École Normale Supérieure de Lyon
Thèse de master — Informatique

Online Learning Algorithms for Fair Ad Auctions

Krishna Acharya

Under the supervision of:
Patrick Loiseau LIG, INRIA-Grenoble
Nicolas Gast LIG, INRIA-Grenoble

Grenoble - August, 2020

Abstract

Online advertisements are increasingly prevalent on the internet. The allocation of ad slots to advertisers has been typically studied from the perspective of a platform designer, with the goal being maximization of metrics such as revenue, social welfare etc.

However, we study the problem of repeated ad auctions from the perspective of a constrained advertiser. Specifically, we show what an advertiser with non-discrimination constraints should do given the auction mechanism is prespecified. We provide results for the full information setting using dynamic programming for MDPs and deal with the partial information setting via online reinforcement learning algorithms.

Keywords: *Fairness, Non-discrimination, Online Learning, Reinforcement Learning, Markov Decision Process*

Contents

1 Introduction

Online advertisements are prevalent on the internet, with companies like Google, Facebook deriving a major portion of their revenue by displaying advertisements to the website users. After every **discrete event** for e.g searching an item on Google search, scrolling through "x" amount of posts on the Facebook homepage: an ad is displayed. Each advertiser has a target demographic and if a website user fits this demographic (based on data collected by Google, Facebook) an ad is placed. Which advertiser's ad to display is determined by an auction. Each interested advertiser bids some amount and the Ad Exchange platform e.g Google Ads, Facebook Ads [2] [1] after obtaining all the bids declares the winner(who gets to display its ad) and the process repeats. Thus the **system can be viewed as repeated auctions**, until some finite number of repetitions.

In repeated auction literature we usually deal with 2 types of constraints - 1) **finite budget constraint** and the more recently studied - 2) **non-discrimination constraint (fairness)**. Fairness constraints arise due to many reasons [8], some are even legally required for e.g no discrimination by the advertiser based on the user's gender, race etc.

Most Ad exchange platforms run a second price auction (see section 2.1) to determine the winning ad. In the absence of any constraints this repeated auction is easy to analyse, however as soon as we add budget constraints and/or fairness constraints the optimal bidding strategy for the advertiser is not trivial. [6], [10] specifically deal with budget constraints. To deal with fairness constraints [7] changes the auction mechanism. Like [14], **our focus is on what an advertiser with non-discrimination constraints but no budget constraint should do given the ad auction is fixed to second price**. Note that we do not consider game theoretic dynamics, like [6], where other bidders adapt to the constrained bidders strategy. Thus the implicit assumption is the number of advertisers is "large enough" so that this competitition does not matter and other bidders can be modelled by stationary distributions.

1.1 Contribution

This work is an extension of [14] in which the stationary distributions for the other bidders were estimated from past auction data. This enabled them to precompute the optimal bidding strategy for the current series of auctions. **We relax this assumption and do not need any knowledge from past auctions**. Our contributions are to 1) make a minor but useful change in the full-information setting of [14] to instead deal with undiscounted cumulative rewards. 2) apply online reinforcement learning(RL) algorithms to obtain near optimal policies. In particular, we apply both model-based and model-free algorithms. The model-based algorithms [17][16][4] are recent techniques and have an added advantage of having a theoretical regret bound. The model-free algorithms are based on Temporal difference learning, a classic reinforcement learning paradigm. We see empirically that the (a) model-based algorithms perform close to optimal. (b) model-free algorithms always performed worse compared to model-based, even failing badly in certain scenarios.

1.2 Organisation of the thesis

In Chapter 2 we begin with a description of the Ad auction platform. Chapter 3 models ad auctions with fairness as a Markov Decision Process in the full information setting. Chapter 4 deals with partial information and applies various online RL algorithms. Chapter 5 contains the numerical results and comparison for all the algorithms. Appendix A has the theoretical regret bounds and Appendix B important results for a 2^{nd} Price auction.

2 Ad auction platform

In this section we describe the Ad auction platform, its corresponding parameters and the fairness constraint the advertiser has to follow.

2.1 Basic auction terminology

Consider a total of N advertisers, each indexed by $i \in \{1, \ldots, N\}$. Advertiser i values the user at v_i, known as its **true value**[1], and makes a bid b_i. The platform collects all the bids and runs an auction to determine the winning advertiser i'. The price p is determined by a function $f(.)$ in the auction mechanism i.e $p := f(b_1, \ldots, b_N)$. The corresponding reward $r_{i'}$ obtained is defined as $\boldsymbol{r_{i'} := v_{i'} - p}$. Many auction mechanisms have been studied in auction theory literature [13] but the one we focus on and which is prevalent in internet ad platforms is the **"Sealed bid second price auction"**. It is described as follows, 1) Collect bids from all the advertisers; 2) The highest bidding advertiser **wins** the auction (i.e gets to display its ad); 3)the price paid by the winner is the second highest bid.

$$\begin{aligned} \textbf{Winner } i' &:= \arg\max_i b_i \\ p &:= \max_{j \neq i'} b_j \\ r_{i'} &:= v_{i'} - p \end{aligned} \tag{2.1}$$

