Supplement to ODM report

Krishna Acharya

December 2021

Recall the definitions, results for 2^{nd} Price auction We are considering a bidding model in which:

- We assume true values of ad slot and bids for ad slots $\in [0, 1]$.
- gender $\theta = \{m, f\}$, Absolute parity K
- State $s = (k, \theta)$ where $k \in \{-K \dots K\}$ and $\theta \in \{m, f\}$. The state space is denoted by $S = \{-K \dots K\} \times \{m, f\}$. θ_s refers to the gender in state s.
- p_m is the probability that the user is male, $p_f := 1 - p_m$ is probability that the user is female.
- immediate reward $R(\theta, b)$ only depends on gender and action(bid placed). $R(\theta, b)$ follows some probability distribution on $[-1, 1]$, we also have an analytic form for its expected value.
- When you bid b:
  - Your probability of winning the auction is $P_{\text{win},\theta}(b)$.
  - Expected reward for bidding b when gender is θ, $\bar{R}(\theta, b) := \mathbb{E}[R(\theta, b)] = (v_\theta - b)P_{\text{win},\theta}(b) + \int_0^b P_{win,\theta}(u)du$
- Total reward is $\sum_{t=1}^{T} R(\theta_t, a_t)$, gender at time $t := \theta_t$ and action taken at time $t := a_t$
- Regret after T steps, for the learning algorithm $\mathcal{L}$ starting at s, $\Delta(M, \mathcal{L}, s, T) := T\rho_M^* - \sum_{t=1}^{T} R(\theta_t, a_t)$, a_t is the amount to bid at step t (chosen on the fly by $\mathcal{L}$) . ρ_M^* represents the optimal average reward for the MDP M.

Regarding the bids:

1. if they lie in the set $\{b_1, b_2, \dots, b_l | b_i \in [0, 1] \forall i\}$ its the **"discrete bids" setting**
2. if they lie in the set $[0, 1]$ its the **"continous bids" setting**

About notation S, A refers to the state space and action space. With some abuse of notation they also refer to size of state space and size of action space. The usage will be clear by context.

0.1 Regret bound for discrete bids

UCRL2 is an online learning algorithm for finite state space and finite action space MDPs. We modify the UCRL2 algorithm for our MDPs structure and obtain a regret bound of $\tilde{O}(D\sqrt{AT})$ which is tighter than if we directly applied UCRL2 [1]. Here D refers to the diameter of the MDP M' i.e $D = D_{M'}$. $|S|$ is size of state space, $|A|$ is size of action space.

0.1.1 Description of the Algorithm

First lets see the modified UCRL2 algorithm. Initial state $s_1 = (0, \theta_1)$ where $\theta_1 \sim$ Bernoulli(p_f)[2]
the number of times (gender= θ and action $= a$) in episode k is denoted by $v_k(\theta, a)$

Algorithm 1 UCRL 2 adapted

Input: Confidence parameter $\delta \in (0, 1)$,
Initialize: Set $t := 1$, s_1 as defined earlier
for episodes $k = 1, 2 \ldots$ **do**
Initialize episode k:
1. Set start time of episode $t_k = t$
2. $\forall(\theta, a) \in \{m, f\} \times A,\ v_k(\theta, a) := 0$
Also $N_k(\theta, a) := \#\{\tau < t_k : \theta_\tau = \theta, a_\tau = a\}$
3. $R_k(\theta, a) := \sum_{\tau=1}^{t_k-1} r_\tau \mathbb{1}_{\theta_\tau=\theta, a_\tau=a}$
$P_{win,k}(\theta, a) := \#\{\tau < t_k : \theta = \theta, a_t = a$ & won auction$\}$
Compute estimates $\hat{r}_k(\theta, a) := \frac{R_k(\theta,a)}{\max\{1,N_k(\theta,a)\}}$ $\hat{p}_{win,k}(\theta, a) := \frac{P_{win,k}(\theta,a)}{\max\{1,N_k(\theta,a)\}}$
4. **Compute Policy** $\tilde{\pi}_k$ that is average reward optimal among all $\mathcal{M}_k$ ▷ *Extended value iteration*
5. **Execute policy** $\tilde{\pi}_k$
while $v_k(\theta_t, \tilde{\pi}_k(s_t)) < \max\{1, N_k(\theta_t, \tilde{\pi}_k(s_t))\}$ **do**
Action $a_t = \tilde{\pi}_k(s_t)$, obtain reward r_t and observe next state s_{t+1}
$v_k(\theta_t, a_t) = v_k(\theta_t, a_t) + 1$
t := t+1
end while
end for

