Model-Free Approach to Evaluate Reinforcement Learning Algorithms

Denis Belomestny [1 2] **Ilya Levin** [2] **Eric Moulines** [3 2] **Alexey Naumov** [2] **Sergey Samsonov** [2] **Veronika Zorina** [2]

1. Introduction

The key objective of Reinforcement Learning (RL) is to learn an optimal agent's behaviour in an unknown environment. A natural performance metric is given by the value function V^π which is the expected total reward of the agent following π. Unfortunately, even a precise knowledge of V^π does not provide information on how far is the policy π from the optimal one. To address this issue a popular quality measures are the *regret bounds* of the algorithm (Jaksch et al., 2010) and *suboptimality gap (policy error)* (Szepesvári, 2010; Pires & Szepesvári, 2016). However, available estimates of both quantities are typically pessimistic and rely on the unknown quantities of the underlying Markov Decision Processes (MDP). Moreover, even if the bounds are known, they does not apply to the general policy π and depends significantly on the particular algorithm which produced it (Jin et al., 2018; Azar et al., 2017).

In this paper we are interested in deriving agnostic (model independent) bounds for the policy error using the concept of *upper solutions* to the Bellman optimality equation. Our approach is substantially different from the ones known in literature as it can be used to estimate the suboptimality gap for an arbitrary given policy π. The concept of upper solutions is closely related to martingale duality in optimal control and information relaxation approach, see (Belomestny & Schoenmakers, 2018), and references therein. The concept of upper solutions has also a connection to distributional RL, as it can be formulated pathwise or using distributional Bellman operator, see e.g. (Lyle et al., 2019).

The contributions of this paper are three-fold:

- We propose a novel approach to construct model free confidence bounds for the optimal value function $V^\star$ based on a notion of upper solutions.
- Given a policy π, we propose an upper value iterative procedure (UVIP) for constructing an a.s. upper bound for V^π such that it coincides with $V^\star$ if $\pi = \pi^\star$.
- We study convergence properties of the proposed algorithm. In particular, we show that the variance of the resulting upper bound is small if π is close to π^* leading to the tight confidence bounds for $V^\star$.

[1]Duisburg-Essen University, Germany [2]National Research University Higher School of Economics, Russia [3]Ecole Polytechnique, France. Correspondence to: Denis Belomestny <denis.belomestny@uni-due.de>, Veronika Zorina <vozorina@hse.ru>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021. Copyright 2021 by the author(s).

Definitions and notations A *Markov Decision Process* (MDP) is a tuple $(\mathsf{X}, \mathsf{A}, \mathscr{P}, r, \gamma)$, where X is the state space, A is the action space, $\mathscr{P} = (\mathsf{P}^a)_{a \in \mathsf{A}}$ is the *transition probability kernel*, $r = (r^a)_{a \in \mathsf{A}}$ is the *reward function*, $0 < \gamma < 1$ is the *discount factor*. A *policy*, is denoted as π. An *optimal policy* $\pi^\star$ is one that achieves the maximum possible value amongst all policies in each state $x \in \mathsf{X}$. The *optimal value* for state x is denoted by $V^\star(x)$. The *value function* of a policy π in a state $x \in \mathsf{X}$ is denoted by $V^\pi(x)$, similarly in a state $x \in \mathsf{X}$ and $a \in \mathsf{A}$ the *action-value function* $Q^\pi(x,a)$. Let us denote the space of bounded measurable functions with domain X by $\mathcal{B}(\mathsf{X})$ equipped with the norm $\|f\|_{\mathsf{X}} = \sup_{x \in \mathsf{X}} |f(x)|$ for any $f \in \mathcal{B}(\mathsf{X})$. The *Bellman return operator* w.r.t P, $\mathsf{T}_{\mathsf{P}} : \mathcal{B}(\mathsf{X}) \to \mathcal{B}(\mathsf{X} \times \mathsf{A})$, is defined by $(\mathsf{T}_{\mathsf{P}} V)(x,a) = r^a(x) + \gamma \mathsf{P}^a V(x)$, where $(\mathsf{P}^a V)(x) = \int V(x') \mathsf{P}^a(dx'|x)$. We also define the *maximum selection* operator $\mathsf{M} : \mathcal{B}(\mathsf{X} \times \mathsf{A}) \to \mathcal{B}(\mathsf{X})$ by $(\mathsf{M}V^{\cdot})(x) = \max_a V^a(x)$. Then $\mathsf{M}\mathsf{T}_{\mathsf{P}}$ corresponds to the *Bellman optimality operator*. The optimal value function $V^\star$ satisfies a non-linear fixed-point equation