The reward for all advertisers except the winner is defined to be zero i.e $r_{i|i \neq i'} := 0$.
Concisely, for any i $\boxed{\boldsymbol{r_i = (v_i - \max(b_{-i})) \cdot \mathbb{1}_{b_i > \max b_{-i}}}}$. For 2^{nd} price auctions bidding the true value (**truthful bidding**) is a weakly dominant strategy (see **Appendix** 8.1). In other words if bidding truthfully obtains a reward r then bidding anything else can only give a reward $\leq r$

2.2 Ad auction platform

In a typical Ad exchange platform, e.g Google Ads, users keep arriving and their attributes (e.g age, gender, location) are relayed to the advertisers. This process goes on for $\boldsymbol{T}$ slots, known as the **ad campaign duration**. The attributes of the user observed in the current slot($t \in \{1, \ldots, T\}$) determines each advertiser's true value. An interested advertiser j (i.e with $v_j > 0$) bids b_j, all uninterested advertisers are assumed to be bidding zero. Since the auction mechanism is fixed to second price, the optimal strategy for an advertiser with no budget or fairness constraints, is to bid truthfully. This holds because from the perspective of the unconstrained advertiser each user can be treated as a new second price auction.

[1]This is only known to advertiser i, so it is also called its private value

Our focus is on computing the optimal bidding strategy from the perspective of a fairness constrained advertiser. Section 2.3 describes quantitatively the absolute parity constraint.

2.3 Fairness - Absolute parity constraint

As in [14] Let ω denote a set of **sensitive attributes** with respect to which the advertiser due to legal or other reasons wants to be fair. We will focus only on one sensitive attribute i.e gender of the user. Thus $w := \{\textbf{gender} \in \{\text{male}, \text{female}\}\}$. $\boldsymbol{n_m(t), n_f(t)}$ **denotes the number of male, female users that the advertiser has won till time slot** t.

Definition 1. *K-parity w.r.t gender*

Advertiser follows a K-strict absolute parity constraint with respect to gender iff, after each auction round $t \in \{1, \ldots, T\}$, $|n_m(t) - n_f(t)| \leq K$

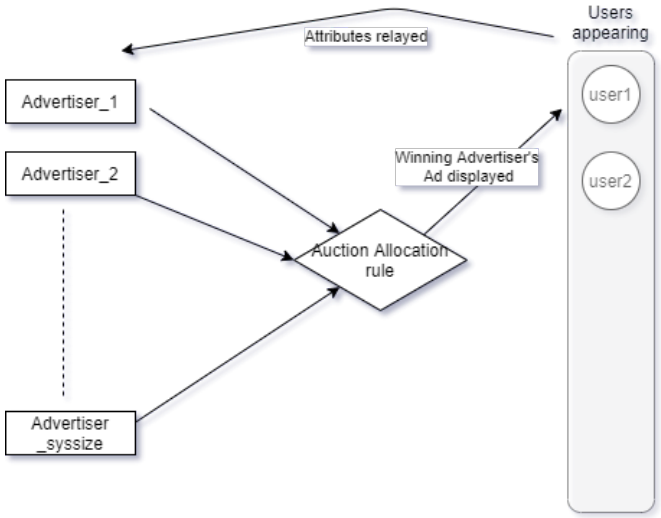


Figure 2.1: Typical Ad exchange platform

2.4 System parameters and Goal

Let index j denote the constrained advertiser. With Figure 2.1 in mind: A single user arrives at each time slot. We assume that the probability of it being male is $\boldsymbol{p_m}$, probability of it being female is $\boldsymbol{p_f := 1 - p_m}$. The constrained advertiser's true values for male and female are drawn from probability distributions $\boldsymbol{V_m, V_f}$(both with support $\in [0, 1]$). $\boldsymbol{v_m, v_f}$ are the expected values of V_m and V_f respectively. The maximum of the other advertiser's bids when the user is male is $\boldsymbol{D_m}$. $\boldsymbol{D_f}$ is similarly defined for a female user. To obtain the optimal strategy it is sufficient to know $\boldsymbol{g_m, g_f}$: the cumulative distribution function(cdf) for the random variables D_m, D_f (the supports are again $\in [0, 1]$). Quantitatively, for user gender $\theta \in \{m, f\}$ and $B_{i,\theta}$ denoting the corresponding bid by advertiser i.

$$
\begin{aligned}
D_\theta &:= \max_{i|i \neq j} B_{i,\theta} \\
g_\theta(x) &= P(D_\theta \leq x) = \prod_{i|i \neq j} P(B_{i,\theta} \leq x)
\end{aligned} \tag{2.2}
$$

The immediate **reward** to the advertiser is $\mathbb{1}_{\boldsymbol{b(t)>D_\theta}} \cdot \boldsymbol{(V_\theta - D_\theta)}$. Here $\boldsymbol{b(t)} \in [0,1]$ is the bid by the constrained advertiser at time t. The following is a straightforward result by taking the expectation of the immediate reward.

