our IPOC analysis to the appendix available at https://arxiv.org/abs/1811.03056 but the main advances leverage that $[\underline{Q}_{k,h}(s,a), \tilde{Q}_{k,h}(s,a)]$ is an *observable* confidence interval for both $Q^\star_h(s,a)$ and $Q^{\pi_k}_h(s,a)$. Specifically, our main novel insights are:

- While prior works (e.g. Lattimore & Hutter, 2012; Dann & Brunskill, 2015) control the suboptimality $Q^\star_h - Q^{\pi_k}_h$ of the policy by recursively bounding $\tilde{Q}_{k,h} - Q^{\pi_k}_h$, we instead recursively bound $\tilde{Q}_{k,h} - \underline{Q}_{k,h} \leq 2\psi_{k,h} + \hat{P}_k(\tilde{V}_{k,h+1} - \underline{V}_{k,h+1})$ which is not only simpler but also controls both the suboptimality of the policy and the size of the certificates simultaneously.
- As existing work (e.g. Azar et al., 2017; Jin et al., 2018), we use empirical Bernstein-type concentration inequalities to construct $\tilde{Q}_{k,h}(s,a)$ as an upper bound to $Q^\star_h(s,a) = r(s,a) + P(s,a)V^\star_{h+1}$. This results in a dependency of the upper bound on the variance of the optimal next state value $\sigma_{\hat{P}_k(s,a)}(V^\star_{h+1})^2$ under the empirical model. Since $V^\star_{h+1}$ is unknown this has to be upper-bounded by $\sigma_{\hat{P}_k(s,a)}(\tilde{V}_{k,h+1})^2 + B$ with an additional bonus $B$ to account for the difference between the values, $\tilde{V}_{k,h+1} - V^\star_{h+1}$, which is again unobservable. Azar et al. (2017) now constructs an observable bound on $B$ through an intricate regret analysis that involves additional high-probability bounds on error terms (see their $\mathcal{E}_{fr}/\mathcal{E}_{az}$ events) which causes the suboptimal $\sqrt{H^3T}$ term in their regret bound. Instead, we use the fact that $\tilde{V}_{k,h+1} - \underline{V}_{k,h+1}$ is an observable upper bound on $\tilde{V}_{k,h+1} - V^\star_{h+1}$ which we can directly use in our confidence widths $\psi_{k,h}$ (see the last term in Line 9 of Alg. 1). Hence, availability of lower bounds through certificates improves also our upper confidence bounds on $Q^\star_h$ and yields more sample-efficient exploration with improved performance bounds as we avoid additional high-probability bounds of error terms.

[2] These model-free and model-based methods have the best known bounds in our problem class. Q-learning with UCB and UBEV allow time-dependent dynamics. One might be able to improve their regret bound by $\sqrt{H}$ when adapting them to our setting. Note that by augmenting our state space with a time index, our algorithm also achieves minimax optimality with $\tilde{O}(\sqrt{SAH^3T})$ regret up to lower order terms in their setting.

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

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

$\phi(n) = 1 \wedge \sqrt{\frac{0.52}{n}\left(1.4 \ln\ln(e \vee n) + \ln\frac{26SA(H+1+S)}{\delta}\right)}$; $\quad \tilde{V}_{k,H+1}(s) = 0$; $\quad \underset{\sim}{V}_{k,H+1}(s) = 0 \quad \forall s \in \mathcal{S}, k \in \mathbb{N}$;

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

  **for** $s', s \in \mathcal{S}, a \in \mathcal{A}$ **do** // update empirical model and number of observations

    $n_k(s,a) = \sum_{i=1}^{k-1}\sum_{h=1}^{H} \mathbf{1}\{s_{i,h} = s, a_{i,h} = a\}$ ; // number of times (s,a) was observed

    $\hat{r}_k(s,a) = \frac{1}{n_k(s,a)}\sum_{i=1}^{k-1}\sum_{h=1}^{H} r_{i,h}\mathbf{1}\{s_{i,h} = s, a_{i,h} = a\}$ ; // avg. reward observed for (s,a)

    $\hat{P}_k(s'|s,a) = \frac{1}{n_k(s,a)}\sum_{i=1}^{k-1}\sum_{h=1}^{H} \mathbf{1}\{s_{i,h} = s, a_{i,h} = a, s_{i,h+1} = s'\}$