$$V^\star(x) = \mathsf{M}\mathsf{T}_{\mathsf{P}} V^\star(x). \tag{1}$$

which is the *Bellman optimality equation*. We write $Y^{x,a}, x \in \mathsf{X}, a \in \mathsf{A}$ for a random variable generated according to $\mathsf{P}^a(\cdot|x)$, and define a random Bellman operator $(\widetilde{\mathsf{T}}_{\mathsf{P}} V)(x) \mapsto r^a(x) + \gamma V(Y^{x,a})$.

2. UVIP Algorithm

A straightforward approach to bound the policy error $\Delta_\pi(x) \overset{\text{def}}{=} V^\star(x) - V^\pi(x)$ requires the estimation of the optimal value function $V^\star(x)$. Unfortunately, (1) does not allow to represent $V^\star$ as an expectation. Thus the problem of estimating $V^\star$ can not be naturally reduced to a stochastic approximation problem. Moreover, for sequence of value iteration procedure $V_{k+1} = \mathsf{M}\mathsf{T}_{\mathsf{P}} V_k$, if $(\mathsf{P}^a)_{a \in \mathsf{A}}$ is replaced by its empirical estimate $\hat{\mathsf{P}}^a$ the desired upper biasedness property $V_k(x) \geq V^\star(x)$ is lost. Below we describe our approach, which is based on the following key assumptions:

- we consider infinite-horizon MDPs with discount factor $\gamma < 1$;
- we can sample from the conditional distribution $\mathsf{P}^a(\cdot|x)$ for any $x \in \mathsf{X}$ and $a \in \mathsf{A}$.

The key concept of our algorithm is an *upper solution*, introduced below.

Definition 2.1. *We call a function V^{up} an upper solution to the Bellman optimality equation* (1) *if*

$$V^{\mathrm{up}}(x) \geq \mathsf{M}\mathsf{T}_{\mathsf{P}} V^{\mathrm{up}}(x)\,, \forall x \in \mathsf{X}\,.$$

Upper solutions can be used to build tight upper bounds for the optimal value function $V^\star$. Let $\Phi \in \mathcal{B}(\mathsf{X})$ be a *martingale function* w.r.t. the operator P^a, that is, $\mathsf{P}^a\Phi(x) = 0$ for all $a \in \mathsf{A}, x \in \mathsf{X}$. Define V^{up} as a solution to the following fixed point equation:

$$V^{\mathrm{up}}(x) = \mathbb{E}[\max_a\{r^a(x) + \\ + \gamma(V^{\mathrm{up}}(Y^{x,a}) - \Phi(Y^{x,a}))\}], \quad (2)$$

where $Y^{x,a} \sim \mathsf{P}^a(\cdot|x)$. In terms of the random Bellman operator $\widetilde{\mathsf{T}}_{\mathsf{P}}$, we can rewrite (2) as $V^{\mathrm{up}} = \mathbb{E}[\mathsf{M}\widetilde{\mathsf{T}}_{\mathsf{P}}(V^{\mathrm{up}} - \Phi)]$. It is easy to see that (2) defines an upper solution. Indeed, for any $x \in \mathsf{X}$,

$$\begin{aligned} V^{\mathrm{up}}(x) &\geq \max_a \mathbb{E}[r^a(x) + \gamma(V^{\mathrm{up}}(Y^{x,a}) - \Phi(Y^{x,a}))] \\ &= \max_a\{r^a(x) + \gamma \mathsf{P}^a V^{\mathrm{up}}(x)\} = \mathsf{M}\mathsf{T}_{\mathsf{P}} V^{\mathrm{up}}(x)\,. \end{aligned}$$