$$\mathbb{E}[\mathbb{1}_{b(t)>D_\theta}(V_\theta - D_\theta)] = P(D_\theta < b(t))v_\theta - \int_0^{b(t)} x g'_\theta(x) = \boxed{(v_\theta - b(t))g_\theta(b(t)) + \int_0^{b(t)} g_\theta(u)du} \quad (2.3)$$

T is the duration of the ad campaign. **The $b(t)$'s for $t \in \{1 \dots T\}$ that provide maximum expected cumulative reward given the parity constraint in definition (1) are what we seek.**

3 A full information model for fair ad auctions

Here we extend the model in [14] to work with an undiscounted [1] ad campaign of duration T. The learning algorithms (in section 4.2) work with undiscounted rewards so its a good common ground for comparison.

3.1 Counterexample to the truthful strategy

We can question why **bidding truthfully fails**, here's why: Consider an ad campaign of 3 slots with *K-parity* $= 1$ and the advertiser's value for male and female $v_m = 0.5$, $v_f = 0.5$ respectively. The advertiser knows from past history that the other bidders probably value male more than female. However it ignores this data and just bids truthfully.
Consider the sequence in the table, the optimal strategy obtains a cumulative reward of 0.3, it overbids[2] and wins in the first slot, wins in the second and also wins in the third. .
Truthful bidding obtains a cumulative reward of 0.2, it loses the first auction, wins the second, cannot bid in the third(as the parity $|n_m(t) - n_f(t)| \leq K = 1$ has to be satisfied).

	$t = 1$ (male)	$t = 2$ (female)	$t = 3$(female)
Highest bid others($\max b_{-j}$)	0.6	0.3	0.3
reward-optimal at slot t	-0.1	0.2	0.2
reward-truthful at slot t	0	0.2	0

3.2 Markov Decision Process

Markov Decision Process(MDP) and techniques to solve it are used in this chapter and Chapter 4. For our purposes we need basic definitions and 2 algorithms.

Definition 2. *MDP is a 4-tuple $\boldsymbol{M = (S, A, P, R)}$, where S is a set of states called the state space, A is a set of actions called the action space, $P(s'|s,a) = P(s_{t+1} = s' \mid s_t = s, a_t = a)$ is the probability that action a in state s at time t will lead to state s' at time $t+1$, $R(s,a)$ is immediate reward received after taking action a in state s, $\bar{R}(s,a)$ is its expected value.*

The goal is to find a policy π_t[3] that will maximize the expected sum: $\mathbb{E}[\sum_{t=1}^{T} R(s_t, a_t)]$, where we choose the action at time t according to the policy π_t, i.e $a_t = \pi_t(s_t)$. Value iteration can be defined by

[1] reward at $t = 1$ is same as reward at $t = t'$
[2] bidding higher than its true value for male which was 0.5 - **Overbid**
[3] For each $t \in \{1, \dots, T\}, \pi_t : S \to A$

the form of update:

$$J_{i+1}(s) := \max_a \left\{ \bar{R}(s,a) + \sum_{s'} P(s'|s,a) J_i(s') \right\} \tag{3.1}$$

There is a special significance to $J_i(s)$, when dealing with finite T it represents the expected cumulative reward starting from state s with i steps remaining.

For the average reward maximization criterion, a stationary ϵ- optimal policy can be obtained for a unichain MDP. [4] [18] [3].

Accordingly Algorithm 1 deals with expected cumulative reward maximization (finite T). Algorithm 2 is used for finding the policy that gives average reward ϵ - close to $\boldsymbol{\rho^*_M}$(the optimal average reward).

Algorithm 1 Value iteration finite horizon

procedure GET POLICY
  $J_0(s) = \max_a \bar{R}(s,a)\ \forall s$
  for $t = 1 \dots T$ **do**
    $\pi_t(s) = \arg\max_a [\bar{R}(s,a) + \sum_{s'} P(s'|s,a) J_{t-1}(s')]\ \ \forall s$
    $J_t(s) = \bar{R}(s, \pi_t(s)) + \sum_{s'} P(s'|s, \pi_t(s)) J_{t-1}(s')\ \ \forall s$
  end for
  return $\pi_t\ \forall t \in \{1, \dots, T\}$ ▷ each $\pi_t : S \to A$
end procedure

Algorithm 2 Value iteration for Unichain MDP

procedure GET POLICY
  Input optgap ϵ
  $t = 0, J_{prev} \in \text{Uniform}[0,1]^S$ ▷ All states arbitrarily set to Uniform draws $\in [0,1]$
  while True **do**
    $\pi_t(s) = \arg\max_a [\bar{R}(s,a) + \sum_{s'} P(s'|s,a) J_{prev}(s')]\ \ \forall s$
    $J_{next}(s) = \bar{R}(s, \pi_t(s)) + \sum_{s'} P(s'|s, \pi_t(s)) J_{prev}(s')\ \ \forall s$
    if $\max_s(J_{next}(s) - J_{prev}(s)) - \min_s(J_{next}(s) - J_{prev}(s)) < \epsilon$ **then** ▷ $span(J_{next} - J_{prev}) < \epsilon$
      break
    end if
    $J_{prev} = J_{next}, t = t + 1$
  end while
  return π_t
end procedure

