Notes of [1]

Chao Tao

Jan. 31, 2020

1 Problem Setup

There is a *tabular episodic* MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathbb{P}, R, H, s_1)$ where we assume the reward function R is bounded within $[0, 1]$ and for simplicity we also assume R is *deterministic*. In other words, only the transition probability $\mathbb{P}$ is *unknown*. We want to find a policy such that the *expected* 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

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

where V denotes the value function and the initial state $x_{k,1}$ can be either randomized or *adversarial*.

Remark 1. *There exists an optimal policy which is Markov and deterministic (may depend on time $t \in [H]$).*

2 Notations and Definitions

$[n]$	$\{1, 2, \ldots, n\}$
$\mathcal{A}$	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
$\pi = (\pi_1, \ldots, \pi_K)$	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 at the hth time step of the kth episode
$\mathcal{H}_k$	history before the kth episode $(x_{1,1}, a_{1,1}, \ldots, x_{1,H+1}, \ldots, x_{k-1,1}, a_{k-1,1} \ldots, x_{k-1,H+1})$
$n_k : \mathcal{S} \times \mathcal{A} \to \mathbb{N}$	number of hits of state-action pair before the kth episode
$n_k(y \mid x, a)$	number of hits of state y when taking action a at state x before the kth episode
$\widehat{\mathbb{P}}_k$	empirical transition probability using $\mathcal{H}_k$
$\widetilde{Q}_{k,h}$	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: UCBVI-CH ([1])

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 $a_{k,h} = \mathrm{argmax}_{a\in\mathcal{A}} \widetilde{Q}_{k,h}(x_{k,h}, a)$

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

initialization: $\widetilde{Q}_{k,H+1}(x,a) = \widetilde{V}_{k,H+1}(x,a) = 0$ and $\widehat{\mathbb{P}}_k(y \mid x,a) = \frac{n_k(y \mid x,a)}{n_k(x,a)}$ for every $(x,a,y) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$
for *step* $h = H$ *downto* 1 **do**
  for *every state-action pair* (x,a) **do**
    if $(x,a) = (x_{k-1,h}, a_{k-1,h})$ **then**
      let $b_k(x,a) = c_1 H \sqrt{\frac{\ln(SAT/\delta)}{n_k(x,a)}}$
      $\widetilde{Q}_{k,h}(x,a) = R(x,a) + (\widehat{\mathbb{P}}_k \widetilde{V}_{k,h+1})(x,a) + b_k(x,a)$
    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 when event $\mathcal{E}_1$ is defined.

Remark 2. *Algorithm 1 needs to know the horizon* T.

4 Proofs

4.1 Favorable Events

4.1.1 $\mathcal{E}_1$

Given any $(x,a,t) \in \mathcal{S} \times \mathcal{A} \times [T]$, define *i.i.d.* random variables $X_1^{x,a}, \ldots, X_t^{x,a}$ following the distribution $\mathbb{P}(x,a)$. Let

$$\mathcal{E}_1 \stackrel{\text{def}}{=} \left\{ \forall (h,x,a,t) \in [H] \times \mathcal{S} \times \mathcal{A} \times [T], \left| \frac{\sum_{i=1}^t V_h^*(X_i^{x,a})}{t} - \sum_{y\in\mathcal{S}} \mathbb{P}(y \mid x,a) V_h^*(y) \right| \le c_1 H \sqrt{\frac{\ln(SAT/\delta)}{t}} \right\},$$

where c_1 is a constant which will be defined later.

By Hoeffding's inequality (Lemma 12) and a union bound, there exists a constant c_1 such that $\mathbf{Pr}(\mathcal{E}_1) \ge 1 - \delta/4$.