$\mathcal{M}_k$ is defined as the set of all MDPs with probability of winning $\tilde{p}_{win}(\theta, a)$ close to $\hat{p}_{win,k}(\theta, a)$ and mean reward $\tilde{r}(\theta, a)$ close to $\hat{r}_k(\theta, a)$, Quantitatively the "closeness" is as follows, we must have $\forall\theta, \forall a$:

$$|\tilde{r}(\theta, a) - \hat{r}_k(\theta, a)| \leq d'(\theta, a) = \sqrt{\frac{c_1 \log(c_1' A t_k/\delta)}{\max\{1, N_k(\theta, a)\}}} \tag{1}$$

$$|\tilde{p}_{win}(\theta, a) - \hat{p}_{win,k}(\theta, a)| \leq d(\theta, a) = \sqrt{\frac{c_2 \log(c_2' A t_k/\delta)}{\max\{1, N_k(\theta, a)\}}} \tag{2}$$

[3]

In the above $c_1 = 14, c_1' = 2$, $c_2 = 7/2, c_2' = 2$, these constants are picked for ease of analysis in the proof of lemma 3.

[1] which has a regret bound of $\tilde{O}(DS\sqrt{AT})$
[2] $\theta = 1$ with probability $p_f, \theta = 0$ with probability $p_m, p_m + p_f = 1$
[3] See the relation between p_{win} and $p(.|s, a)$ in Lemma 1, note this gets rid of the explicit l^1 norm condition in UCRL2

Also note how the mean reward for (state,action), transitions from (state,action) are defined:

$$\tilde{r}(s,a) \coloneqq \tilde{r}(\theta_s, a)$$

$$\tilde{p}(s'|s=(\mathit{diff},\theta),a) \coloneqq \begin{cases} p_m \tilde{p}_{win}(\theta,a) & \text{if } s' = (\mathit{diff}+(-1)^\theta, m) \\ p_f \tilde{p}_{win}(\theta,a) & \text{if } s' = (\mathit{diff}+(-1)^\theta, f) \\ p_m(1-\tilde{p}_{win}(\theta,a)) & \text{if } s' = (\mathit{diff}, m) \\ p_f(1-\tilde{p}_{win}(\theta,a)) & \text{if } s' = (\mathit{diff}, f) \\ 0 & \text{for any other } s' \end{cases} \tag{3}$$

Lemma 1. *If* $\forall(\theta,a)|\hat{p}_{win,k}(\theta,a) - p_{win}(\theta,a)| \leq \epsilon$ *then* $\forall(s,a)\ \|\hat{p}_k(.|s,a) - p(.|s,a)\|_1 \leq 2\epsilon$ [4]
Here $p_{win}(\theta,a)$ *is the "true" probability of winning the auction for user of type* θ *by bidding* a. $p(.|s,a)$ *is the corresponding transition probability vector.* $\hat{p}_k(.|s,a)$ *is the estimated transition probability vector, it uses* $\hat{p}_{win,k}(\theta,a)$ *as an estimate for auction win probability.*

Proof. For any non-edge state s $\|\hat{p}(.|s,a) - p(.|s,a)\|_1$ can be broken down into 4 terms corresponding to the four transitions, θ_s denotes the gender in state s.

$$\begin{aligned} &|p_m\Big(\hat{p}_{win,k}(\theta_s,a) - p_{win}(\theta_s,a)\Big)| + |p_f\Big(\hat{p}_{win,k}(\theta_s,a) - p_{win}(\theta_s,a)\Big)| + \\ &|p_m\Big((1-\hat{p}_{win,k}(\theta_s,a)) - (1-p_{win}(\theta_s,a))\Big)| + |p_f\Big((1-\hat{p}_{win,k}(\theta_s,a)) - (1-p_{win}(\theta_s,a))\Big)| \\ &= 2p_m|\hat{p}_{win,k}(\theta_s,a) - p_{win}(\theta_s,a)| + 2p_f|\hat{p}_{win,k}(\theta_s,a) - p_{win}(\theta_s,a)| = 2|\hat{p}_{win,k}(\theta_s,a) - p_{win}(\theta_s,a)| \end{aligned}$$

□

The main step in extended value iteration (see UCRL2 paper) is the following maximization, the second equation(4) is what it looks like for our MDP

$$u_{i+1}(s) = \max_{a\in A}\Big\{\tilde{r}(s,a) + \max_{p(.)\in \text{Polytope}}\Big\{\sum_{s'\in S} p(s')u_i(s')\Big\}\Big\}$$

$$u_{i+1}(s) = \max_{a\in A}\Big\{\tilde{r}(s,a) + \max_{\tilde{p}_{win}(\theta,a)\in \text{Polytope}}\Big\{\tilde{p}_{win}(\theta,a)\phi(s) + c(s)\Big\}\Big\} \tag{4}$$

Where the polytope is given by Eq(2)(which requires $\tilde{p}_{win}(\theta,a) \in [0,1]$ and to be within $d(\theta,a)$ of $\hat{p}_{win,k}(\theta,a)$). The inner maximization is easy to do for our MDP since it can be written as $\boxed{\tilde{p}_{win}(\theta,a)\psi(s) + c(s)}$