The optimal state-action value function $\boldsymbol{Q^*(s,a)}$ is defined as the optimal total expected reward obtained starting from state s given action a is chosen initially. The relation between $J^*(s)$ and $Q^*(s,a)$ is as follows.

$$\begin{aligned} J^*(s) &= \max_{a \in A} Q^*(s,a) \\ Q^*(s,a) &= \bar{R}(s,a) + \sum_{s' \in S} P(s'|s,a) J^*(s') \end{aligned} \tag{3.2}$$

[4] i.e any stationary deterministic policy induces a single ergodic Markov chain

Definition 3. *Diameter of MDP* $\boldsymbol{D_M}$

For the stochastic process generated by stationary deterministic policies $\pi : S \to A$ *on MDP* M *with initial state* s, $T(s'|M, \pi, s)$ *is the random variable for the first time step in which state* s' *is reached in this process.*

$$D_M := \max_{s \neq s'} \min_{\pi: S \to A} \mathbb{E}[T(s'|M, \pi, s)] \tag{3.3}$$

3.3 Absolute Parity MDP

Given the K-parity constraint(1) and the system parameters 2.4 we can **represent the problem as a MDP**.

The state is a tuple containing 1) The difference: $n_m - n_f \in \{-K, \dots K\}$; 2) gender of the current user $\theta \in \{m, f\}$, thus the state space $S = \{-K, \dots, K\} \times \{m, f\}$. The action is a bid $b \in [0, 1]$, thus the **action space** $A = [0, 1]$. The **expected immediate reward** $\bar{R}(s, b) = (v_{\theta_s} - b)g_{\theta_s}(b) + \int_0^b g_{\theta_s}(u)du$, here θ_s is gender in state s. For **transition probabilities** $P(s'|s, b)$ it is useful to refer Fig 3.1. Any transition from s_t to s_{t+1} can be seen as a two step process: 1) make a bid, observe the corresponding change in difference ; 2) observe gender of the user in the next slot.

For the edge states $(-K, f)$ and (K, m) the advertiser does not bid as that could violate the K-parity constraint, thus the difference remains the same. In any other state the advertiser always bids and if it wins the auction the difference changes. Which means for the edge states we have exactly 2 transitions. Whereas the others states give us 4 **"types" of transitions**, all the bids transition to the same 4 states but with different probabilities depending on the amount bid.

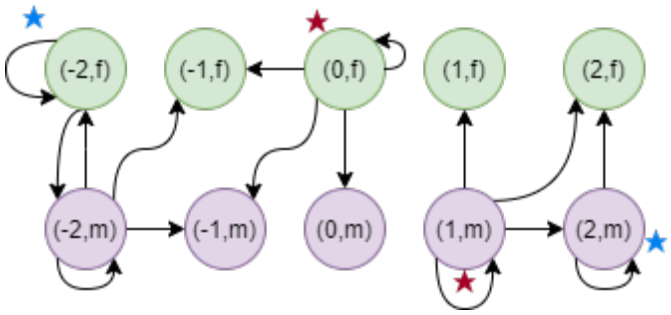


Figure 3.1: $K = 2$, blue stars are edge states, red stars are normal states.(All transitions not shown)

3.4 Bidding strategies

Here we compute the optimal bidding strategy that maximizes expected cumulative reward. We have the absolute parity MDP and a finite horizon(ad campaign duration T), thus the optimal bid $\pi_t(s)$ in state $s = (\mathit{diff}, \theta)$ with t slots remaining can be computed by using Algorithm 1.

$$\begin{aligned}
\pi_0(s) &= \arg\max_b \bar{R}(s, b) = \arg\max_b \left[(v_{\theta_s} - b)g_{\theta_s}(b) + \int_0^b g_{\theta_s}(u)du\right] \\
\pi_t(s) &= \arg\max_b \left[(v_{\theta_s} - b)g_{\theta_s}(b) + \int_0^b g_{\theta_s}(u)du + \sum_{s'} P(s'|s, b)J_{t-1}(s')\right]
\end{aligned} \tag{3.4}$$

$g_\theta(u)$ is the cumulative distribution function, its non-decreasing, so for the continuous action space maximization we can use the following lemma.

Lemma 4. *$f_\phi(x) = (\phi - x)h(x) + \int_0^x h(u)du$ where $h(x)$ is a non-decreasing function is maximized at $x = \phi$*

In case of discrete b, $\arg\max_b \left[(\phi - b)h(b) + \int_0^b h(u)du\right]$ would either be $\lceil\phi\rceil_{closest}$ or $\lfloor\phi\rfloor_{closest}$ i.e the nearest discrete bid above or below ϕ respectively (proof in Appendix 8.1) i.e we need not do a search through all the discrete b values.