4.1.2 $\mathcal{E}_2$

Given any $(x, a, y, t) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times [T]$, suppose *i.i.d.* random variables $X_1^{x,a,y}, \ldots, X_t^{x,a,y}$ follow the Bernoulli distribution $\mathcal{B}(\mathbb{P}(y \mid x, a))$. Let

$$\mathcal{E}_2 \stackrel{\text{def}}{=} \Bigg\{ \forall (x, a, y, t) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times [T] \text{ satisfying } \mathbb{P}(y \mid x, a)t \geq c_2 H^2 \ln(SAT/\delta), \\ \frac{\sum_{i=1}^t X_i^{x,a,y}}{t} \leq (1 + 1/H)\mathbb{P}(y \mid x, a) \Bigg\},$$

where c_2 is a constant which will be defined later.

By Multiplicative Chernoff bound (Lemma 13) and a union bound, there exists a constant c_2 such that $\mathbf{Pr}(\mathcal{E}_2) \geq 1 - \delta/4$.

4.1.3 $\mathcal{E}_3$

Given any $(x, a, y, t) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times [T]$, suppose *i.i.d.* random variables $X_1^{x,a,y}, \ldots, X_t^{x,a,y}$ follow the Bernoulli distribution $\mathcal{B}(\mathbb{P}(y \mid x, a))$. Let

$$\mathcal{E}_3 \stackrel{\text{def}}{=} \Bigg\{ \forall (x, a, y, t) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times [T] \text{ satisfying } \mathbb{P}(y \mid x, a)t \leq c_2 H^2 \ln(SAT/\delta), \\ \frac{\sum_{i=1}^t X_i^{x,a,y}}{t} \leq \frac{c_3 H \ln(SAT/\delta)}{t} \Bigg\},$$

where c_3 is a constant which will be defined later.

By Bernstein's inequality (Lemma 14) and a union bound, there exists a constant c_3 such that $\mathbf{Pr}(\mathcal{E}_3) \geq 1 - \delta/4$.

4.2 Main Theorem

Theorem 3. *With probability at least $1-\delta$, the regret incurred by Algorithm 1 is bounded by*

$$O\left(H\sqrt{SAT\ln(SAT/\delta)} + H^2S^2A\ln\left(\frac{T}{SA}\right)\ln(SAT/\delta)\right).$$

Remark 4. *When T is large, with probability at least $1-\delta$, the regret is bounded by $\widetilde{O}(H\sqrt{SAT})$.*

Corollary 5. *There exists an algorithm who does not need to know the horizon T and its expected regret is bounded by $\widetilde{O}(H\sqrt{SAT})$.*

Proof. We leverge doubling trick to design the new algorithm. Denote Algorithm 1 by $\mathbb{A}(\delta, T)$. Since we do not know the horizon, a good strategy is to guess. Specifically, we divide the whole time steps into several episodes and in episode i, run $\mathbb{A}(1/2^i, 2^i)$ for 2^i steps. The algorithm continues until the end of the horizon.

Now we analyze the regret. It is easy to see the expected regret of episode i is upper bounded by $\widetilde{O}(H\sqrt{SA2^i})$. Let I be the minimum integer such that $\sum_{i=1}^{I} 2^i \geq T$. And we have $2^I \leq T+2$. Hence the total expected regret is upper bounded by

$$\sum_{i=1}^{I}\widetilde{O}(H\sqrt{SA2^i}) = \widetilde{O}(H\sqrt{SA2^I}) = \widetilde{O}(H\sqrt{SAT}).$$

□

Remark 6. *The optimal upper bound is $\widetilde{O}(\sqrt{HSAT})$ [1]. And the lower bound is $\Omega(\sqrt{HSAT})$ [3].*

Proof. The following arguments are conditioned on event $\mathcal{E} \stackrel{\text{def}}{=} \mathcal{E}_1 \wedge \mathcal{E}_2 \wedge \mathcal{E}_3 \wedge \mathcal{E}_4$, where $\mathcal{E}_4$ is 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 $\widetilde{Q}_{k,h}$ is optimistic.