Note that unlike the optimal state value function $V^\star$, the upper solution V^{up} is represented as an expectation, which allows us to use various stochastic approximation methods to compute V^{up}. The Banach's fixed-point theorem implies that for iterates

$$V^{\mathrm{up}}_{k+1} = \mathbb{E}[\mathsf{M}\widetilde{\mathsf{T}}_{\mathsf{P}}(V^{\mathrm{up}}_k - \Phi)], \quad k \in \mathbb{N},$$

we have convergence $V^{\mathrm{up}}_k \to V^{\mathrm{up}}$ as $k \to \infty$. Moreover, V^{up} does not depend on V^{up}_0 and $V^{\mathrm{up}}_k(x) \geq V^\star(x)$ for any $k \in \mathbb{N}, x \in \mathsf{X}$, provided that $V^{\mathrm{up}}_0(x) \geq V^\star(x)$. Given a policy π and the corresponding value function V^π, we set $\Phi^{x,a}_\pi(y) \stackrel{\mathrm{def}}{=} V^\pi(y) - (\mathsf{P}^a V^\pi)(x)$. It is easy to check that $\mathsf{P}^a \Phi^{x,a}_\pi(x) = 0$. This leads to the upper value iterative procedure (UVIP):

$$V^{\mathrm{up}}_{k+1}(x) = \mathbb{E}[\mathsf{M}\widetilde{\mathsf{T}}_{\mathsf{P}}(V^{\mathrm{up}}_k - \Phi^{x,\cdot}_\pi)(x)] = \\ \mathbb{E}\big[\max_a\{r^a(x) + \gamma(V^{\mathrm{up}}_k(Y^{x,a}) - \Phi^{x,a}_\pi(Y^{x,a}))\}\big] \quad (3)$$

with $V^{\mathrm{up}}_0 \in \mathcal{B}(\mathsf{X})$. Further note that by taking $\Phi^{x,a}(y) \stackrel{\mathrm{def}}{=} V^\star(y) - (\mathsf{P}^a V^\star)(x)$, we get with probability 1 :

$$V^\star(x) = (\mathsf{M}\widetilde{\mathsf{T}}_{\mathsf{P}}(V^\star - \Phi^{x,\cdot}))(x) = \\ \max_a\{r^a(x) + \gamma(V^\star(Y^{x,a}) - \Phi^{x,a}(Y^{x,a}))\}, \quad (4)$$

that is, (4) can be viewed as an almost sure version of the Bellman equation $V^\star = \mathsf{M}\mathsf{T}_{\mathsf{P}} V^\star$. The upper solutions can be used to evaluate the quality of the policies and to construct confidence intervals for $V^\star$. It is clear that

$$V^\pi(x) \leq V^\star(x) \leq V^{\mathrm{up}}_k(x)$$

for any $k \in \mathbb{N}$ and $x \in \mathsf{X}$, thus a policy π can be evaluated by computing the difference $\Delta^{\mathrm{up}}_{\pi,k}(x) \doteq V^{\mathrm{up}}_k(x) - V^\pi(x) \geq \Delta_\pi(x)$. Representations (3) and (4) imply

$$\left\|V^{\mathrm{up}}_{k+1} - V^\star\right\|_{\mathsf{X}} \leq \gamma \left\|V^{\mathrm{up}}_k - V^\star\right\|_{\mathsf{X}} + 2\gamma \left\|V^\pi - V^\star\right\|_{\mathsf{X}},$$

$k \in \mathbb{N}$. Hence, we derive that $\Delta^{\mathrm{up}}_\pi \doteq \lim_{k\to\infty} \Delta^{\mathrm{up}}_{\pi,k}$ satisfies

$$\|\Delta_\pi\|_{\mathsf{X}} \leq \|\Delta^{\mathrm{up}}_\pi\|_{\mathsf{X}} \leq \left(1 + 2\gamma(1-\gamma)^{-1}\right)\|V^\star - V^\pi\|_{\mathsf{X}}. \quad (5)$$

As a result $\Delta^{\mathrm{up}}_\pi = 0$ if $\pi = \pi^\star$ and the corresponding confidence intervals collapses into one point. The quantity $\Delta^{\mathrm{up}}_{\pi,k}$ can be used to measure the quality of policies π obtained by many well-known algorithms like Reinforce, A2C, etc.

For simplicity, below we will describe all the results for finite state and action spaces ($|\mathsf{X}|, |\mathsf{A}| \leq \infty$), providing a short remark on a generalization of these results to continuous ones. Basically, the general iteration procedure is given by (3).

For all expectations in (3) we use empirical counterparts. Algorithm 1 contains the pseudocode of UVIP.