In the following analysis we continue with bids $\in [0, 1]$, noting that the discrete analysis is not much different. $\pi_0(s) = v_{\theta_s}$, in other words the advertiser should bid truthfully for the last slot.
For state $s = (d, \theta)$, win difference(wd) is the difference if the advertiser wins the auction, therefore, if $\theta = m, wd = d + 1$ and if $\theta = f, wd = d - 1$. $\theta' \in \{m, f\}$ is next user's gender, the probability of winning user of type θ given $bid = b$ can be obtained using the cumulative distribution $g_\theta(b)$. Thus the transition probabilities $P(s'|s, b)$

$$P(s' = (wd, \theta')|s = (d, \theta), bid = b) = g_\theta(b) \cdot p_{\theta'}$$
$$P(s' = (d, \theta')|s = (d, \theta), bid = b) = (1 - g_\theta(b)) \cdot p_{\theta'}$$

To get $\pi_t(s)$ the key observation is $\sum_{s'} P(s'|s, b)J_{t-1}(s')$ in Eqn (3.4) can be conveniently written as $\psi_t(s) \cdot g_\theta(b) + c_t(s)$.

$$\pi_t(s) = \arg\max_b \left[(v_\theta - b)g_\theta(b) + \int_0^b g_\theta(u)du + \overbrace{\sum_{s'} P(s'|s, b)J_{t-1}(s')}^{\text{term2}}\right]$$

$$\text{term2} = g_\theta(b)p_m J_{t-1}(wd, m) + g_\theta(b)p_f J_{t-1}(wd, f) + (1 - g_\theta(b))p_m J_{t-1}(d, m) + (1 - g_\theta(b))p_f J_{t-1}(d, f)$$

$$\text{term2} = g_\theta(b)\overbrace{\left\{p_m J_{t-1}(wd, m) + p_f J_{t-1}(wd, f) - p_m J_{t-1}(d, m) - p_f J_{t-1}(d, f)\right\}}^{\psi_t(s)}$$
$$+ \quad \overbrace{p_m J_{t-1}(d, m) + p_f J_{t-1}(d, f)}^{c_t(s)} \tag{3.5}$$

$$\pi_t(s) = \arg\max_b \left[(v_\theta + \psi_t(s) - b)g_\theta(b) + \int_0^b g_\theta(u)du + c_t(s)\right] \tag{3.6}$$

By Lemma 4 $\boxed{\pi_t(s) = v_\theta + \psi_t(s)}$, Thus the **recursive equations** are:

$$J_0(d, \theta) = \int_0^{v_\theta} g_\theta(u)du \quad \forall d \in \{-K, \ldots, K\}, \text{ except } J_0(-K, f) = J_0(K, m) = 0$$
$$J_t(d, \theta) = p_m J_{t-1}(d, m) + p_f J_{t-1}(d, f) + \int_0^{\pi_t(s)} g_\theta(u)du \tag{3.7}$$

Recall $\pi_t(s) := 0$ for the edge cases. Thus $J_t(K, m)$ and $J_t(-K, f)$ are $p_m J_{t-1}(K, m) + p_f J_{t-1}(K, f)$ and $p_m J_{t-1}(-K, m) + p_f J_{t-1}(-K, f)$ respectively.

3.4.1 Truthful bidding - a baseline

We now evaluate the truthful bidding strategy i.e bidding the true value each time. We can follow a similar analysis as done earlier. In fact, we do not need Lemma 4 as the bidding strategy is fixed. We obtain the following recursive equations.

$$J_0(d,\theta) = \int_0^{v_\theta} g_\theta(u)du \quad \forall d \in \{-K,\dots,K\} \text{ except } J_0(-K,f) = J_0(K,m) = 0$$

$$J_t(d,\theta) = p_m J_{t-1}(d,m) + p_f J_{t-1}(d,f) + \psi_t(s) g_\theta(v_\theta) + \int_0^{v_\theta} g_\theta(u)du$$

Even though optimal bidding always provides higher expected cumulative reward, truthful bidding has a nice property i.e it always gives positive cumulative reward in expectation. Thus truthful bidding serves as a baseline which the online learning algorithms in the next section should beat.

4 Online Learning

In this section we deal with the case when we do not know the bidding behaviour of the other bidders. Thus, we cannot precompute the optimal bids like in the earlier section. First we describe what is known to the learner (the constrained advertiser) and what types of feedback it can receive. We also introduce regret, which is a metric for how close the learner is to the optimal total reward. After this we provide the list of algorithms that were implemented.