  **for** $h = H$ **to** $1$ ***and*** $s \in \mathcal{S}$ **do** // optimistic planning with upper and lower confidence bounds

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

      $\psi_{k,h}(s,a) = (1+\sqrt{12}\sigma_{\hat{P}_k(s,a)}(\tilde{V}_{k,h+1}))\phi(n_k(s,a)) + 45SH^2\phi(n_k(s,a))^2 + \frac{1}{H}\hat{P}(s,a)(\tilde{V}_{k,h+1} - \underset{\sim}{V}_{k,h+1})$;

      $\tilde{Q}_{k,h}(s,a) = 0 \vee \left(\hat{r}_k(s,a) + \hat{P}_k(s,a)\tilde{V}_{k,h+1} + \psi_{k,h}(s,a)\right) \wedge V_h^{\max}$ ; // UCB of $Q^\star_{h+1}$

      $\underset{\sim}{Q}_{k,h}(s,a) = 0 \vee \left(\hat{r}_k(s,a) + \hat{P}_k(s,a)\underset{\sim}{V}_{k,h+1} - \psi_{k,h}(s,a)\right) \wedge V_h^{\max}$ ; // LCB of $Q^{\pi_k}_{h+1}$

    $\pi_k(s,h) = \operatorname{argmax}_a \tilde{Q}_{k,h}(s,a)$; $\quad \tilde{V}_{k,h}(s) = \tilde{Q}_{k,h}(s, \pi_k(s,h))$; $\quad \underset{\sim}{V}_{k,h}(s) = \underset{\sim}{Q}_{k,h}(s, \pi_k(s,h))$;

  **output** policy $\pi_k$ with certificates $\mathcal{I}_k = [\underset{\sim}{V}_{k,1}(s_{1,1}), \tilde{V}_{k,1}(s_{1,1})]$ and $\epsilon_k = |\mathcal{I}_k|$;

  **sample episode** $k$ with policy $\pi_k$; // Observe $s_{k,1}, a_{k,1}, r_{k,1}, s_{k,2}, \ldots, s_{k,H}, a_{k,H}, r_{k,H}$

| Algorithm | Regret | PAC | Mistake IPOC |
|---|---|---|---|
| UCBVI-BF (Azar et al., 2017) | $\tilde{O}(\sqrt{SAH^2T} + \sqrt{H^3T} + S^2AH^2)$ | - | - |
| Q-l. w/ UCB$^2$ (Jin et al., 2018) | $\tilde{O}(\sqrt{SAH^4T} + S^{1.5}A^{1.5}H^{4.5})$ | - | - |
| UCFH (Dann & Brunskill, 2015) | - | $\tilde{O}\left(\frac{S^2AH^2}{\epsilon^2}\right)$ | - |
| UBEV [2] (Dann et al., 2017) | $\tilde{O}(\sqrt{SAH^4T} + S^2AH^3)$ | $\tilde{O}\left(\frac{SAH^4}{\epsilon^2} + \frac{S^2AH^3}{\epsilon}\right)$ | - |
| ORLC (this work) | $\tilde{O}(\sqrt{SAH^2T} + S^2AH^3)$ | $\tilde{O}\left(\frac{SAH^2}{\epsilon^2} + \frac{S^2AH^3}{\epsilon}\right)$ | $\tilde{O}\left(\frac{SAH^2}{\epsilon^2} + \frac{S^2AH^3}{\epsilon}\right)$ |
| Lower bounds | $\Omega\left(\sqrt{SAH^2T}\right)$ | $\Omega\left(\frac{SAH^2}{\epsilon^2}\right)$ | $\Omega\left(\frac{SAH^2}{\epsilon^2}\right)$ |

*Table 1.* Comparison of the state of the art and our bounds for episodic RL in tabular MDPs. A dash means that the algorithm does not satisfy a non-trivial bound without modifications. $T$ is the number of episodes and $\ln(1/\delta)$ factors are omitted for readability. For an empirical comparison of the sample-complexity of these approaches, see Appendix E.2 available at https://arxiv.org/abs/1811.03056.