Algorithm 1 UVIP

Input: $V^\pi, \widehat{V}^{\mathrm{up}}_0, \gamma, \varepsilon, M_1, M_2$
Output: V^{up}
  for $x \in \mathsf{X}, a \in \mathsf{A}$ **do**
    $\overline{V}(x,a) = M_1^{-1}\sum_{i=1}^{M_1} V^\pi(Y^{x,a}_i),\ Y^{x,a}_i \sim \mathsf{P}^a(\cdot|x)$
    for $y \in \mathsf{X}$ **do**
      $\Phi^{x,a}_\pi(y) = V^\pi(y) - \overline{V}(x,a)$
    end for
  end for
  $k = 1$
  while $\|\widehat{V}^{\mathrm{up}}_k - \widehat{V}^{\mathrm{up}}_{k-1}\|_{\mathsf{X}} > \varepsilon$ **do**
    for $x \in \mathsf{X}$ **do**
      $\widehat{V}^{\mathrm{up}}_{k+1}(x) = M_2^{-1}\sum_{i=1}^{M_2}[\max_a\{r^a(x) +$
        $+\gamma(\widehat{V}^{\mathrm{up}}_k(Y^{x,a}_i) - \Phi^{x,a}_\pi(Y^{x,a}_i))\}],$
        $Y^{x,a}_i \sim \mathsf{P}^a(\cdot|x)$
    end for
    $k = k + 1$
  end while
  $V^{\mathrm{up}} = \widehat{V}^{\mathrm{up}}_k$

3. Convergence results

In this section, we analyze the distance between $(\widehat{V}^{\mathrm{up}}_k)_{k\in\mathbb{N}}$ and $V^\star$, where $\widehat{V}^{\mathrm{up}}_k(x)$ is the k-th iterate of Algorithm 1.

Note that with $\overline{V}_k^{\mathrm{up}}(x) \overset{\mathrm{def}}{=} \mathsf{E}\bigl[\widehat{V}_k^{\mathrm{up}}(x)\bigr]$ we have

$$\overline{V}_k^{\mathrm{up}}(x) \geq \max_a \bigl\{ r^a(x) + \gamma \mathsf{P}^a \overline{V}_{k-1}^{\mathrm{up}}(x) \bigr\}, \qquad (6)$$

$x \in \mathsf{X}$, $k \in \mathbb{N}$. Furthermore, if $\widehat{V}_0^{\mathrm{up}}(x) \geq V^\star(x)$ for $x \in \mathsf{X}$, then $\overline{V}_k^{\mathrm{up}}(x) \geq V^\star(x)$ for any $x \in \mathsf{X}$ and $k \in \mathbb{N}$. Hence $\widehat{V}_k^{\mathrm{up}}$ is an upper-biased estimate of $V^\star$ for any $k \geq 0$. Before stating our convergence results, we first state a number of technical assumptions.

A1. *There exists a measurable mapping $\psi : \mathsf{X} \times \mathsf{A} \times \mathbb{R}^m \to \mathsf{X}$ such that $Y^{x,a} = \psi(x, a, \xi)$, where ξ is a random variable with values in $\Xi \subseteq \mathbb{R}^m$ and distribution P_ξ on Ξ, that is, $\psi(x, a, \xi) \sim \mathsf{P}^a(\cdot|x)$.*

A2. *For some $R_{\max} > 0$ and all $a \in \mathsf{A}$, $\|r^a\|_{\mathsf{X}} \leq R_{\max}$.*

Suppose that for each $k = 1, \ldots, K$ we use an i.i.d. sample $\boldsymbol{\xi}_k = (\xi_{k,1} \ldots, \xi_{k,M_1+M_2}) \sim \mathrm{P}_\xi^{\otimes(M_1+M_2)}$ to generate $Y_j^{x,a} = \psi(x, a, \xi_{k,j}), j = 1, \ldots, M_1 + M_2$, and these samples are independent for different k. We now state main theorems that can be proved.

Theorem 3.1. *Assume A1, A2. Then for any $k \in \mathbb{N}$ and $\delta \in (0, 1)$ it holds with probability at least $1 - \delta$ that*

$$\|\widehat{V}_k^{\mathrm{up}} - V^*\|_{\mathsf{X}} \lesssim \gamma^k \bigl\|\widehat{V}_0^{\mathrm{up}} - V^*\bigr\|_{\mathsf{X}} + \\ + \|V^\pi - V^*\|_{\mathsf{X}} + \sqrt{\frac{\log(|\mathsf{X}||\mathsf{A}|/\delta)}{M_1}} . \qquad (7)$$

In the above bound $\lesssim$ stands for inequality up to a constant depending on γ and $R_{\max}$.