[4] For the edge states the difference is exactly zero, since we know p_m and p_f

$c(s), \psi(s)$ are terms we get by collecting $u_i(s')$ i.e the u_i values of the next 4 states from s.
So if $\psi(s) \geq 0$, set $\tilde{p}_{win}(\theta, a) = \min\{1, \hat{p}_{win,k}(\theta, a) + d(\theta, a)\}$
If $\psi(s) < 0$, set $\tilde{p}_{win}(\theta, a) = \max\{0, \hat{p}_{win,k}(\theta, a) - d(\theta, a)\}$
Also set $\tilde{r}(s, a) = \tilde{r}(\theta_s, a) = \hat{r}_k(\theta_s, a) + d'(\theta_s, a)$.

The following is theorem 7 from UCRL2

Theorem 2. *Let $\mathcal{M}$ be the set of all MDPs with state space S, action space A, transition probabilities $\tilde{p}(.|s,a)$ and mean rewards $\tilde{r}(s,a)$ that satisfy $\|\tilde{p}(.|s,a) - \hat{p}(.|s,a)\|_1 \leq d(s,a)$ and $|\tilde{r}(s,a) - \hat{r}(s,a)| \leq d'(s,a), \forall s, \forall a$. Where the probability distributions $\hat{p}(.|s,a)$, values $\hat{r}(s,a) \in [0,1]$ and $d(s,a) > 0, d'(s,a) \geq 0$. If $\mathcal{M}$ contains at least one communicating MDP, extended value iteration converges.*
Further by stopping extended value iteration when $span(u_{i+1} - u_i) < \epsilon$, then the greedy policy wrt to u_i is ϵ-optimal

About convergence of extended value iteration for UCRL2 adapted If $|\tilde{p}_{win}(\theta, a) - \hat{p}_{win,k}(\theta, a)| \leq d(\theta, a)$ and $|\tilde{r}(\theta, a) - \hat{r}_k(\theta, a)| \leq d'(\theta, a)$ then all the conditions for the above theorem are satisfied.

So, we run extended value iteration at the start of episode k to obtain a $1/\sqrt{t_k}$ - optimal policy $\tilde{\pi}_k$

Steps to bound regret:

1. Splitting into episodes
2. Bound the regret when the true MDP $M \notin \mathcal{M}_k$
3. Consider the case when the true MDP $M \in \mathcal{M}_k$
4. Combine results from step 1,2,3

0.1.2 Splitting into Episodes

$\sum_{t=1}^{T} R(\theta_t, a_t)$ is a random variable, but it can be appropriately bounded using the Hoeffding inequality

- Immediate reward $R(s,a) \coloneqq R(\theta_s, a)$ i.e it only depends on gender of state s and action. Similarly the expected immediate reward $\bar{R}(s,a) \coloneqq \bar{R}(\theta_s, a)$
- Each $R(\theta, a)$ is a probability distribution on $[-1, 1]$, thus $\bar{R}(\theta, a) \in [-1, 1]$
- $v_k(\theta, a)$ denotes the number of times ($\theta_t = \theta$ and $a_t = a$) in episode k of UCRL2 adapted.
- $N(\theta, a)$ is the $\#(\theta, a)$ after T steps. therefore $\sum_{\theta,a} N(\theta, a) = T$
- $\sum_{k=1}^{m} v_k(\theta, a) = N(\theta, a)$, m is the total number of episodes.

Recap of Hoeffding inequality, for $S_n = X_1 + \dots X_n$ where each $X_i \in [a, b]$
$P(S_n \leq \mathbb{E}[S_n] - t) \leq \exp(-\frac{2t^2}{n(b-a)^2})$

$$P\left(\sum_{t=1}^{T} R(\theta_t, a_t) \leq \sum_{\theta,a} N(\theta, a)\bar{R}(\theta, a) - \sqrt{z_1 T log(\frac{z_2 T}{\delta})}\right)$$
$$\leq \exp\left(-\frac{z_1 log(z_2 T/\delta)}{2}\right) \tag{5}$$

For $z_1 = 5/2$ and $z_2 = 8$ the rhs is $\exp(-\frac{5}{4}log(8T/\delta)) = (\frac{\delta}{8T})^{5/4} < \frac{\delta}{12T^{5/4}}$

Therefore $T\rho^* - \sum_{t=1}^{T} R(\theta_t, a_t) < T\rho^* - \sum_{\theta,a} N(\theta, a)\bar{R}(\theta, a) + \sqrt{\frac{5}{2}T\log(\frac{8T}{\delta})}$ with probability atleast $1 - \frac{\delta}{12T^{5/4}}$