Lemma 7. *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 8. *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) and note that

$$\begin{aligned}\widetilde{Q}_{k,h}(x,a) - Q_h^*(x,a) &= (\widehat{\mathbb{P}}_k\widetilde{V}_{k,h+1})(x,a) - (\mathbb{P}V_{h+1}^*)(x,a) + b_k(x,a)\\ &= (\widehat{\mathbb{P}}_k(\widetilde{V}_{k,h+1} - V_{h+1}^*))(x,a) + ((\widehat{\mathbb{P}}_k - \mathbb{P})V_{h+1}^*)(x,a) + b_k(x,a)\end{aligned}$$

By event $\mathcal{E}_1$, we have $|(\widehat{\mathbb{P}}_k - \mathbb{P})V_{h+1}^*(x,a)| \leq b_k(x,a)$. Using mathematical induction, we prove this lemma. □

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

$$\begin{aligned}\mathcal{R}_K^\pi &= \sum_{k=1}^{K}(V_1^* - V_1^{\pi_k})(x_{k,1})\\ &\leq \sum_{k=1}^{K}(\widetilde{V}_{k,1} - V_1^{\pi_k})(x_{k,1})\\ &= \sum_{k=1}^{K}\widetilde{\delta}_{k,1}.\end{aligned}$$

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

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.

$$\widetilde{\delta}_{k,h} = ((\widehat{\mathbb{P}}_k - \mathbb{P})\widetilde{V}_{k,h+1})(x_{k,h}, a_{k,h}) + ((\mathbb{P}(\widetilde{V}_{k,h+1} - V_{h+1}^{\pi_k}))(x_{k,h}, a_{k,h}) - \widetilde{\delta}_{k,h+1}) + \widetilde{\delta}_{k,h+1} + b_k(x_{k,h}, a_{k,h})$$

The idea to write in this way is that the expectation of $(\mathbb{P}(\widetilde{V}_{k,h+1} - V_{h+1}^{\pi_k}))(x_{k,h}, a_{k,h}) - \widetilde{\delta}_{k,h+1}$ is 0 conditioned on history $\mathcal{H}_k, x_{k,1}, a_{k,1}, \ldots, x_{k,h}$.

Proof. Just note that

$$\begin{aligned}
\widetilde{\delta}_{k,h} &= \widetilde{V}_{k,h}(x_{k,h}) - V_h^{\pi_k}(x_{k,h}) \\
&= (\widehat{\mathbb{P}}_k \widetilde{V}_{k,h+1})(x_{k,h}, a_{k,h}) - (\mathbb{P}V_{h+1}^{\pi_k})(x_{k,h}, a_{k,h}) + b_k(x_{k,h}, a_{k,h}) \\
&= ((\widehat{\mathbb{P}}_k - \mathbb{P})\widetilde{V}_{k,h+1})(x_{k,h}, a_{k,h}) + ((\mathbb{P}(\widetilde{V}_{k,h+1} - V_{h+1}^{\pi_k}))(x_{k,h}, a_{k,h}) - \widetilde{\delta}_{k,h+1}) + \widetilde{\delta}_{k,h+1} + b_k(x_{k,h}, a_{k,h})
\end{aligned}$$

□

We next focus on bounding

$$((\widehat{\mathbb{P}}_k - \mathbb{P})\widetilde{V}_{k,h+1})(x_{k,h}, a_{k,h}) \tag{1}$$

and show

Lemma 10 (One Step Transition Probability Error)**.**

$$\begin{aligned}
(1) \leq \frac{1}{H}\widetilde{\delta}_{k,h+1} + c_1 H \sqrt{\frac{\ln(SAT/\delta)}{n_k(x_{k,h}, a_{k,h})}} \\
+ \frac{1}{H}\left((\mathbb{P}(\widetilde{V}_{k,h+1} - V_{h+1}^{\pi_k}))(x_{k,h}, a_{k,h}) - \widetilde{\delta}_{k,h+1}\right) + \frac{\max\{c_2, c_3\} H^2 S \ln(SAT/\delta)}{n_k(x_{k,h}, a_{k,h})}.
\end{aligned}$$

