

Notes of [\[2\]](#)

Chao Tao

Feb. 14, 2020

1 Problem Setup

There is a tabular *episodic* MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathbb{P}, R, H)$ where R is bounded within $[0, 1]$ and only the transition probability \mathbb{P} is *unknown*. For simplicity, we also assume the reward function R is *deterministic*. We want to find a policy such that the regret incurred by this policy after K episodes is minimized. Given a policy $\pi = (\pi_1, \dots, \pi_K)$, the regret incurred by this policy is defined by

$$\mathcal{R}_K^\pi \stackrel{\text{def}}{=} \mathbb{E} \left[\sum_{k=1}^K (V_1^* - V_1^{\pi_k})(x_{k,1}) \right],$$

where the initial state for each episode can be either randomized or *adversarial*.

2 Notations and Definitions

$[n]$	$\{1, 2, \dots, n\}$
\mathcal{A}	action space
A	$ \mathcal{A} $
\mathcal{S}	state space
S	$ \mathcal{S} $
H	horizon
K	# of episodes
T	HK
$R : \mathcal{S} \times \mathcal{A} \rightarrow [0, 1]$	<i>known</i> reward function
$\mathbb{P} : \mathcal{S} \times \mathcal{A} \rightarrow \Delta(\mathcal{S})$	transition probability of the underlying MDP
$\pi = (\pi_1, \dots, \pi_K)$	an arbitrary policy where π_k is the policy in the k th episode
$Q_h^{\pi_k} : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$	Q -value function of policy π_k starting from time h
$V_h^{\pi_k} : \mathcal{S} \rightarrow \mathbb{R}$	value function of policy π_k starting from time h
Q_h^*	Q -value function of the optimal policy starting from time h
V_h^*	value function of the optimal policy starting from time h
$x_{k,1}$	initial state of the k th episode
$(x_{k,h}, a_{k,h})$	state-action pair in the k th episode and at the h th time step
\mathcal{H}_k	history before the k th episode $(x_{1,1}, a_{1,1}, \dots, x_{1,H+1}, \dots, x_{k-1,1}, a_{k-1,1}, \dots, x_{k-1,H+1})$
$n_{k,h}(x, a)$	number of hits of state-action pair (x, a) at the h th time step <i>before</i> the k th episode
$n_k(x, a)$	number of hits of state-action pair (x, a) <i>before</i> the k th episode
$\tilde{Q}_{k,h}$	estimate of the optimal Q -value function starting from the h th step of the k th episode
$\tilde{V}_{k,h}$	estimate of the optimal value function starting from the h th step of the k th episode
ρ	an arbitrary transition probability
V	an arbitrary value function
$(\rho V)(x, a)$	$\sum_{y \in \mathcal{S}} \rho(y x, a) V(y)$
\mathcal{R}_K^π	regret incurred by policy π

3 Algorithm

Algorithm 1: Q-learning with UCB-Hoeffding ([2])

```

1 initialization:  $\tilde{Q}_{1,h}(x, a) = H - h + 1$  for every  $(h, x, a) \in [H] \times \mathcal{S} \times \mathcal{A}$ 
2 for episode  $k = 1$  to  $K$  do
3   if  $k > 1$  then
4     call Algorithm 2 to compute  $\tilde{Q}_{k,\cdot}(\cdot, \cdot)$  and  $\tilde{V}_{k,\cdot}(\cdot)$ 
5   for step  $h = 1$  to  $H$  do
6     observe state  $x_{k,h}$ 
7     take action  $a_{k,h} = \operatorname{argmax}_{a \in \mathcal{A}} \tilde{Q}_{k,h}(x_{k,h}, a)$ 