Therefore regret $\Delta(s_1, T) = T\rho^* - \sum_{t=1}^{T} R(\theta_t, a_t) \leq \boxed{\sum_{k=1}^{m} \Delta_k + \sqrt{\frac{5}{2}T\log(8T/\delta)}}$ wp atleast $1 - \frac{\delta}{12T^{5/4}}$.
Here $\Delta_k = \sum_{\theta,a} v_k(\theta, a)(\rho^* - \bar{R}(\theta, a))$

The boxed term can be rewritten as

$$\boxed{\sum_{k=1}^{m} \Delta_k \mathbb{1}_{M \notin \mathcal{M}_k} + \sum_{k=1}^{m} \Delta_k \mathbb{1}_{M \in \mathcal{M}_k} + \sqrt{\frac{5}{2}T\log(\frac{8T}{\delta})}} \tag{6}$$

0.1.3 Episodes with $M \notin \mathcal{M}_k$

Lets upper bound the regret for UCRL2 episodes in which the set of plausible MDPs $\mathcal{M}_k$ does not contain the true MDP M

Analysis The while loop stopping criteria ensures the following
$\sum_{\theta,a} v_k(\theta, a) \leq \sum_{\theta,a} N_k(\theta, a) = t_k - 1$ Note that the optimal average reward $\rho^* \leq 1$, $\bar{R}(\theta, a) \in [-1, 1]$ therefore $\rho^* - \bar{R}(\theta, a) \leq 2$. Thus we can build the following sequence of inequalities

$$\sum_{k=1}^{m} \Delta_k \mathbb{1}_{M \notin \mathcal{M}_k} \leq \sum_{k=1}^{m} \mathbb{1}_{M \notin \mathcal{M}_k} \sum_{\theta,a} v_k(\theta, a)(\rho^* - \bar{R}(\theta, a))$$
$$\leq 2\sum_{k=1}^{m} t_k \mathbb{1}_{M \notin \mathcal{M}_k} = 2\sum_{t=1}^{T} t \sum_{k=1}^{m} \mathbb{1}_{t_k = t, M \notin \mathcal{M}_k} \leq 2\sum_{t=1}^{T} t\mathbb{1}_{M \notin M(t)}$$

$$\leq 2 \overbrace{\sum_{t=1}^{\lfloor T^{1/4} \rfloor}}^{\leq \sqrt{T}} t\mathbb{1}_{M \notin M(t)} + 2\sum_{t=\lfloor T^{1/4} \rfloor + 1}^{T} t\mathbb{1}_{M \notin M(t)} \leq 2\sqrt{T} + 2 \overbrace{\sum_{t=\lfloor T^{1/4} \rfloor + 1}^{T}}^{\to 0 \text{ with high prob}} t\mathbb{1}_{M \notin M(t)}$$

The idea is if $P(M \notin M(t)) \leq \delta/t^n$ where n is a "large enough" positive integer, then over the course of $t = \lfloor T^{1/4} \rfloor + 1$ to T, we can ensure probability of $M \in M(t)$ is high, then indicator $\mathbb{1}_{M \notin M(t)} = 0$ with high

probability. Giving the final result that $\sum_{k=1}^{m} \Delta_k \mathbb{1}_{M \notin \mathcal{M}_k} \leq 2\sqrt{T}$

Lemma 3. $P(M \notin M(t)) \leq \frac{\delta}{15t^6}$

Proof. Recall Hoeffding $P(|\bar{X} - \mathbb{E}[\bar{X}]| \geq t) \leq 2\exp\left(\frac{-2n^2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right)$
$M(t)$ denotes the set of MDPs with prob of winning and mean reward, $\tilde{p}_{win}(\theta, a)$ and $\tilde{r}(\theta, a)$ in the sets defined by (2) and (1) (recall $c_1 = 14, c_1' = 2$). $M \notin M(t)$ if $\bar{R}(\theta, a), p_{win}(\theta, a)$ do not lie in (1), (2) for any (θ, a).
Using hoeffding inequality for $\bar{X} - E[\bar{X}]$ and the fact that $\delta \in (0, 1]$. Also note that the n below is a placeholder for $N(\theta, a)$

$$\because \sqrt{\frac{2}{n}\log\Big(\frac{120At^7}{\delta}\Big)} \leq \sqrt{\frac{14}{n}\log\Big(\frac{2At}{\delta}\Big)}$$
$$\therefore P\left(|\hat{r}(\theta, a) - \bar{R}(\theta, a)| \geq \sqrt{\frac{14}{n}\log\Big(\frac{2At}{\delta}\Big)}\right) \leq 2\exp\Big(-\frac{2n^2}{4n} \cdot \frac{2}{n} \cdot \log\Big(\frac{120At^7}{\delta}\Big)\Big) \leq \frac{\delta}{60At^7}$$

