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

3 Algorithm

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

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

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

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

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 hth time step for the ith 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

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

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 $\mathbf{Pr}(\mathcal{E}_1) \geq 1 - \delta/2$.

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 $\widetilde{\mathcal{O}}\left(H^2\sqrt{SAT}\right)$.*

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 ith time step.*

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

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

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

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

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

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

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

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

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

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

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

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

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* $\frac{1}{\sqrt{t}} \le \sum_{i=1}^t \frac{\alpha_t^i}{\sqrt{i}} \le \frac{2}{\sqrt{t}}$ *for every* $t \ge 1$ *,*
(b) $\sum_{i=1}^t (\alpha_t^i)^2 \le \frac{2H}{t}$ *for every* $t \ge 1$ *,*
(c) $\sum_{t=i}^{+\infty} \alpha_t^i = 1 + \frac{1}{H}$ *for every* $i \ge 1$ *.*

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

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

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

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

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

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

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

Proof. Note that

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

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

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

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

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

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

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

Corollary 10.

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

Proof. When $n_{k,h}(x_{k,h}, a_{k,h}) = 0$, we apply the naive upper bound i.e., $\widetilde{\delta}_{k,h} \leq H$. 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 $\sum_{k \in \mathcal{K}} \widetilde{\delta}_{k,h} \leq SAH$. Together with Lemma 9, we prove this corollary. □

We next focus on bounding

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

in Corollary 10 and show

Lemma 11.

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

Proof. Rewrite (6) we have

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

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. □

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

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

Hence by recursion, we further obtain

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

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 $(X_1, \ldots, X_t)$ is $\boldsymbol{c}$-Lipschitz if $|X_i - X_{i-1}| \leq c_i$ where $\boldsymbol{c} = (c_1, \ldots, c_t)$. The following lemma states Azuma's inequality.

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

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

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

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.