```

Algorithm 2: Computation of $\tilde{Q}_{k,\cdot}(\cdot, \cdot)$ and $\tilde{V}_{k,\cdot}(\cdot)$

```

1 initialization:  $\tilde{Q}_{k,H+1}(x, a) = \tilde{V}_{k,H+1}(x, a) = 0$  for every  $(x, a) \in \mathcal{S} \times \mathcal{A}$ 
2 for step  $h = 1$  to  $H$  do
3   for every state-action pair  $(x, a)$  do
4     if  $(x, a) = (x_{k-1,h}, a_{k-1,h})$  then
5       let  $t = n_{k,h}(x, a)$ ,  $\alpha_t = \frac{H+1}{H+t}$  and  $\beta_t = c_1 \sqrt{2} \cdot \sqrt{\frac{H^3 \ln(SAH/\delta)}{t}}$ 
6        $\tilde{Q}_{k,h}(x, a) = \tilde{Q}_{k-1,h}(x, a) + \alpha_t(R(x, a) + \tilde{V}_{k-1,h+1}(x_{k-1,h+1}) + \beta_t - \tilde{Q}_{k-1,h}(x, a))$ 
7     else
8        $\tilde{Q}_{k,h}(x, a) = \tilde{Q}_{k-1,h}(x, a)$ 
9   for every state  $x \in \mathcal{S}$  do
10     $\tilde{V}_{k,h}(x) = \min\{H + 1 - h, \max_{a \in \mathcal{A}} \tilde{Q}_{k,h}(x, a)\}$ 

```

Here c_1 is a constant which will be defined later.

Remark 1. Algorithm 1 is model free since it does not explicitly calculate the transition probability. Hence its running time during each time step is $\mathcal{O}(SA)$.

4 Proofs

4.1 Favorable Events

4.1.1 \mathcal{E}_1

Given any $(h, x, a) \in [K] \times [H] \times \mathcal{S} \times \mathcal{A}$, let k_i be the episode within which state-action pair (x, a) was hit at the h th time step for the i th time. Define t be the total number of hits. Note that t depends on (h, x, a) .

But for cleaner presentation, we have dropped that dependency in the notation. Let

$$\mathcal{E}_1 \stackrel{\text{def}}{=} \left\{ \forall (h, x, a) \in [K] \times [H] \times \mathcal{S} \times \mathcal{A}, \left| \sum_{i=1}^t \alpha_t^i (V_h^*(x_{k_i, h+1}) - (\mathbb{P}V_h^*)(x, a)) \right| \leq c_1 \sqrt{H^2 \sum_{i=1}^t (\alpha_t^i)^2 \cdot \ln(SAH/\delta)} \right\},$$

where c_1 is a constant which will be defined later.

By Azuma's inequality (Lemma 12) and a union bound, there exists a constant c_1 such that $\Pr(\mathcal{E}_1) \geq 1 - \delta/2$.

4.2 Main Theorem

Theorem 2. *With probability at least $1 - \delta$, the regret incurred by Algorithm 1 is bounded by*

$$\mathcal{O} \left(SAH^2 + H^2 \sqrt{SAT \ln(SAH/\delta)} + \sqrt{TH^2 \ln(\delta^{-1})} \right).$$

Remark 3. *When T is large, the upper bound becomes $\tilde{\mathcal{O}} \left(H^2 \sqrt{SAT} \right)$.*

Remark 4. *The proof can be applied to the MDP where $\mathbb{P}_i \neq \mathbb{P}_j$ for $i \neq j$. Here \mathbb{P}_i denotes the transition probability at the i th time step.*

Remark 5. *There exists a refined proof giving an upper bound $\tilde{\mathcal{O}} \left(\sqrt{H^3 SAT} \right)$ [2].*

Proof. The following arguments are conditioned on event $\mathcal{E} \stackrel{\text{def}}{=} \mathcal{E}_1 \wedge \mathcal{E}_2$, where \mathcal{E}_2 will be defined later. And for simplicity, we use $\pi = (\pi_1, \dots, \pi_K)$ to represent Algorithm 1.

We first prove that the estimated Q -value function $\tilde{Q}_{k,h}(x, a)$ is optimistic.

Lemma 6. *For every $(k, h, x, a) \in [K] \times [H] \times \mathcal{S} \times \mathcal{A}$, it holds that*

$$\tilde{Q}_{k,h}(x, a) \geq Q_h^*(x, a).$$

Corollary 7. *For every $(k, h, x) \in [K] \times [H] \times \mathcal{S}$, it holds that $\tilde{V}_{k,h}(x) \geq V_h^*(x)$.*