Variance of the estimator and confidence bounds. Our next step is to bound the variance of the estimator $\widehat{V}_k^{\mathrm{up}}(x)$. Denote

$$\sigma_k \overset{\mathrm{def}}{=} \gamma^k \bigl\|\widehat{V}_0^{\mathrm{up}} - V^*\bigr\|_{\mathsf{X}} + \\ + \|V^\pi - V^*\|_{\mathsf{X}} + \sqrt{\frac{\log(|\mathsf{X}||\mathsf{A}|)}{M_1}} . \qquad (8)$$

Note that under A1, A2, and $\bigl\|\widehat{V}_0^{\mathrm{up}}\bigr\|_{\mathsf{X}} \leq R_{\max}(1-\gamma)^{-1}$,

$$\sigma_k \lesssim \|V^\pi - V^\star\|_{\mathsf{X}} \ , \qquad (9)$$

provided that k and M_1 are large enough. The next theorem implies that $\mathsf{Var}\bigl[\widehat{V}_k^{\mathrm{up}}(x)\bigr]$ can be much smaller than the standard rate $1/M_2$, provided that V^π is close to V^* and M_1, K are large enough.

Theorem 3.2. *Assume A1, A2. Then for any $k \in \mathbb{N}$,*

$$\max_{x \in \mathsf{X}} \mathsf{Var}\bigl[\widehat{V}_k^{\mathrm{up}}(x)\bigr] \lesssim \sigma_k^2 \log(\mathrm{e} \vee \sigma_\mathrm{k}^{-1}) \mathrm{M}_2^{-1} \ , \qquad (10)$$

Corollary 3.1. *Recall that $\widehat{V}_k^{\mathrm{up}}$ is an upper biased estimate of $V^\star$ in a sense that $\overline{V}_k^{\mathrm{up}}(x) \geq V^\star(x)$ provided $\widehat{V}_0^{\mathrm{up}}(x) \geq V^\star(x)$ for $x \in \mathsf{X}$. Together with Theorem 3.2, it implies that for any $\delta \in (0, 1)$, with probability at least $1 - \delta$,*

$$V^\pi(x) \leq V^\star(x) \leq \widehat{V}_k^{\mathrm{up}}(x) \\ + \sigma_k \sqrt{\mathrm{C} \log(\mathrm{e} \vee \sigma_k^{-1}) \delta^{-1} M_2^{-1}} . \qquad (11)$$

Note that bounds of type (11) are known in the literature only in the case of specific policies π. For example, (Wainwright, 2019) proves bounds of this type for greedy policies in tabular Q-learning. At the same time, (11) holds for arbitrary policy π.

4. Numerical Results

In this section we demonstrate the performance of Algorithm 1 on several tabular state-space RL problems. We construct the upper confidence bounds for policy π, coming from the Value iteration procedure and Reinforce algorithm. Recall that the closer policy π is to the optimal one $\pi^\star$, the smaller is the difference between $V^\pi(x)$ and $V^{\mathrm{up},\pi}(x)$.

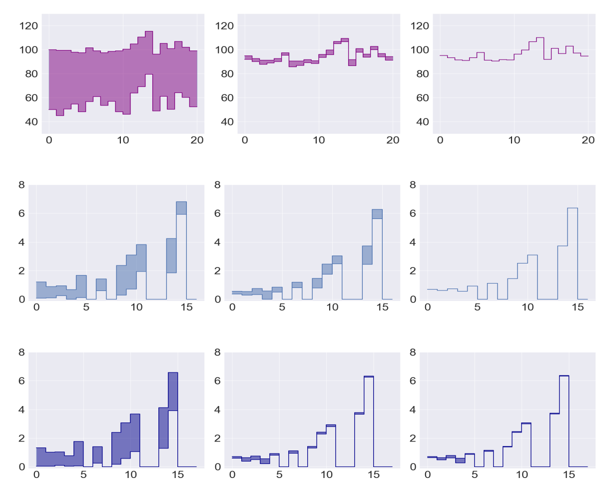

Figure 1. The difference between $V^{\mathrm{up},\pi_i}(x)$ and $V^{\pi_i}(x)$. X-axis represents states in a discrete environment for all pictures. Each group of three pictures of the same color demonstrates the process of learning the policy from the first iteration to the last. *First and second rows*: Evaluation of the policies during the process of Value iteration for Garnet and Frozen Lake. The policies are obtained greedily from $Q_i(x, a)$ function at the i-th step. *Third row*: Reinforce algorithm during learning for Frozen Lake.