Similarly to lie outside set (2), (but here in the hoeffding inequaility the interval $[a_i, b_i] = [0, 1]$). Since $c_2 = 7/2$ and $c_2' = 2$

$$\because \sqrt{\frac{1}{2n}\log\Big(\frac{120At^7}{\delta}\Big)} \leq \sqrt{\frac{7}{2n}\log(2At/\delta)}$$
$$\therefore P\left(|\hat{p}_{win}(\theta, a) - p_{win}(\theta, a)| \geq \sqrt{\frac{7}{2n}\log\Big(\frac{2At}{\delta}\Big)}\right) \leq 2\exp\Big(-2n \cdot \frac{1}{2n} \cdot \log(\frac{120At^7}{\delta})\Big) \leq \frac{\delta}{60At^7}$$

Now the next steps follow Lemma 17(Appendix C.1) in UCRL2 (Union bound over all possible values of $n = 1, 2, \ldots t-1$).

$$P\left(|\hat{r}(\theta, a) - \bar{R}(\theta, a)| \geq \sqrt{\frac{14}{\max\{1, N(\theta, a)\}}\log(\frac{2At}{\delta})}\right) \leq \sum_{N(\theta,a)=1}^{t-1} \frac{\delta}{60At^7} \leq \frac{\delta}{60At^6}$$
$$P\left(|\hat{p}_{win}(\theta, a) - p_{win}(\theta, a)| \geq \sqrt{\frac{7}{2\max\{1, N(\theta, a)\}}\log(\frac{2At}{\delta})}\right) \leq \frac{\delta}{60At^6}$$

$M \notin M(t)$ occurs if $(|\bar{R}(\theta, a) - \hat{r}(\theta, a)| \geq d'(\theta, a))$ or $|p_{win}(\theta, a) - \hat{p}_{win}(\theta, a)| \geq d(\theta, a)$ for any (θ, a). So we sum the above error probabilities over all (θ, a). Thus $P(M \notin M(t)) \leq \frac{\delta}{30t^6} \leq \frac{\delta}{15t^6}$ □

0.1.4 Episodes with $M \in \mathcal{M}_k$

$v_k(\theta, a)$ defined earlier denotes the number of times (θ, a) occurs in episode k. Similarly $v_k(s, a)$ denotes the number of times Algorithm 1 was in state s and took action a during episode k. [5]

Theorem 2 ensures that $\tilde{\pi}_k$ is $\frac{1}{\sqrt{t_k}}$ optimal. Let $\tilde{\rho}_k$ denote the average reward estimate obtained after convergence. Since $M \in \mathcal{M}_k$, this means the average reward for the true MDP $\rho^* \leq \tilde{\rho}_k + \frac{1}{\sqrt{t_k}}$

$$\Delta_k = \sum_{\theta,a} v_k(\theta, a)(\rho^* - \bar{R}(\theta, a)) \leq \boxed{\sum_{\theta,a} v_k(\theta, a)(\tilde{\rho}_k - \bar{R}(\theta, a)) + \sum_{\theta,a} \frac{v_k(\theta, a)}{\sqrt{t_k}}}$$

The boxed term can be rewritten [6] as $\boxed{\sum_{s,a} v_k(s, a)(\tilde{\rho}_k - \bar{R}(s, a)) + \sum_{s,a} \frac{v_k(s, a)}{\sqrt{t_k}}}$.

Convergence criteria gives $|u_{i+1}(s) - u_i(s) - \tilde{\rho}_k| \leq \frac{1}{\sqrt{t_k}}\ \forall s$

Also $u_{i+1}(s) = \tilde{r}_k(s, \tilde{\pi}_k(s)) + \sum_{s'} \tilde{p}_k(s'|s, \tilde{\pi}_k(s)) \cdot u_i(s')$, So by expanding we get

$$\Big|\Big(\tilde{\rho}_k - \tilde{r}_k(s, \tilde{\pi}_k(s))\Big) - \Big(\sum_{s'} \tilde{p}_k(s'|s, \tilde{\pi}_k(s)) \cdot u_i(s') - u_i(s)\Big)\Big| \leq \frac{1}{\sqrt{t_k}}$$

$$\Delta_k = \sum_{s,a} v_k(s, a)(\tilde{\rho}_k - \tilde{r}_k(s, a)) + \sum_{s,a} v_k(s, a)(\tilde{r}_k(s, a) - \bar{R}(s, a)) + \sum_{s,a} \frac{v_k(s, a)}{\sqrt{t_k}} \tag{7}$$

$$\Delta_k \leq \overbrace{\boldsymbol{v_k(\tilde{P}_k - I)u_i}}^{\text{term 1}} + \overbrace{\sum_{s,a} v_k(s, a)(\tilde{r}_k(s, a) - \bar{R}(s, a))}^{\text{term 2}} + \overbrace{2\sum_{s,a} \frac{v_k(s, a)}{\sqrt{t_k}}}^{\text{term 3}} \tag{8}$$