- As opposed to the upper bounds, we cannot simply apply concentration inequalities to construct $\underset{\sim}{Q}_{k,h}(s,a)$ as a lower bound to $Q^{\pi_k}$ because the estimation target $Q^{\pi_k}(s,a) = r(s,a) + P(s,a)V^{\pi_k}_{h+1}$ is itself random. The policy $\pi_k$ depends in highly non-trivial ways on all samples from which we also estimate the empirical model $\hat{P}_k, \hat{r}_k$. A prevalent approach in model-based policy evaluation (Strehl & Littman, 2008; Ghavamzadeh et al., 2016, e.g.) to deal with this challenge is to instead apply a concentration argument on the $\ell_1$ distance of the transition estimates $\|P(s,a) - \hat{P}_k(s,a)\|_1 \le \sqrt{S}\phi(n_k(s,a))$. This yields confidence intervals that shrink at a rate of $H\sqrt{S}\phi(n_k(s,a))$. Instead, we can exploit that $\pi_k$ is generated by a sample-efficient algorithm and construct $\underset{\sim}{Q}_{k,h}$ as a lower bound to the non-random quantity $r(s,a) + P(s,a)V^\star_{h+1}$. We account for the difference $P(s,a)(V^\star_{h+1} - V^{\pi_k}_{h+1}) \le P(s,a)(\tilde{V}_{k,h+1} - \underset{\sim}{V}_{k,h+1})$ explicitly, again through a recursive bound. This allows us to achieve confidence intervals that shrink at a faster rate of $\psi_{k,h} \approx H\phi(n_k(s,a)) + SH^2\phi(n_k(s,a))^2$ without the $\sqrt{S}$ dependency in the dominating $\phi(n_k(s,a))$ term (recall $\phi(n_k(s,a)) \le 1$ and goes to 0). Hence, by leveraging that $\pi_k$ is computed by a sample-efficient approach, we improve the tightness of the certificates.

### 4.2. MDPs With Linear Side Information

We now present an algorithm for the more general setting with side information, which, for example, allows us to take 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 & Neu (2014). Its overall structure is the same as the tabular Algorithm 1 but here the empirical model are least-squares estimates of the model parameters evaluated at the current contexts. Specifically, the

**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 $\lambda > 0$

$\xi_{\theta^{(r)}} = \sqrt{d}$; $\xi_{\theta^{(p)}} = \sqrt{d}$; $\tilde{V}_{k,H+1}(s) = 0$; $\underset{\sim}{V}_{k,H+1}(s) = 0 \quad \forall s \in \mathcal{S}, k \in \mathbb{N}$;

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

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

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

  **for** $s, s' \in \mathcal{S}, a \in \mathcal{A}$ **do** // estimate model with least-squares

    $N_{k,s,a}^{(q)} = \lambda I + \sum_{i=1}^{k-1}\sum_{h=1}^{H} \mathbf{1}\{s_{i,h} = s, a_{i,h} = a\} x_k^{(q)} (x_k^{(q)})^\top$ for $q \in \{r, p\}$;

    $\hat{\theta}_{k,s,a}^{(r)} = (N_{k,s,a}^{(r)})^{-1} \sum_{i=1}^{k-1}\sum_{h=1}^{H} \mathbf{1}\{s_{i,h} = s, a_{i,h} = a\} x_k^{(r)} r_{i,h}$; $\hat{r}_k(s,a) = 0 \vee (x_k^{(r)})^\top \hat{\theta}_{k,s,a}^{(r)} \wedge 1$;

    $\hat{\theta}_{s',s,a}^{(p)} = (N_{k,s,a}^{(p)})^{-1} \sum_{i=1}^{k-1}\sum_{h=1}^{H} \mathbf{1}\{s_{i,h} = s, a_{i,h} = a, s_{i,h+1} = s'\} x_k^{(p)}$;

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

  **for** $h = H$ **to** $1$ ***and*** $s \in \mathcal{S}$ **do** // optimistic planning with ellipsoid confidence bounds

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

      $\psi_{k,h}(s,a) = \|\tilde{V}_{k,h+1}\|_1 \phi(N_{k,s,a}^{(p)}, x_k^{(p)}, \xi_{\theta^{(p)}}) + \phi(N_{k,s,a}^{(r)}, x_k^{(r)}, \xi_{\theta^{(r)}})$;

      $\tilde{Q}_{k,h}(s,a) = 0 \vee \left(\hat{r}_k(s,a) + \hat{P}_k(s,a)\tilde{V}_{k,h+1} + \psi_{k,h}(s,a)\right) \wedge V_h^{\max}$; // UCB of $Q^\star_{h+1}$