We consider 2 popular tabular environments: Garnet ((Archibald et al., 1995)) and AI Gym Frozen Lake ((Brockman et al., 2016)). For each environment we perform K updates of the Value iteration with known transition kernel P^a. We denote the k-th step estimate of the action-value function as $\hat{Q}_k(x, a)$ and denote π_k the greedy policy w.r.t. $\hat{Q}_k(x, a)$. Then we evaluate the policies π_k with the Algorithm 1 for certain iteration numbers k. Figure 1 displays the gap between $V^{\pi_k}(x)$ and $V^{\mathrm{up},\pi_k}(x)$, which converges to zero while π^k converges to the optimal policy $\pi^\star$. Data for the upper bounds estimation is generated using *off-policy* method. On the Frozen Lake environment we also apply the tabular version of the Reinforce algorithm. We evaluate policies π_k obtained from the $k-$th Reinforce iteration. On the Figure 1 we display $V^{\pi_k}(x)$ and $V^{\mathrm{up},\pi_k}(x)$ for different time steps k. The difference $V^{\mathrm{up},\pi_k}(x) - V^{\pi_k}(x)$ does not converge to zero, indicating suboptimality of the Reinforce policy.

5. Conclusion

We propose a new approach towards model-free evaluation of the agent's policies in RL, based on upper solutions to the Bellman optimality equation (1). To the best of our knowledge, the UVIP is the first procedure which allows to construct the non-asymptotic confidence intervals for the optimal value function $V^\star$ based on the value function corresponding to an arbitrary policy π. In our analysis we consider only infinite-horizon MDPs and assume that sampling from the conditional distribution $\mathsf{P}^a(\cdot|x)$ is feasible for any $x \in \mathsf{X}$ and $a \in \mathsf{A}$. A promising future research direction is to generalize the algorithm to the case of finite-horizon MDPs combining it with the idea of Real-time dynamic programming (see (Efroni et al., 2019)).

It is worth to highlight that the Theorems 3.1 and 3.2 have a generalization for the case of infinite state and action spaces, which requires the introduction of the covering number of set $\mathsf{X} \times \mathsf{A}$, the Dudley's integral, along with the proper approximation for $V^{\mathrm{up}}(x)$. Moreover, the UVIP can be adapted for RL benchmarks with continuous state space and discrete action space by performing an additional approximation step. It implies that for an arbitrary policy we can construct the upper confidence bounds for such environments as AI Gym CartPole and Acrobot. Nevertheless, the success of the procedure relies on the policy evaluation methods. We have to choose the approximation points properly to be able to assess the next states quality after one step of the agent.

References

Archibald, T. W., McKinnon, K. I. M., and Thomas, L. C. On the generation of Markov Decision Processes. *The Journal of the Operational Research Society*, 46(3):354–361, 1995. ISSN 01605682, 14769360.

Azar, M. G., Osband, I., and Munos, R. Minimax regret bounds for reinforcement learning. *Proceedings of the 34th International Conference on Machine Learning*, 70: 263–272, 06–11 Aug 2017.

Belomestny, D. and Schoenmakers, J. *Advanced Simulation-Based Methods for Optimal Stopping and Control: With Applications in Finance*. Springer, 2018.

Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., and Zaremba, W. Openai gym, 2016.

Efroni, Y., Merlis, N., Ghavamzadeh, M., and Mannor, S. Tight regret bounds for model-based reinforcement learning with greedy policies. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Jaksch, T., Ortner, R., and Auer, P. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(51):1563–1600, 2010.

Jin, C., Allen-Zhu, Z., Bubeck, S., and Jordan, M. I. Is q-learning provably efficient? In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Lyle, C., Bellemare, M. G., and Castro, P. S. A comparative analysis of expected and distributional reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 4504–4511, 2019.

Pires, B. Á. and Szepesvári, C. Policy error bounds for model-based reinforcement learning with factored linear models. In *Conference on Learning Theory*, pp. 121–151. PMLR, 2016.

Szepesvári, C. Algorithms for reinforcement learning. *Synthesis lectures on artificial intelligence and machine learning*, 4(1):1–103, 2010.

Wainwright, M. J. Variance-reduced q-learning is minimax optimal. *CoRR*, 2019.