Proof. Fix (k, h, x, a) where $k > 1$ and $n_{k,h}(x, a) > 0$ and let $t = n_{k,h}(x, a)$ which shares the same definition as that in Algorithm 2. Note that

$$\begin{aligned} \tilde{Q}_{k,h}(x, a) &= (1 - \alpha_t) \tilde{Q}_{prev(k),h}(x, a) + \alpha_t (R(x, a) + \tilde{V}_{prev(k),h+1}(x_{prev(k),h+1}) + \beta_t) \\ &= \dots \\ &= \alpha_t^0 \cdot \tilde{Q}_{1,h}(x, a) + \sum_{i=1}^t \alpha_t^i \cdot (R(x, a) + \tilde{V}_{k_i,h+1}(x_{k_i,h+1})) + \sum_{i=0}^t \alpha_t^i \beta_i, \end{aligned} \quad (1)$$

where we have defined k_i and $prev(k)$ as the episode when the i th time and the last time that state-action pair (x, a) was hit at the h th time step *before* the k th episode respectively and

$$\alpha_t^0 \stackrel{\text{def}}{=} \prod_{j=1}^t (1 - \alpha_j), \quad \alpha_t^i \stackrel{\text{def}}{=} \prod_{j=i+1}^t (1 - \alpha_j) \cdot \alpha_i.$$

Subtracting both sides of (1) by $Q_h^*(x, a)$, we obtain

$$\begin{aligned}
\tilde{Q}_{k,h}(x, a) - Q_h^*(x, a) &= \alpha_t^0 \cdot (\tilde{Q}_{1,h}(x, a) - Q_h^*(x, a)) \\
&\quad + \sum_{i=1}^t \alpha_t^i \cdot (R(x, a) + \tilde{V}_{k_i, h+1}(x_{k_i, h+1}) - Q_h^*(x, a)) + \sum_{i=0}^t \alpha_t^i \beta_i \\
&= \alpha_t^0 \cdot (\tilde{Q}_{1,h}(x, a) - Q_h^*(x, a)) + \sum_{i=1}^t \alpha_t^i \cdot (\tilde{V}_{k_i, h+1}(x_{k_i, h+1}) - V^*(x_{k_i, h+1})) \\
&\quad + \sum_{i=1}^t \alpha_t^i \cdot (V^*(x_{k_i, h+1}) - (\mathbb{P}V^*)(x, a)) + \sum_{i=0}^t \alpha_t^i \beta_i, \tag{2}
\end{aligned}$$

where in the last equality we have used the Bellman Optimality Equation $Q_h^*(x, a) = R(x, a) + (\mathbb{P}V_{h+1}^*)(x, a)$.

Lemma 8. α_t^i 's satisfy the following properties (Lemma 4.1 of [2]):

- (a) $\alpha_t^0 = 0$ and $\frac{1}{\sqrt{t}} \leq \sum_{i=1}^t \frac{\alpha_t^i}{\sqrt{i}} \leq \frac{2}{\sqrt{t}}$ for every $t \geq 1$,
- (b) $\sum_{i=1}^t (\alpha_t^i)^2 \leq \frac{2H}{t}$ for every $t \geq 1$,
- (c) $\sum_{t=i}^{+\infty} \alpha_t^i = 1 + \frac{1}{H}$ for every $i \geq 1$.

By event \mathcal{E}_1 and Lemma 8(b), we have $|\sum_{i=1}^t \alpha_t^i \cdot (V^*(x_{k_i, h+1}) - (\mathbb{P}V^*)(x, a))| \leq c_1 \sqrt{2} \cdot \sqrt{\frac{H^3 \ln(SAH/\delta)}{t}}$. Further by Lemma 8(a), we have $\sum_{i=0}^t \alpha_t^i \beta_i \geq c_1 \sqrt{2} \cdot \sqrt{\frac{H^3 \ln(SAH/\delta)}{t}} \geq |\sum_{i=1}^t \alpha_t^i \cdot (V^*(x_{k_i, h+1}) - (\mathbb{P}V^*)(x, a))|$. Using mathematical induction, we are able to show $\tilde{Q}_{k,h}(x, a) - Q_h^*(x, a) \geq 0$ and conclude the proof of this lemma. \square

With optimistic guarantee, we can give a direct upper bound of \mathcal{R}_K^π . Note that

$$\begin{aligned}
\mathcal{R}_K^\pi &= \mathbb{E} \left[\sum_{k=1}^K (V_1^* - V_1^{\pi_k})(x_{k,1}) \right] \\
&\leq \mathbb{E} \left[\sum_{k=1}^K (\tilde{V}_{k,1} - V_1^{\pi_k})(x_{k,1}) \right] \\
&= \sum_{k=1}^K \mathbb{E} \tilde{\delta}_{k,1}.
\end{aligned}$$

where we have defined $\tilde{\delta}_{k,h} \stackrel{\text{def}}{=} (\tilde{V}_{k,h} - V_h^{\pi_k})(x_{k,h})$.