      $\underset{\sim}{Q}_{k,h}(s,a) = 0 \vee \left(\hat{r}_k(s,a) + \hat{P}_k(s,a)\underset{\sim}{V}_{k,h+1} - \psi_{k,h}(s,a)\right) \wedge V_h^{\max}$; // LCB of $Q^{\pi_k}_{h+1}$

    $\pi_k(s,h) = \operatorname{argmax}_a \tilde{Q}_{k,h}(s,a)$; $\tilde{V}_{k,h}(s) = \tilde{Q}_{k,h}(s, \pi_k(s,t))$; $\underset{\sim}{V}_{k,h}(s) = \underset{\sim}{Q}_{k,h}(s, \pi_k(s,t))$;

  **output** policy $\pi_k$ with certificates $\mathcal{I}_k = [\underset{\sim}{V}_{k,1}(s_{1,1}), \tilde{V}_{k,1}(s_{1,1})]$ and $\epsilon_k = |\mathcal{I}_k|$;

  **sample episode** $k$ with policy $\pi_k$ ; // Observe $s_{k,1}, a_{k,1}, r_{k,1}, s_{k,2}, \ldots, s_{k,H}, a_{k,H}, r_{k,H}$

empirical transition probability $\hat{P}_k(s'|s,a)$ is $(x_k^{(p)})^\top \hat{\theta}_{s',s,a}$ 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 $[0,1]$. This model is estimated separately for each $(s',s,a)$-triple, but generalizes across different contexts. The confidence widths $\psi_{k,h}$ are derived using ellipsoid confidence sets on model parameters. We show the following IPOC bound:

**Theorem 5** (Cumulative IPOC Bound for Alg. 2 )**.** *For any $\delta \in (0,1)$ and regularizer $\lambda > 0$, Alg. 2 satisfies the following cumulative IPOC bound in any MDP with contexts of dimensions $d^{(r)}$ and $d^{(p)}$ and bounded parameters $\xi_{\theta^{(r)}} \le \sqrt{d^{(p)}}, \xi_{\theta^{(p)}} \le \sqrt{d^{(p)}}$. With prob. at least $1-\delta$ all return certificates contain the return of $\pi_k$ and optimality certificates are upper bounds on the optimality gaps and their total sum after $T$ episodes is bounded for all $T$ by*

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

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 & 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 [3] and can likely be transferred to their setting. The mistake-type PAC bound by Modi et al. (2018) is not comparable because our cumulative IPOC bound does not imply a mistake-type PAC bound.[4] Nonetheless, loosely translating our result to a PAC-like bound yields $\tilde{O}\left(\frac{d^2S^3AH^5}{\epsilon^2}\right)$ which is much smaller than their $\tilde{O}\left(\frac{d^2SAH^4}{\epsilon^5}\max\{d^2, S^2\}\right)$ bound for small $\epsilon$.

The confidence bounds in Alg. 2 are more general but looser than those for the tabular case of Alg. 1. Instantiating the IPOC bound for Alg. 2 from Theorem 5 for tabular MDPs ($x_k^{(r)} = x_k^{(p)} = 1$) yields $\tilde{O}(\sqrt{S^3AH^4T})$ which is worse than the cumulative IPOC bound $\tilde{O}(\sqrt{SAH^2T} + S^2AH^3)$ of Alg. 1 implied by Thm. 4 and Prop. 3.

By Prop. 3, a mistake IPOC bound is stronger than the cumulative version we proved for Algorithm 2. One might wonder if Alg. 2 also satisfies a mistake bound, but in Appendix D (at https://arxiv.org/abs/1811.03056) we show that this is not the case because of its non-decreasing ellipsoid confidence sets. There could be other algorithms with mistake IPOC bounds for this setting, but they they would likely require entirely different confidence sets.

[3] 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).

[4] An algorithm with a sub-linear cumulative IPOC bound can output a certificate larger than a threshold $\epsilon_k \ge \epsilon$ infinitely often as long as it does so sufficiently less frequently (see Section D).