4.1 Why consider online performance for repeated auctions?

The ad campaign by the advertiser lasts for T time slots. At the end of the ad campaign many system parameters could change - for e.g g_m, g_f or p_m, p_f. A practical example is when the advertiser switches ad platforms - going for GoogleAds to FacebookAds, or even within the same platform when new competitors enter. **Thus motivated by such scenarios the goal of the learning algorithm is to maximize cumulative reward till slot T, given that it does not have a complete description of the system.**

In the following we assume the advertiser knows its own value distributions V_m, V_f and their expected values v_m, v_f. It also knows the probabilities of a male or female user appearing $p_m, 1-p_m$. But the advertiser does not know g_m and g_f - the cumulative distribution functions for D_m and D_f (See Eq(2.2)). At the end of each round the learner receives feedback about the other bids depending on the action chosen. We consider two possible auction feedbacks [1]

Type 1 The first type of feedback is when the learner **observes the exact draws** of D_θ, $\theta \in \{m, f\}$ in each time slot. This kind of feedback is obtained when the ad auction platform makes the bid of all advertisers public after each round. Intuitively we can see that a Bayesian update could work here.

Type 2 The second kind of feedback is when the learner only knows whether it **won or lost** the auction. This scenario occurs when the ad auction platform does not make all the bids public. In essence, the learner observes two kinds of data 1) exact 2) censored. If the learner wins the auction it receives reward $(V_\theta - D_\theta)$ and since V_θ is known, the draw D_θ can be **exactly** found out. If it loses, it

[1] **Note the UCRL2 adapted algorithm in Appendix A does not use this side information, it updates reward, transition estimates only for the action taken**

only knows that the maximum of the other bids is greater than its own bid i.e $D_\theta \in (b_t, 1.0]$ denoted by b_t^+ known as **right censored** data. We can still insist on doing Bayesian updates with the exact draws and discarding the censored observations, however this is a clear wastage of samples. There is efficient way of estimating cumulative distribution functions in the presence of censored data [12]. The survival function $\boldsymbol{S(x) := P(X > x) = 1 - cdf(x)}$ has been studied from the context of lifetime analysis and $\hat{S}_{km}(x)$, the Kaplan- Meir(KM) estimator is a classic non-parametric statistic for it. Adapted to our problem it is as follows.

$$\hat{S}_{km}(x) = \prod_{i:\ x_i \leq x} \left(1 - \frac{h_i}{n_i}\right)$$

Where $x_i \in [0, 1]$ is a point at which **at least one exact draw** of D_θ has been observed. h_i is the number of exact draws at x_i, these are obtained only **when the learner wins** the auction. n_i is the total observations that are $\geq x_i$, thus n_i includes both exact and censored observations.
Thus, we maintain two estimators $\boldsymbol{1 - \hat{S}_{kmm}(x)}$ and $\boldsymbol{1 - \hat{S}_{kmf}(x)}$ for g_m and g_f respectively. If we receive type 2 feedback then the update phase in the model-based algorithms updates the KM estimator.

4.2 Regret

As in [5],[4],[16] here we describe regret, a metric which the model based algorithms try to minimize. A learning algorithm $\mathcal{L}$ starting in an initial state s of the MDP M generates a stochastic process. This stochastic process is described by (s_t, a_t, r_t) - the state at step t, the action taken by $\mathcal{L}$ in step t, and the reward obtained at step t. The corresponding **total reward for $\mathcal{L}$ in T steps** is

$$R(M, \mathcal{L}, s, T) := \sum_{t=1}^{T} r_t$$

From $R(M, \mathcal{L}, s, T)$, we obtain $\lim_{T\to\infty} \frac{1}{T}\, \mathbb{E}[R(M, \mathcal{L}, s, T)]$ the average reward. The average reward can be maximized by some stationary deterministic policy $\pi : S \to A$, the learning algorithm of course doesn't know this apriori, the regret captures how close we are to optimal performace. The **optimal average reward** ρ_M^* for a communicating MDP(i.e having finite diameter) is independent of the start state and is given by:

$$\rho_M^* := \max_{\boldsymbol{\pi \in \Pi_{stationary}}} \sum_{s \in S} \mu_\pi(s) r(s, \pi(s)) \tag{4.1}$$

Where the stationary distribution $\mu_\pi \in [0, 1]^S$ is a probability vector that is obtained by following the stationary policy π [2]. $r(s, \pi(s))$ is the expected reward obtained by taking action $\pi(s)$ in state s. **Regret** Δ for the learning algorithm $\mathcal{L}$ after T steps is defined as

$$\Delta(M, \mathcal{L}, s, T) := T\rho_M^* - \sum_{t=1}^{T} r_t \tag{4.2}$$

We now describe the following list of learning algorithms.

[2]Do not confuse the stationary policy and stationary distribution, the former describes actions, the latter describes steady state probabilities of the induced Markov chain

Model-Based Algorithms

- Posterior Sampling for Reinforcement Learning
- Thompson Sampling with Dynamic Episodes
- Upper Confidence Reinforcement learning 2

Model-free Algorithms

- Sarsa
- Expected Sarsa
- Q Learning
- Double Q Learning

4.3 Model based algorithms

Model based learning algorithms $\mathcal{L}$ that provide non trivial upper bounds on Regret $\Delta(M, \mathcal{L}, s, T)$ have been obtained with two main techniques - 1) Bayesian learning 2) Optimism in the face of uncertainty(OFU).