The next step idea is to rewrite $\tilde{\delta}_{k,h}$ using $\tilde{\delta}_{k,h+1}$ and then use recursion to calculate an upper bound of $\sum_{k=1}^K \tilde{\delta}_{k,h}$. We first show

Lemma 9. When $n_{k,h}(x_{k,h}, a_{k,h}) > 0$, it holds that

$$\begin{aligned}
\tilde{\delta}_{k,h} &\leq \sum_{i=1}^t \alpha_t^i \cdot (\tilde{V}_{k_i, h+1}(x_{k_i, h+1}) - V^*(x_{k_i, h+1})) - (\tilde{V}_{k, h+1} - V_{h+1}^*)(x_{k, h+1}) + \tilde{\delta}_{k, h+1} \\
&\quad + 2c_1 \sqrt{2} \cdot \sqrt{\frac{H^3 \ln(SAH/\delta)}{t}} + (\mathbb{P}(V_{h+1}^* - V_{h+1}^{\pi_k}))(x_{k,h}, a_{k,h}) - (V_{h+1}^* - V_{h+1}^{\pi_k})(x_{k,h+1}).
\end{aligned}$$

Proof. Note that

$$\begin{aligned}
\tilde{\delta}_{k,h} &= \tilde{V}_{k,h}(x_{k,h}) - V_h^{\pi_k}(x_{k,h}) \\
&= \tilde{Q}_{k,h}(x_{k,h}, a_{k,h}) - Q_h^{\pi_k}(x_{k,h}, a_{k,h}) \\
&= \tilde{Q}_{k,h}(x_{k,h}, a_{k,h}) - Q_h^*(x_{k,h}, a_{k,h}) + Q_h^*(x_{k,h}, a_{k,h}) - Q_h^{\pi_k}(x_{k,h}, a_{k,h}).
\end{aligned} \tag{3}$$

Plugging (2) and $\alpha_t^0 = 0$ from Lemma 8(a) in (3), we obtain

$$\begin{aligned}
\tilde{\delta}_{k,h} &= \sum_{i=1}^t \alpha_t^i \cdot (\tilde{V}_{k_i,h+1}(x_{k_i,h+1}) - V^*(x_{k_i,h+1})) \\
&\quad + \sum_{i=1}^t \alpha_t^i \cdot (V^*(x_{k_i,h+1}) - (\mathbb{P}V^*)(x, a)) + \sum_{i=0}^t \alpha_t^i \beta_i + Q_h^*(x_{k,h}, a_{k,h}) - Q_h^{\pi_k}(x_{k,h}, a_{k,h}) \\
&\leq \sum_{i=1}^t \alpha_t^i \cdot (\tilde{V}_{k_i,h+1}(x_{k_i,h+1}) - V^*(x_{k_i,h+1})) + \underbrace{Q_h^*(x_{k,h}, a_{k,h}) - Q_h^{\pi_k}(x_{k,h}, a_{k,h})}_{(I)} \\
&\quad + 2c_1\sqrt{2} \cdot \sqrt{\frac{H^3 \ln(SAH/\delta)}{t}},
\end{aligned} \tag{4}$$

where we have used $\sum_{i=1}^t \alpha_t^i \cdot (V^*(x_{k_i,h+1}) - (\mathbb{P}V^*)(x, a)) \leq c_1\sqrt{2} \cdot \sqrt{\frac{H^3 \ln(SAH/\delta)}{t}}$ and $\sum_{i=0}^t \alpha_t^i \beta_i \leq c_1\sqrt{2} \cdot \sqrt{\frac{H^3 \ln(SAH/\delta)}{t}}$. Both of them have been proved in the analysis of Lemma 6.