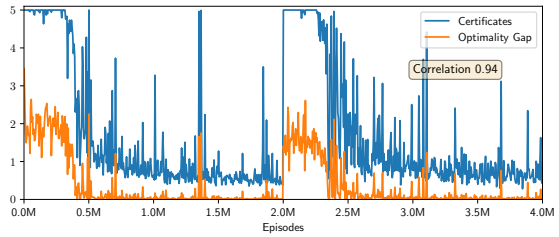


*Figure 1.* Certificates and (unobserved) optimality gaps of Algorithm 2 for 4M episodes on an MDP with context distribution shift after 2M (episodes sub-sampled for better visualization)

## 5. Simulation Experiment

One important use case for certificates is to detect sudden performance drops when the distribution of contexts changes. For example, in a call center dialogue system, there can be a sudden increase of customers calling due to a certain regional outage. We demonstrate that certificates can identify such performance drops caused by context shifts. 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, this causes a spike in optimality gap as well as in the optimality certificates. While our certificates need to upper bound the optimality gap / contain the return in each episode up to a small failure probability, even for the worst case, our algorithm reliably can detect this sudden decrease of performance. In fact, the optimality certificates have a very high correlation of 0.94 with the unobserved optimality gaps.

One also may wonder if our algorithms leads to improvements over prior approaches in practice or only in the theoretical bounds. To help answer this, we present results in Appendix E at `https://arxiv.org/abs/1811.03056` both on analyzing the policy certificates provided, and examining `ORLC`'s performance in tabular MDPs versus other recent papers with similar regret (Azar et al., 2017) or PAC (Dann et al., 2017) bounds. Encouragingly in the small simulation MDPs considered, we find that our algorithms lead to faster learning and better performance. Therefore while our primary contribution is theoretical results, these simulations suggest the potential benefits of the ideas underlying our proposed framework and algorithms.

## 6. Related Work

The connection of IPOC to other frameworks is formally discussed in Section 3. Our algorithms essentially compute confidence bounds as in OFU methods, and then use those in model-based policy evaluation to obtain policy certificates. There are many works on off-policy policy evaluation (e.g., Jiang & Li, 2016; Thomas & Brunskill, 2016; Mahmood et al., 2017), some of which 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 set of episodes sampled by fixed policies is given. Many approaches rely on importance weights that require stochastic data-collecting policies but most sample-efficient algorithms for which we would like to provide certificates 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.

Interestingly, there is very recent work (Zanette & Brunskill, 2019) that also observed the benefits of using lower bounds in optimism-based exploration in tabular episodic RL. Though both their and our work obtain improved theoretical results, the specific forms of the optimistic bonuses are distinct and the analyses differ in many parts (e.g., we provide (Uniform-)PAC and regret bounds instead of only regret bounds). Most importantly, our work provides policy certificate guarantees as a main contribution whereas that work focuses on problem-dependent regret bounds.

Approaches on safe exploration (Kakade & Langford, 2002; Pirotta et al., 2013; Thomas et al., 2015a; Ghavamzadeh et al., 2016) guarantee monotonically increasing performance by operating in a batch loop. Our work is orthogonal, as we are not restricting exploration but rather exposing its impact to the users and give them the choice to intervene.

## 7. Conclusion and Future Work

We introduced policy certificates to improve accountability in RL by enabling users to intervene if the guaranteed performance is deemed inadequate. Bounds in our new theoretical framework IPOC ensure that certificates indeed bound the return and suboptimality in each episode and prescribe the rate at which certificates and policy improve. By combining optimism-based exploration with model-based policy evaluation, we have created two algorithms for RL with policy certificates, including for tabular MDPs with side information. For tabular MDPs, we demonstrated that policy certificates help optimism-based policy learning and vice versa. As a result, our new algorithm is the first to achieve minimax-optimal PAC bounds up to lower-order terms for tabular episodic MDPs, and, also the first to have both, minimax PAC and regret bounds, for this setting.

Future areas of interest include scaling up these ideas to continuous state spaces, extending them to model-free RL, and to provide per-episode risk-sensitive guarantees on the reward obtained.

## Acknowledgements

Part of this work were completed while Christoph Dann was an intern at Google. We appreciate support from a Microsoft Faculty Fellowship and a NSF Career award.

## References

Abbasi-Yadkori, Y. and Neu, G. Online learning in MDPs with side information. *arXiv preprint arXiv:1406.6812*, 2014.

Azar, M. G., Osband, I., and Munos, R. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, pp. 263–272, 2017.