Remark 11. *There exists an easier way to bound (1) which leavages Hölder's inequality to derive*

$$(1) \leq \left\|(\widehat{\mathbb{P}}_k - \mathbb{P})(x_{k,h}, a_{k,h})\right\|_1 \cdot \left\|\widetilde{V}_{k,h+1}\right\|_\infty$$

and then uses the inequality in [4] to bound the ℓ_1-norm deviation of the transition probability. Using this method will lead to an extra $\sqrt{S}$ in the final conclusion.

Proof. Rewrite (1) we have

$$(1) = \underbrace{((\widehat{\mathbb{P}}_k - \mathbb{P})V_{h+1}^*)(x_{k,h}, a_{k,h})}_{(I)} + \underbrace{((\widehat{\mathbb{P}}_k - \mathbb{P})(\widetilde{V}_{k,h+1} - V_{h+1}^*))(x_{k,h}, a_{k,h})}_{(II)} \tag{2}$$

Consider (I) first. Note that

$$\begin{aligned}
(I) &= \sum_{y \in \mathcal{S}} \left(\widehat{\mathbb{P}}_k(y \mid x_{k,h}, a_{k,h}) - \mathbb{P}(y \mid x_{k,h}, a_{k,h})\right) V_{h+1}^*(y) \\
&= \left(\sum_{y \in \mathcal{S}} \widehat{\mathbb{P}}_k(y \mid x_{k,h}, a_{k,h}) V_{h+1}^*(y)\right) - \left(\sum_{y \in \mathcal{S}} \mathbb{P}(y \mid x_{k,h}, a_{k,h}) V_{h+1}^*(y)\right)
\end{aligned} \tag{3}$$

The first part of (3) can be seen as the empirical mean of $\sum_{y \in \mathcal{S}} \mathbb{P}(y \mid x_{k,h}, a_{k,h}) V^*_{h+1}(y)$ after $n_k(x_{k,h}, a_{k,h})$ trials. By event $\mathcal{E}_1$, we conclude that

$$|(I)| \le c_1 H \sqrt{\frac{\ln(SAT/\delta)}{1 \vee n_k(x_{k,h}, a_{k,h})}}. \tag{4}$$

We now take care of (II). Note that

$$(II) = \sum_{y \in \mathcal{S}} \left(\widehat{\mathbb{P}}_k(y \mid x_{k,h}, a_{k,h}) - \mathbb{P}(y \mid x_{k,h}, a_{k,h}) \right) (\widetilde{V}_{k,h+1} - V^*_{h+1})(y). \tag{5}$$

Let $\mathcal{S}'$ be the set of states such that

$$\mathbb{P}(y \mid x_{k,h}, a_{k,h})(1 \vee n_k(x_{k,h}, a_{k,h})) \ge c_2 H^2 \ln(SAT/\delta).$$

Rewrite (5) we get

$$\begin{aligned}
(II) \le \; & \frac{1}{H} \widetilde{\delta}_{k,h+1} \\
& + \underbrace{\sum_{y \in \mathcal{S}'} (\widehat{\mathbb{P}}_k(y \mid x_{k,h}, a_{k,h}) - \mathbb{P}(y \mid x_{k,h}, a_{k,h}))(\widetilde{V}_{k,h+1} - V^{\pi_k}_{h+1})(y) - \frac{1}{H} \widetilde{\delta}_{k,h+1}}_{(III)} \\
& + \underbrace{\sum_{y \in (\mathcal{S} - \mathcal{S}')} \left(\widehat{\mathbb{P}}_k(y \mid x_{k,h}, a_{k,h}) - \mathbb{P}(y \mid x_{k,h}, a_{k,h}) \right) (\widetilde{V}_{k,h+1} - V^*_{h+1})(y)}_{(IV)},
\end{aligned} \tag{6}$$