We next take care of (I) and try to expand it. Notice that

$$\begin{aligned}
(I) &= (\mathbb{P}(V_{h+1}^* - V_{h+1}^{\pi_k}))(x_{k,h}, a_{k,h}) \\
&= (\mathbb{P}(V_{h+1}^* - V_{h+1}^{\pi_k}))(x_{k,h}, a_{k,h}) - (V_{h+1}^* - V_{h+1}^{\pi_k})(x_{k,h+1}) + \tilde{\delta}_{k,h+1} - (\tilde{V}_{k,h+1} - V_{h+1}^*)(x_{k,h+1}).
\end{aligned} \tag{5}$$

The intuition to expand (I) in this way is that the expectation of $(\mathbb{P}(V_{h+1}^* - V_{h+1}^{\pi_k}))(x_{k,h}, a_{k,h}) - (V_{h+1}^* - V_{h+1}^{\pi_k})(x_{k,h+1})$ equals 0 conditioning on the history \mathcal{H}_k and $(x_{k,1}, a_{k,1}, \dots, x_{k,h})$.

Finally, plugging (5) back into (4), we prove this lemma. \square

Corollary 10.

$$\begin{aligned}
\sum_{k=1}^K \tilde{\delta}_{k,h} &\leq SAH + \sum_{k=1}^K \sum_{i=1}^t \alpha_t^i \cdot (\tilde{V}_{k_i,h+1}(x_{k_i,h+1}) - V^*(x_{k_i,h+1})) - \sum_{k=1}^K (\tilde{V}_{k,h+1} - V_{h+1}^*)(x_{k,h+1}) + \sum_{k=1}^K \tilde{\delta}_{k,h+1} \\
&\quad + \sum_{k=1}^K 2c_1\sqrt{2} \cdot \sqrt{\frac{H^3 \ln(SAH/\delta)}{t}} + \sum_{k=1}^K \left((\mathbb{P}(V_{h+1}^* - V_{h+1}^{\pi_k}))(x_{k,h}, a_{k,h}) - (V_{h+1}^* - V_{h+1}^{\pi_k})(x_{k,h+1}) \right).
\end{aligned}$$

Proof. When $n_{k,h}(x_{k,h}, a_{k,h}) = 0$, we apply the naive upper bound i.e., $\tilde{\delta}_{k,h} \leq H$. Let \mathcal{K} be the set of k 's such that $n_{k,h}(x_{k,h}, a_{k,h}) = 0$. Hence $|\{(x_{k,h}, a_{k,h}) : k \in \mathcal{K}\}| \leq SA$. So $\sum_{k \in \mathcal{K}} \tilde{\delta}_{k,h} \leq SAH$. Together with Lemma 9, we prove this corollary. \square

We next focus on bounding

$$\sum_{k=1}^K \sum_{i=1}^t \alpha_t^i \cdot (\tilde{V}_{k_i, h+1}(x_{k_i, h+1}) - V^*(x_{k_i, h+1})) - \sum_{k=1}^K (\tilde{V}_{k, h+1} - V_{h+1}^*)(x_{k, h+1}) \quad (6)$$

in Corollary 10 and show

Lemma 11.

$$(6) \leq \frac{1}{H} \cdot \left(\sum_{k=1}^K (\tilde{V}_{k, h+1} - V_{h+1}^*)(x_{k, h+1}) \right).$$

Proof. Rewrite (6) we have

$$(6) = \sum_{t=(i+1)}^K \alpha_t^i \cdot \left(\sum_{k=1}^K (\tilde{V}_{k, h+1} - V_{h+1}^*)(x_{k, h+1}) \right) - \left(\sum_{k=1}^K (\tilde{V}_{k, h+1} - V_{h+1}^*)(x_{k, h+1}) \right).$$

By Lemma 5, we have $\sum_{t=(i+1)}^K \alpha_t^i \leq 1 + \frac{1}{H}$. Using aforementioned inequality, we are able to show this lemma. \square

By Corollary 10, Lemma 11 and the fact that $V_{h+1}^*(x) \geq V_{h+1}^{\pi_k}(x)$, we have

$$\begin{aligned} \sum_{k=1}^K \tilde{\delta}_{k, h} &\leq SAH + \left(1 + \frac{1}{H}\right) \cdot \sum_{k=1}^K \tilde{\delta}_{k, h+1} + \sum_{k=1}^K 2c_1 \sqrt{2} \cdot \sqrt{\frac{H^3 \ln(SAH/\delta)}{t}} \\ &\quad + \sum_{k=1}^K \left((\mathbb{P}(V_{h+1}^* - V_{h+1}^{\pi_k}))(x_{k, h}, a_{k, h}) - (V_{h+1}^* - V_{h+1}^{\pi_k})(x_{k, h+1}) \right). \end{aligned}$$