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

Dann, C., Lattimore, T., and Brunskill, E. Unifying PAC and regret: Uniform PAC bounds for episodic reinforcement learning. In *Advances in Neural Information Processing Systems*, pp. 5713–5723, 2017.

Dann, C., Jiang, N., Krishnamurthy, A., Agarwal, A., Langford, J., and Schapire, R. E. On oracle-efficient PAC RL with rich observations. In *Advances in Neural Information Processing Systems*, pp. 1422–1432, 2018.

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

Hallak, A., Di Castro, D., and Mannor, S. Contextual Markov decision processes. *arXiv:1502.02259*, 2015.

Howard, S. R., Ramdas, A., Mc Auliffe, J., and Sekhon, J. Uniform, nonparametric, non-asymptotic confidence sequences. *arXiv preprint arXiv:1810.08240*, 2018.

Jabbari, S., Joseph, M., Kearns, M., Morgenstern, J., and Roth, A. Fair learning in Markovian environments. *arXiv preprint arXiv:1611.03071*, 2016.

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

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

Jiang, N., Krishnamurthy, A., Agarwal, A., Langford, J., and Schapire, R. E. Contextual decision processes with low Bellman rank are PAC-learnable. In *International Conference on Machine Learning*, pp. 1704–1713, 2017.

Jin, C., Allen-Zhu, Z., Bubeck, S., and Jordan, M. I. Is Q-learning provably efficient? *arXiv preprint arXiv:1807.03765*, 2018.

Joseph, M., Kearns, M., Morgenstern, J. H., and Roth, A. Fairness in learning: Classic and contextual bandits. In *Advances in Neural Information Processing Systems*, pp. 325–333, 2016.

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

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

Kannan, S., Kearns, M., Morgenstern, J., Pai, M., Roth, A., Vohra, R., and Wu, Z. S. Fairness incentives for myopic agents. In *Proceedings of the 2017 ACM Conference on Economics and Computation*, pp. 369–386. ACM, 2017.

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

Lattimore, T. and Czepesvari, C. *Bandit Algorithms*. Cambridge University Press, 2018.

Lattimore, T. and Hutter, M. PAC bounds for discounted MDPs. In *International Conference on Algorithmic Learning Theory*, pp. 320–334. Springer, 2012.

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

Mahmood, A. R., Yu, H., and Sutton, R. S. Multi-step off-policy learning without importance sampling ratios. *arXiv preprint arXiv:1702.03006*, 2017.

Modi, A., Jiang, N., Singh, S., and Tewari, A. Markov decision processes with continuous side information. In *Algorithmic Learning Theory*, pp. 597–618, 2018.

Osband, I., Russo, D., and Van Roy, B. (More) efficient reinforcement learning via posterior sampling. In *Advances in Neural Information Processing Systems*, pp. 3003–3011, 2013.

Osband, I., Van Roy, B., and Wen, Z. Generalization and exploration via randomized value functions. In *International Conference on Machine Learning*, pp. 2377–2386, 2016.

Pirotta, M., Restelli, M., Pecorino, A., and Calandriello, D. Safe policy iteration. In *International Conference on Machine learning*, pp. 307–315, 2013.

Raghavan, M., Slivkins, A., Vaughan, J. W., and Wu, Z. S. The externalities of exploration and how data diversity helps exploitation. *arXiv preprint arXiv:1806.00543*, 2018.

Sajed, T., Chung, W., and White, M. High-confidence error estimates for learned value functions. *arXiv preprint arXiv:1808.09127*, 2018.

Strehl, A. L. and Littman, M. L. An analysis of model-based interval estimation for Markov decision processes. *Journal of Computer and System Sciences*, 74(8):1309–1331, 2008.

Strehl, A. L., Li, L., Wiewiora, E., Langford, J., and Littman, M. L. PAC model-free reinforcement learning. In *International Conference on Machine Learning*, 2006.

Strehl, A. L., Li, L., and Littman, M. L. Reinforcement learning in finite MDPs: PAC analysis. *Journal of Machine Learning Research*, 10:2413–2444, 2009.

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

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

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

Thomas, P. S., Theocharous, G., and Ghavamzadeh, M. High-confidence off-policy evaluation. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pp. 3000–3006, 2015b.

Zanette, A. and Brunskill, E. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. *https://arxiv.org/abs/1901.00210*, 2019.

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