$\boldsymbol{v_k} \coloneqq v_k\big(\big(s, \tilde{\pi}_k(s)\big)\big)_s$ is a row vector, containing visit count for each state s and corresponding action $\tilde{\pi}_k(s)$. $\boldsymbol{\tilde{P}_k} \coloneqq \big(\tilde{p}_k(s'|s, \tilde{\pi}_k(s))\big)_{s,s'}$ is the $S \times S$ transition matrix, each row of this matrix has exactly 4 non zero entries (see (3)).

term 3 $= 2\sum_{\theta,a} \frac{v_k(\theta,a)}{\sqrt{t_k}}$, term 2 [7] $\leq 2\sum_{s,a} v_k(s, a)\sqrt{\frac{c_1 \log(c_1' A t_k/\delta)}{\max\{1, N_k(\theta_s, a)\}}} = 2\sum_{\theta,a} v_k(\theta, a)\sqrt{\frac{c_1 \log(c_1' A t_k/\delta)}{\max\{1, N_k(\theta, a)\}}}$, also $\max\{1, N_k(\theta, a)\} \leq t_k \leq T$

$$\text{term2+term3} \leq \Big(2\sqrt{c_1 \log\Big(\frac{c_1' AT}{\delta}\Big)} + 2\Big) \sum_{\theta,a} \frac{v_k(\theta, a)}{\sqrt{\max\{1, N_k(\theta, a)\}}} \tag{9}$$

$w_k(s) \coloneqq u_i(s) - \frac{\min_s u_i(s) + \max_s u_i(s)}{2}$

[5] Note that episode ends are triggered by some $v_k(\theta, a) \geq N_k(\theta, a)$, $v_k(s, a)$ is introduced for the sake of analysis

[6] $\because \bar{R}(s, a) = \bar{R}(\theta_s, a)$

[7] We use $\tilde{r}_k(s, a) - \bar{R}(s, a) \leq |\tilde{r}_k(\theta_s, a) - \hat{r}_k(\theta_s, a)| + |\hat{r}_k(\theta_s, a) - \bar{R}(\theta_s, a)| \leq 2d'(\theta_s, a)$, Similarly $\|\tilde{p}_k(.|s, \tilde{\pi}_k(s)) - p(.|s, \tilde{\pi}_k(s))\|_1 \leq \|\tilde{p}_k(.|s, \tilde{\pi}_k(s)) - \hat{p}_k(.|s, \tilde{\pi}_k(s))\|_1 + \|\hat{p}_k(.|s, \tilde{\pi}_k(s)) - p(.|s, \tilde{\pi}_k(s))\|_1 \leq 4d(\theta_s, \tilde{\pi}_k(s))$

Just as in the UCRL2 analysis, term 1 i.e $\boldsymbol{v_k(\tilde{P}_k - I)u_i}$ can be rewritten as $\boldsymbol{v_k(\tilde{P}_k - I)w_k}$
Also $\boldsymbol{v_k(\tilde{P}_k - I)w_k = v_k(\tilde{P}_k - P_k)w_k + v_k(P_k - I)w_k}$. Here $\boldsymbol{P_k} \coloneqq \big(p(s'|s, \tilde{\pi}_k(s))\big)_{s,s'}$
First we bound $\boldsymbol{v_k(\tilde{P}_k - P_k)w_k}$

$$
\begin{aligned}
&v_k(\tilde{P}_k - P_k)w_k \leq \sum_s v_k(s, \tilde{\pi}_k(s)) \cdot \|\tilde{p}_k(.|s, \tilde{\pi}_k(s)) - p(.|s, \tilde{\pi}_k(s))\|_1 \cdot \|w_k\|_\infty \\
&\leq \sum_s v_k(s, \tilde{\pi}_k(s)) \cdot 4\sqrt{\frac{c_2 \log(c_2' A t_k/\delta)}{\max\{1, N_k(\theta_s, \tilde{\pi}_k(s))\}}} \cdot \frac{D}{2} \\
&\leq 2D\sqrt{c_2 \log(c_2' A t_k/\delta)} \boxed{\sum_s \frac{v_k(s, \tilde{\pi}_k(s))}{\sqrt{\max\{1, N_k(\theta_s, \tilde{\pi}_k(s))\}}}} = 2D\sqrt{c_2 \log(c_2' A t_k/\delta)} \sum_{\theta,a} \frac{v_k(\theta, a)}{\sqrt{\max\{1, N_k(\theta, a)\}}} \\
&\leq 2D\sqrt{c_2 \log(c_2' A T/\delta)} \sum_{\theta,a} \frac{v_k(\theta, a)}{\sqrt{\max\{1, N_k(\theta, a)\}}} \qquad (10)
\end{aligned}
$$