where we have used $V^{\pi_k}_{h+1}(y) \le V^*_{h+1}(y)$. Due to event $\mathcal{E}_2$, we have

$$\begin{aligned}
(III) &\le \frac{1}{H} \left(\sum_{y \in \mathcal{S}'} \mathbb{P}(y \mid x_{k,h}, a_{k,h})(\widetilde{V}_{k,h+1} - V^{\pi_k}_{h+1})(y) - \widetilde{\delta}_{k,h+1} \right) \\
&\le \frac{1}{H} \left((\mathbb{P}(\widetilde{V}_{k,h+1} - V^{\pi_k}_{h+1}))(x_{k,h}, a_{k,h}) - \widetilde{\delta}_{k,h+1} \right)
\end{aligned} \tag{7}$$

Next we upper bound (IV). By event $\mathcal{E}_3$ and plugging in inequality $\mathbb{P}(y \mid x_{k,h}, a_{k,h}) \le \frac{c_2 H^2 \ln(SAT/\delta)}{1 \vee n_k(x_{k,h}, a_{k,h})}$, we have

$$\begin{aligned}
(IV) &\le \frac{c_3 H S \ln(SAT/\delta)}{1 \vee n_k(x_{k,h}, a_{k,h})} + \frac{c_2 H^2 S \ln(SAT/\delta)}{1 \vee n_k(x_{k,h}, a_{k,h})} \\
&\le \frac{\max\{c_2, c_3\} H^2 S \ln(SAT/\delta)}{1 \vee n_k(x_{k,h}, a_{k,h})}
\end{aligned}$$

Plugging in upper bounds of $(I), (III), (IV)$ to (1), we prove this lemma. $\square$

Combining Lemma 9 and Lemma 10, we get

$$\begin{aligned}
\widetilde{\delta}_{k,h} \le \left(1 + \frac{1}{H}\right) \widetilde{\delta}_{k,h+1} + \left(1 + \frac{1}{H}\right) (\mathbb{P}(\widetilde{V}_{k,h+1} - V^{\pi_k}_{h+1})(x_{k,h}, a_{k,h}) - \widetilde{\delta}_{k,h+1}) \\
+ 2 c_1 H \sqrt{\frac{\ln(SAT/\delta)}{1 \vee n_k(x_{k,h}, a_{k,h})}} + \frac{\max\{c_2, c_3\} H^2 S \ln(SAT/\delta)}{1 \vee n_k(x_{k,h}, a_{k,h})}.
\end{aligned}$$

Hence

$$
\begin{aligned}
\sum_{k=1}^{K} \widetilde{\delta}_{k,1} \le & \left(1+\frac{1}{H}\right)^{H} \Bigg[\underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H} (\mathbb{P}(\widetilde{V}_{k,h+1} - V_{h+1}^{\pi_k})(x_{k,h},a_{k,h}) - \widetilde{\delta}_{k,h+1})}_{(*)} \\
& + 2c_1 H \underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H} \sqrt{\frac{\ln(SAT/\delta)}{1 \vee n_k(x_{k,h},a_{k,h})}}}_{(**)} + \max\{c_2,c_3\} \underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H} \frac{H^2 S \ln(SAT/\delta)}{1 \vee n_k(x_{k,h},a_{k,h})}}_{(***)} \Bigg] \\
& \lesssim H + (*) + H(**) + (***). \qquad (8)
\end{aligned}
$$