4.3.1 Bayesian learning based

The two Bayesian learning algorithms Posterior sampling for reinforcement learning (**PSRL**) [16] and Thompson sampling with dynamic episodes (**TSDE**) [17] can be summarized in 3 steps:
1) **Put prior distributions** on unknown parameters of the MDP, 2) **Compute the optimal policy** for the MDP(with parameters drawn from the priors) and follow this policy for some time steps 3) **Update the priors using the observations**, then go back to step 2)

Algorithm 3 Posterior Sampling for Reinforcement Learning -PSRL

Input: Prior distribution ϕ, Episode length I
for $t = 0, 1, \ldots T-1$ **do** ▷ Time steps
  if $t \equiv 0 \pmod{I}$ **then**
    sample MDP $M_{ep} \sim \phi(.|\mathcal{H}_{t-1})$ ▷ Sample from updated prior
    compute π_{ep} ▷ obtained by value iteration on M_{ep} using Algorithm 2
    $ep \leftarrow ep + 1$ ▷ Increment the episode number
  end if
  action $a_t \leftarrow \pi_{ep}(s_t)$
  observe r_t and s_{t+1}
  $\mathcal{H}_t = \mathcal{H}_{t-1} \cup (a_t, r_t, s_{t+1}, auctionFeedback_t)$ ▷ Add observation to history
end for

For the absolute parity MDP the unknown system parameters are g_m, g_f so we place priors only on those. We also include auction feedback at time t in the history $\mathcal{H}$. T is the duration of the ad campaign. After every I steps the prior is updated. According to [15] PSRL is conjectured[3] to give a $\tilde{O}(IS\sqrt{AT})$ upper bound on $\mathbb{E}[\Delta(M, \mathcal{L}, s, T)]$(the expected regret).

[3]The PSRL paper [16] defines the regret only for the episodic tasks, whereas our learning problem is non-episodic. [15] suggests practical solutions for the non-episodic case, e.g selecting an artifical episode length I

Algorithm 4 Thompson Sampling with Dynamic Episodes - TSDE

1: **Input:** Prior distribution ϕ
2: $t \leftarrow 1, t_{ep} \leftarrow 0$
3: **for** $ep = 1, 2, \ldots$ **do**
4: $tv \leftarrow t - t_{ep}$ ▷ tv is a temporary variable controlling the episode duration
5: $t_{ep} \leftarrow t$
6: **sample** MDP $M_{ep} \sim \phi(.|\mathcal{H}_{t-1})$ ▷ Sample from updated prior
7: **compute** π_{ep} ▷ Using Algorithm 2 on M_{ep}
8: **while** $t \leq t_{ep} + tv$ and $N_t(s,a) \leq 2N_{t_{ep}}(s,a)\ \forall (s,a) \in S \times A$ **do**
9: Action $a_t = \pi_{ep}(s_t)$
10: Observe r_t and s_{t+1}
11: $\mathcal{H}_t = \mathcal{H}_{t-1} \cup (a_t, r_t, s_{t+1}, auctionFeedback_t)$ ▷ Add observation to history
12: **end while**
13: **end for**

$N_t(s,a)$ is number of times the algorithm visited state s and took action a until time t. TSDE [17] is proven to provide a $\tilde{O}(HS\sqrt{AT})$ upper bound on expected regret given that the MDP it operates on is weakly communicating. H is an upper bound on the bias span [4] of J returned by Algorithm 2. From [9] $H \leq D_M$ the diameter. Note that the absolute parity MDP is weakly communicating and has a finite diameter.

In summary, both Algorithm 3 and Algorithm 4, proceed in episodes. Algorithm 3 uses a fixed interval of I after which it resamples the parameters for the MDP. However in Algorithm 4 the episode lengths are dynamic. It depends on two stopping events - 1) $t > t_{ep} + tv$ and 2) $N_t(s,a) > 2N_{t_{ep}}(s,a)$. The first event ensures that the episode length grows at a linear rate and the second event ensures that the number of visits to any state-action pair (s,a) is at most doubled

4.3.2 Optimism in the face of Uncertainty - OFU

Here we give a sketch of the UCRL2 Algorithm, for a full description see [4]. In Appendix A we alter UCRL2 to only maintain a (gender, action) count as that is what matters for the absolute parity MDP [5].

Algorithm 5 UCRL2-Sketch

1: **Input:** Confidence set constructor σ, Episode end signaller E
2: $t \leftarrow 0$
3: **for** $ep = 1, 2, \ldots$ **do**
4: construct confidence set $\mathcal{M}_{ep} = \sigma(\mathcal{H}_{t-1})$
5: find $\pi_{ep} \in \arg\max_{\pi \in \Pi_{stat}} \left[\max_{M \in \mathcal{M}_{ep}} \rho^{\pi}_{M} \right]$ ▷ Using extended value iteration, see [4]
6: **while** $E(\mathcal{H}_{t-1}) = False$ **do**
7: Action $a_t = \pi_{ep}(s_t)$
8: Observe r_t and s_{t+1}
9: $\mathcal{H}_t = \mathcal{H}_{t-1} \cup (a_t, r_t, s_{t+1}),\ t \leftarrow t+1$ ▷ Add observation to history
10: **end while**
11: **end for**

