

# Notes of [1]

Chao Tao

Jan. 31, 2020

## 1 Problem Setup

There is a *tabular episodic* MDP  $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathbb{P}, R, H)$  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*.

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

## 2 Notations and Definitions

$[n]$	$\{1, 2, \dots, n\}$
$\mathcal{A}$	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 $\pi_k$ is the policy in the $k$ th episode for $k \in [K]$
$Q_h^{\pi_k} : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$	$Q$ -value function of policy $\pi_k$ at time $h$
$V_h^{\pi_k} : \mathcal{S} \rightarrow \mathbb{R}$	value function of policy $\pi_k$ at time $h$
$Q_h^*$	$Q$ -value function of the optimal policy
$V_h^*$	value function of the optimal policy
$x_{k,1}$	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 step
$\mathcal{H}_k$	history before the $k$ th episode $(x_{1,1}, a_{1,1}, \dots, x_{1,H+1}, \dots, x_{k-1,1}, a_{k-1,1}, \dots, x_{k-1,H+1})$
$n_k : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{N}$	number of hits of state-action pair before the $k$ th episode
$n_k(y   x, a)$	number of hits of state $y$ when taking action $a$ at state $x$ before the $k$ th episode
$\hat{\mathbb{P}}_k$	empirical transition probability using $\mathcal{H}_k$
$\tilde{Q}_{k,h}$	optimistic estimate of the optimal $Q$ -value function starting from the $h$ th step of $k$ th episode
$\tilde{V}_{k,h}$	optimistic estimate of the optimal value function starting from the $h$ th step of $k$ th episode
$\rho$	an arbitrary transition probability
$V$	an arbitrary value function
$(\rho V)(x, a)$	$\sum_{y \in \mathcal{S}} \rho(y   x, a) V(y)$
$\mathcal{R}_K^\pi$	regret incurred by policy $\pi$

### 3 Algorithm

---

**Algorithm 1:** UCBVI-CH ([1])

---

```

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:  $\tilde{Q}_{k,H+1}(x, a) = \tilde{V}_{k,H+1}(x, a) = 0$  and  $\hat{\mathbb{P}}_k(y | x, a) = \frac{n_k(y | x, a)}{n_k(x, a)}$  for every
    $(x, a, y) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$ 
2 for step  $h = H$  downto 1 do
3   for every state-action pair  $(x, a)$  do
4     if  $(x, a) = \text{then}$ 
5       let  $b_k(x, a) = c_1 H \sqrt{\frac{\ln(SAT/\delta)}{n_k(x, a)}}$ 
6        $\tilde{Q}_{k,h}(x, a) = R(x, a) + (\hat{\mathbb{P}}_k \tilde{V}_{k,h+1})(x, a) + b_k(x, a)$ 
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 when event  $\mathcal{E}_1$  is defined.

### 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  $X_1^{x,a}, \dots, X_t^{x,a}$  following the distribution  $\mathbb{P}(x, a)$ . Let

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

where  $c_1$  is a constant which will be defined later.

By Hoeffding's inequality (Lemma 10) and a union bound, there exists a constant  $c_1$  such that  $\mathbf{Pr}(\mathcal{E}_1) \geq 1 - \delta/4$ .

#### 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  $X_1^{x,a,y}, \dots, X_t^{x,a,y}$  follow the Bernoulli distribution  $\mathcal{B}(\mathbb{P}(y \mid x, a))$ . Let

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

where  $c_2$  is a constant which will be defined later.

By Multiplicative Chernoff bound (Lemma 11) and a union bound, there exists a constant  $c_2$  such that  $\Pr(\mathcal{E}_2) \geq 1 - \delta/4$ .

#### 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  $X_1^{x,a,y}, \dots, X_t^{x,a,y}$  follow the Bernoulli distribution  $\mathcal{B}(\mathbb{P}(y \mid x, a))$ . Let

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

where  $c_3$  is a constant which will be defined later.

By Bernstein's inequality (Lemma 12) and a union bound, there exists a constant  $c_3$  such that  $\Pr(\mathcal{E}_3) \geq 1 - \delta/4$ .

## 4.2 Main Theorem

**Theorem 2.** *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 3.** *When  $T$  is large, the regret is bounded by  $\tilde{O}(H\sqrt{SAT})$ .*

**Remark 4.** *The optimal upper bound is  $\tilde{O}(\sqrt{HSAT})$  [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  $\pi = (\pi_1, \dots, \pi_K)$  to represent Algorithm 1.

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

**Lemma 5.** *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 6.** *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  $(k, h, x, a)$  and note that

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

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

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

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

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  $\tilde{\delta}_{1,h}$ . We first show

**Lemma 7.**

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

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

*Proof.* Just note that

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

□

We next focus on bounding

$$((\hat{\mathbb{P}}_k - \mathbb{P}) \tilde{V}_{k,h+1})(x_{k,h}, a_{k,h}) \quad (1)$$

and show

**Lemma 8** (One Step Transition Probability Error).

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

**Remark 9.** There exists an easier way to bound (1) which just rewrite

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

and then uses the inequality in [4] to bound the  $\ell_1$ -norm deviation of the transition probability. Using this method will lead to an extra  $\sqrt{S}$  in the final regret.

*Proof.* Rewrite (1) we have

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

Consider (I) first. Note that

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

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

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

We now take care of (II). Note that

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

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

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

Rewrite (5) we get

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

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

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

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

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

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

Combining Lemma 7 and Lemma 8, we get

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

Hence

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

(\*) can be seen as a martingale with  $(K-1)H$  r.v.'s and each r.v. is bounded by  $[-H, H]$ . By Azuma's inequality (Lemma 13), with probability at least  $(1 - \delta/4)$ , it holds that

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

And this defines event  $\mathcal{E}_4$ . Rewrite (\*\*), we have

$$(**) \leq \sum_{(x,a) \in \mathcal{S} \times \mathcal{A}} \sum_{t=1}^{n_K(x,a)} \sqrt{\frac{\ln(SAT/\delta)}{t}} \lesssim \sqrt{\ln(SAT/\delta)} \cdot \sum_{(x,a) \in \mathcal{S} \times \mathcal{A}} \sqrt{n_K(x,a)} \leq \sqrt{SAT \ln(SAT/\delta)}, \tag{10}$$

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

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

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.  $\square$



## 5 Probability Tools

The following lemma states Hoeffding's inequality.

**Lemma 10.** *Let  $X_1, X_2, \dots, X_t$  be independent random variables bounded by  $[0, M]$ . Let  $X = \sum_{i=1}^t X_i$ . For every  $\epsilon \geq 0$ , it holds that*

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

The following lemma states a weak Multiplicative Chernoff bound.

**Lemma 11.** *Let  $X_1, X_2, \dots, X_t$  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*

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

The following lemma states Bernstein's inequality.

**Lemma 12.** *Let  $X_1, X_2, \dots, X_t$  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*

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

Assuming  $X_0 = 0$ , a martingale  $(X_1, \dots, X_t)$  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 13.** ([2]) *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] 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  $\ell_1$  deviation of the empirical distribution. *Hewlett-Packard Labs, Tech. Rep*, 2003.