$(*)$ can be seen as a martingale with KH random variables and satisfies H-Lipschitz. By Azuma's inequality (Lemma 15), with probability at least $(1-\delta/4)$, it holds that

$$
|(*)| \lesssim H\sqrt{\ln(1/\delta)T}. \qquad (9)
$$

And this defines event $\mathcal{E}_4$. Let $\mathcal{K}$ be the set of (k,h)'s such that $n_{k,h}(x_{k,h},a_{k,h}) = 0$. Hence $|\mathcal{K}| \le SA$. Rewrite $(**)$, we have

$$
\begin{aligned}
(**) &\le SA\sqrt{\ln(SAT/\delta)} + \sum_{(x,a)\in\mathcal{S}\times\mathcal{A}} \sum_{t=1}^{n_K(x,a)} \sqrt{\frac{\ln(SAT/\delta)}{t}} \\
&\lesssim SA\sqrt{\ln(SAT/\delta)} + \sqrt{\ln(SAT/\delta)} \cdot \sum_{(x,a)\in\mathcal{S}\times\mathcal{A}} \sqrt{n_K(x,a)} \\
&\le SA\sqrt{\ln(SAT/\delta)} + \sqrt{SAT\ln(SAT/\delta)}, \qquad (10)
\end{aligned}
$$

where the last inequality is due to Cauchy–Schwarz inequality. Using a similar way, we get

$$
\begin{aligned}
(***) &\le SA\sqrt{\ln(SAT/\delta)} + \sum_{(x,a)\in\mathcal{S}\times\mathcal{A}} \sum_{t=1}^{n_K(x,a)} \frac{H^2 S\ln(SAT/\delta)}{t} \\
&\lesssim SA\sqrt{\ln(SAT/\delta)} + H^2 S\ln(SAT/\delta) \sum_{(x,a)\in\mathcal{S}\times\mathcal{A}} \ln(n_K(x,a)) \\
&\le H^2 S^2 A \ln\left(\frac{T}{SA}\right) \ln(SAT/\delta), \qquad (11)
\end{aligned}
$$

where the last inequality is due to Jensen's inequality applied to $\ln(\cdot)$ function. Putting back (9), (10), and (11) into (8), we prove this theorem. □

5 Probability Tools

The following lemma states Hoeffding's inequality.

Lemma 12. *Let $X_1, X_2, \ldots, X_t$ be independent random variables bounded by $[0, M]$. Let $X = \sum_{i=1}^t X_i$. For every $\epsilon \geq 0$, it holds that*

$$\mathbf{Pr}\left(|X - \mathbb{E}X| \geq \epsilon\right) \leq 2\exp\left(-\frac{2\epsilon^2}{M^2}\right).$$

The following lemma states a weak Multiplicative Chernoff bound.

Lemma 13. *Let $X_1, X_2, \ldots, X_t$ be independent random variables bounded by $[0, 1]$. Let $X = \sum_{i=1}^t X_i$. For every $\epsilon \in [0, 1]$, it holds that*

$$\mathbf{Pr}\left(X \geq (1+\epsilon)\mathbb{E}X\right) \leq \exp\left(-\frac{\epsilon^2 \mathbb{E}X}{3}\right).$$

The following lemma states Bernstein's inequality.

Lemma 14. *Let $X_1, X_2, \ldots, X_t$ be zero-mean independent random variables bounded by $[-M, M]$. Let $X = \sum_{i=1}^t X_i$. For every $\epsilon \geq 0$, it holds that*

$$\mathbf{Pr}\left(X > \epsilon\right) \leq \exp\left(-\frac{\frac{1}{2}\epsilon^2}{\sum_{i=1}^t \mathbb{E}[X_i^2] + \frac{1}{3}M\epsilon}\right).$$

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

Lemma 15. *([2]) 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] Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *ICML*, pages 263–272, 2017.

[2] Fan Chung and Linyuan Lu. Concentration inequalities and martingale inequalities: a survey. *Internet Mathematics*, 3(1):79–127, 2006.

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

[4] Tsachy Weissman, Erik Ordentlich, Gadiel Seroussi, Sergio Verdu, and Marcelo J Weinberger. Inequalities for the l1 deviation of the empirical distribution. *Hewlett-Packard Labs, Tech. Rep*, 2003.