The upper bound for $\boldsymbol{v_k(P_k - I)w_k}$ exactly follows the UCRL2 analysis, it uses the Azuma-Hoeffding inequality for the martingale difference sequence $X_t \coloneqq (p(.|s_t, a_t) - e_{s_{t+1}})w_{k(t)} \mathbb{1}_{M \in M_{k(t)}}$ (see [**UCRL2**])

$$
\begin{aligned}
v_k(P_k - I)w_k &\leq D + \sum_{t=t_k}^{t_{k+1}-1} X_t \\
\sum_{k=1}^{m} v_k(P_k - I)w_k &\leq mD + \sum_{t=1}^{T} X_t \qquad (11)
\end{aligned}
$$

$\sum_{t=1}^{T} X_t \leq D\sqrt{\frac{5}{2} T \log(\frac{8T}{\delta})}$ with probability atleast $1 - \frac{\delta}{12T^{5/4}}$ (AZ-Hoeffding inequality)

Lemma 4. *Number of episodes m of **UCRL2 adapted** upto step $T \geq 2A$ is upper bounded as*

$$m \leq 2A \log_2(\frac{4T}{A})$$

Proof. $N(\theta, a) \coloneqq \#\{\tau < T + 1 : s_\tau = s, a_\tau = a\}$ be the total number of (θ, a) observations till step T. For each episode $k < m$ the episode end is triggered by some (θ, a) for which either

1. $v_k(\theta, a) = 1$ when $N_k(\theta, a) = 0$
2. or $v_k(\theta, a) = N_k(\theta, a)$

Let $K(\theta, a)$ be the number of episodes with $v_k(\theta, a) = N_k(\theta, a)$ and $N_k(\theta, a) > 0$ then

$$
\begin{aligned}
N(\theta, a) &= \sum_{k=1}^{m} v_k(\theta, a) \geq 2^{K(\theta,a)} - 1 \\
T &= \sum_{\theta,a} N(\theta, a) \geq \sum_{\theta,a} \left(2^{K(\theta,a)} - 1\right) \qquad (12)
\end{aligned}
$$

Also $\sum_{\theta,a} K(\theta,a) \geq m - 1 - |\theta| \cdot A \geq m - 1 - 2A$. [8]
$\sum_{\theta,a} 2^{K(\theta,a)} \geq 2A\Big(\prod_{\theta,a} 2^{K(\theta,a)}\Big)^{1/2A} = 2A \cdot 2^{\sum_{\theta,a} 2^{K(\theta,a)/2A}} \geq 2A2^{\frac{m-1}{2A}-1}$ [9]

Finally from (12) and AmGm inequality $T \geq 2A(2^{\frac{m-1}{2A}-1} - 1)$. And since $T \geq 2A$

$$2^{\frac{m-1}{2A}-1} \leq \Big(\frac{T}{2A} + 1\Big) \leq \Big(\frac{T}{A}\Big)$$

So $m \leq 1 + 2A + 2A\log_2(\frac{T}{A}) \leq 2A(2 + \log_2(\frac{T}{A})) \leq \boxed{2A(\log_2(\frac{4T}{A}))}$. □

0.1.5 Summing over episodes with $M \in \mathcal{M}_k$

Returning to Eq(8) the sum of term1,term2,term3 is upper bounded(see below) with probability atleast $1 - \frac{\delta}{12T^{5/4}}$
We use (10), (11), (9). [10]

$$\begin{aligned}\sum_{k=1}^{m} \Delta_k \mathbb{1}_{M\in\mathcal{M}_k} \leq &\Big(2\sqrt{14\log(\frac{2AT}{\delta})} + 2\Big)\sum_{k=1}^{m}\sum_{\theta,a}\frac{v_k(\theta,a)}{\sqrt{\max\{1,N_k(\theta,a)\}}} \\ &+ 2D\sqrt{\frac{7}{2}\log(\frac{2AT}{\delta})}\sum_{k=1}^{m}\sum_{\theta,a}\frac{v_k(\theta,a)}{\sqrt{\max\{1,N_k(\theta,a)\}}} \\ &+ D\sqrt{\frac{5}{2}T\log(\frac{8T}{\delta})} + 2DA\log_2(\frac{4T}{A})\end{aligned}$$

The main intermediate step here is that

$$\sum_{k=1}^{m}\sum_{\theta,a}\frac{v_k(\theta,a)}{\sqrt{\max\{1,N_k(\theta,a)\}}} \leq (\sqrt{2}+1)\sqrt{2AT}$$

thus with probability atleast $1 - \frac{\delta}{12T^{5/4}}$

$$\sum_{k=1}^{m} \Delta_k \mathbb{1}_{M\in\mathcal{M}_k} \leq D\sqrt{\frac{5}{2}T\log(\frac{8T}{\delta})} + 2DA\log_2(\frac{4T}{A}) + \Big(2D\sqrt{14\log(\frac{2AT}{\delta})} + 2\Big)(\sqrt{2}+1)\sqrt{2AT}$$