Hence by recursion, we further obtain

$$\begin{aligned} \sum_{k=1}^K \tilde{\delta}_{1, h} &\leq \left(1 + \frac{1}{H}\right)^H \cdot \left(SAH^2 + \sum_{h=1}^H \sum_{k=1}^K 2c_1 \sqrt{2} \cdot \sqrt{\frac{H^3 \ln(SAH/\delta)}{t}} \right. \\ &\quad \left. + \sum_{h=1}^H \sum_{k=1}^K \left((\mathbb{P}(V_{h+1}^* - V_{h+1}^{\pi_k}))(x_{k, h}, a_{k, h}) - (V_{h+1}^* - V_{h+1}^{\pi_k})(x_{k, h+1}) \right) \right) \\ &\lesssim SAH^2 + \underbrace{\sum_{h=1}^H \sum_{k=1}^K \sqrt{\frac{H^3 \ln(SAH/\delta)}{t}}}_{(*)} \\ &\quad + \underbrace{\sum_{h=1}^H \sum_{k=1}^K \left((\mathbb{P}(V_{h+1}^* - V_{h+1}^{\pi_k}))(x_{k, h}, a_{k, h}) - (V_{h+1}^* - V_{h+1}^{\pi_k})(x_{k, h+1}) \right)}_{(**)}. \end{aligned} \quad (7)$$

Rewrite (*), we obtain

$$(*) = \sqrt{H^3 \ln(SAH/\delta)} \cdot \sum_{h=1}^H \sum_{(x, a)} \sum_{t=1}^{n_{K, h}(x, a)} \sqrt{\frac{1}{t}}.$$

Further applying $\sum_{i=1}^t \frac{1}{i} \leq 2\sqrt{t}$ and Cauchy–Schwarz inequality, we have

$$\begin{aligned}
(*) &\lesssim \sqrt{H^3 \ln(SAH/\delta)} \cdot \sum_{(x,a)} \sum_{h=1}^H \sqrt{n_{K,h}(x,a)} \\
&\leq \sqrt{H^3 \ln(SAH/\delta)} \cdot \sum_{(x,a)} \sqrt{H \cdot n_K(x,a)} \\
&= \mathcal{O}(H^2 \sqrt{SAT \ln(SAH/\delta)})
\end{aligned} \tag{8}$$

Let $\mathcal{E}_2 \stackrel{\text{def}}{=} \{(**) \leq c_2 \sqrt{TH^2 \ln(\delta^{-1})}\}$, where c_2 is a constant which will be defined later. By Azuma's inequality, we have there exists a constant c_2 such that $\mathbf{Pr}(\mathcal{E}_2) \geq 1 - \delta/2$. According to event \mathcal{E}_2 , it holds that

$$(**) \leq c_2 \sqrt{TH^2 \ln(\delta^{-1})}. \tag{9}$$

Plugging (8) and (9) back into (7), we prove this theorem. \square

5 Probability Tools

Assuming $X_0 = 0$, a martingale (X_1, \dots, X_t) is \mathbf{c} -Lipschitz if $|X_i - X_{i-1}| \leq c_i$ where $\mathbf{c} = (c_1, \dots, c_t)$. The following lemma states Azuma's inequality.

Lemma 12. ([1]) *If a martingale (X_1, \dots, X_t) is \mathbf{c} -Lipschitz, define $X = X_t$, then for every $\epsilon \geq 0$, it holds that*

$$\Pr(|X - \mathbb{E}X| \geq \epsilon) \leq 2 \exp \left(-\frac{\epsilon^2}{2 \sum_{i=1}^t c_i^2} \right),$$

where $\mathbf{c} = (c_1, \dots, c_t)$.

References

- [1] Fan Chung and Linyuan Lu. Concentration inequalities and martingale inequalities: a survey. *Internet Mathematics*, 3(1):79–127, 2006.
- [2] Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is q-learning provably efficient? In *NeurIPS*, pages 4863–4873, 2018.