[8] as in the worst case we fill each (θ, a) bin, before doubling occurs
[9] Arithmetic mean ≥ Geometric mean
[10] Diameter(D) for our MDP $2K \leq D \leq c(2K+1)$, $K \in \mathbb{N}$ is the absolute parity, c is a constant

0.1.6 Sum of Δ_k over all episodes

Recall Eq (6) and the previous results, Thus with probability atleast
$1 - \frac{\delta}{12T^{5/4}} - \frac{\delta}{12T^{5/4}} - \frac{\delta}{12T^{5/4}} = 1 - \frac{\delta}{4T^{5/4}} \geq 1 - \delta$

$$\begin{aligned}\Delta(s_1, T) &\leq \sum_{k=1}^{m} \Delta_k \mathbb{1}_{M \notin M_k} + \sum_{k=1}^{m} \Delta_k \mathbb{1}_{M \in M_k} + \sqrt{\frac{5}{2} T \log(\frac{8T}{\delta})} \\ &\leq \sqrt{\frac{5}{2} T \log(\frac{8T}{\delta})} + 2\sqrt{T} + D\sqrt{\frac{5}{2} T \log(\frac{8T}{\delta})} + 2DA \log_2(\frac{4T}{A}) \\ &+ \Big(2D\sqrt{14 \log(\frac{2AT}{\delta})} + 2\Big)(\sqrt{2}+1)\sqrt{2AT}\end{aligned} \tag{13}$$

Each of the three $\frac{\delta}{12T^{5/4}}$ correspond to the probability of a "bad" event occuring, namely:

1. Probability of landing outside the confidence interval in (5)
2. We know $P(\exists : T^{1/4} < t \leq T : M \notin M(t)) \leq \frac{\delta}{12T^{5/4}}$, this effectively makes the term $\mathbb{1}_{M \notin M_k(t)} \to 0$ with high probability.
3. Probability of landing outside the confidence interval given by the azuma hoeffding inequality

The goal is to prove a bound of $\tilde{O}(D\sqrt{AT})\ \forall T \geq 1$
If $1 \leq T \leq 25\sqrt{AT \log(\frac{T}{\delta})} \iff 1 \leq T \leq 25^2 A \log(\frac{T}{\delta})$ its straightforward:
$\Delta(s_1, T) = T\rho^* - \sum_{t=1}^{T} R(\theta_t, a_t) \leq \sum_{t=1}^{T}(1 - R(\theta_t, a_t)) \leq 2T \leq 50\sqrt{AT \log(\frac{T}{\delta})}$

If $T > 25^2 A \log(\frac{T}{\delta}) \iff A < \frac{1}{25 \log(\frac{T}{\delta})}\sqrt{AT \log(\frac{T}{\delta})}$ [11], also $log_2(4T) \leq 2\log(T)$ therefore $2DA\log_2(\frac{4T}{A}) \leq \frac{4}{25} D\sqrt{AT \log(\frac{T}{\delta})}$
Notice that for $T > 25^2 A \log(\frac{T}{\delta})$ [12], $\log(\frac{2AT}{\delta}) \leq 2\log(\frac{T}{\delta})$ and $\log(\frac{8T}{\delta}) \leq 2\log(\frac{T}{\delta})$
Also $A \geq 2$, $\frac{1}{\sqrt{A}} \leq \frac{1}{\sqrt{2}}$, Thus using Eq (13), we have for $T > 1$ with probability atleast $1 - \delta$

$$\begin{aligned}\Delta(s_1, T) &\leq D\sqrt{AT}\Big(2\sqrt{\frac{1}{A} \cdot \frac{5}{2} \log(\frac{8T}{\delta})} + 2\sqrt{2}(\sqrt{2}+1)\sqrt{14 \log(\frac{2AT}{\delta})} + 2\sqrt{2}(\sqrt{2}+1) + \frac{1}{\sqrt{A}}\Big) \\ &+ 2DA \log_2(\frac{4T}{A}) \\ &\leq D\sqrt{AT \log(\frac{T}{\delta})}\Big(2\sqrt{\frac{5}{2}} + 2\sqrt{2}(\sqrt{2}+1)\sqrt{28} + 2\sqrt{2}(\sqrt{2}+1) + \frac{1}{\sqrt{2}} + \frac{4}{25}\Big) \\ &\leq 46.9904 D\sqrt{AT \log(\frac{T}{\delta})} \leq 50 D\sqrt{AT \log(\frac{T}{\delta})}\end{aligned}$$

[11] Also notice how T has to be > 100 to satisfy $T > 625A\log(\frac{T}{\delta}) > 1250\log(T)$
[12] the constraint implies $T > 2A$, why? simple proof by contradiction