The general structure resembles the Bayesian learning algorithms described earlier, however the **key differences** are line 4 and 5. Instead of sampling MDP parameters from a prior distribution, UCRL2

[4] $\max_s(J_{t+1}(s) - J_t(s)) - \min(J_{t+1}(s) - J_t(s))$
[5] or infact for any 2^{nd} price repeated auction whose constraints result in a MDP with bounded diameter

defines a confidence set $\mathcal{M}_{ep} = \sigma(\mathcal{H}_{t-1})$ of plausible MDP's within which the true MDP lies with high probability. Then an algorithm known as extended value iteration finds a stationary policy that gives optimal average reward amongst all these plausible MDPs, this explains the two maximizers($\arg\max, \max$) on line 5. This policy is run for the episode till the episode signaller ends the episode. Note that the episode end signaller depends on the observations gathered, In this regard UCRL2's Episode end signaller is slightly similar to TSDE and only starts a new episode when the total number of visits to any (state,action) doubles. UCRL2 provides a $\tilde{O}(DS\sqrt{AT})$ upper bound on the regret with high probability. In particular the following is its main theorem

Theorem 5. *With probability at least $1-\delta$, it holds that for any initial state $s \in S$ and any $T > 1$*

$$\Delta(M, UCRL2, s, T) \leq 34DS\sqrt{ATlog(T/\delta)}$$

Note that in practice for our MDP which has similar transition probabilities, rewards - the actual regret for both Bayesian learning and Optimism based methods is far below the upper bounds for regret. In any case its good to have theoretical results, and empirically we shall see that they beat the model-free algorithms.

4.4 Model Free algorithms

The algorithms described earlier worked with MDPs directly, however it is not necessary to do this. In the following we describe some model free online learning algorithms belonging to the class of Temporal difference(TD) learning [6] [19]. These work by maintaining an online estimate of the state-action function Q, which approximates the optimal state action function Q^*. By rewriting Eqn (3.2)

$$Q^*(s,a) = \bar{R}(s,a) + \sum_{s' \in S} P(s'|s,a) \max_{a \in A} Q^*(s',a) \tag{4.3}$$

Here all the algorithms maintain a Q table, updating it at time t for action a_t taken in state s_t. The update is a convex combination of the old and new estimate of $Q(s_t, a_t)$, the learning rate $\alpha \in [0,1]$ decides the weight given to the recent observation.
The states for the absolute parity mdp are $S = (diff \in \{-K, \ldots, K\}, gender = \{m, f\})$ and actions are the discrete bids. Thus the **Q table size is** $2A(2K+1)$.
These TD learning algorithms converge asymptotically to the optimal policy, but are not studied from the perspective of regret minimization. However, they serve as a good baseline which our model-based learning algorithms should beat.

Definition 6. ***Epsilon greedy policy***
Suppose an agent is in state s_t, it is said to follow the ϵ greedy policy with respect to the Q table if it picks an arbitrary action w.p $\boldsymbol{\epsilon}$ and action from $\arg\max_a Q(s_t, a)$ *w.p* $\mathbf{1} - \boldsymbol{\epsilon}$

4.4.1 Sarsa

The S, A, R stand for state, action, reward respectively. It is an **on-policy** algorithm i.e the same policy(ϵ greedy) is used for updating Q and for getting next action.

[6] More specifically TD(0)

Algorithm 6 Sarsa

Algorithm Parameters: learning rate $\alpha \in (0, 1]$, small $\epsilon > 0$
Initialize $Q \sim \text{Uniform}[0, 1]^{S \times A}$ ▷ Initialize arbitrarily
$\textit{diff} \leftarrow 0, \theta \leftarrow \text{Bernoulli}(p_f)$, $\boldsymbol{s} = (\textit{diff}, \theta)$. ▷ diff = 0, no auctions won for either gender
Get $\boldsymbol{a}$ from $Q(s, A)$ acc to ϵ greedy
for $t = 1, 2 \dots T$ **do** ▷ Time steps, T is the ad campaign duration
    Take action a
    Observe $\boldsymbol{r}$ and $\boldsymbol{s'}$
    Find action $\boldsymbol{a'}$ derived from $Q(s', A)$ acc to ϵ greedy
    $\boldsymbol{Q(s, a) \leftarrow (1 - \alpha)Q(s, a) + \alpha(r + Q(s', a'))}$
    $s \leftarrow s', a \leftarrow a'$
end for

4.4.2 Expected Sarsa

Expected sarsa is similar to sarsa, except the Q table is updated using the expectation of next (state, action) pairs. The expectation helps reduce overall variance compared to sarsa. $\boldsymbol{p_{\epsilon g}(a|s)}$ is the probability of taking action a at state s when following the epsilon greedy policy.

Algorithm 7 Exp Sarsa

Algorithm Parameters: learning rate $\alpha \in (0, 1]$, small $\epsilon > 0$
Initialize $Q \sim \text{Uniform}[0, 1]^{S \times A}$
$\textit{diff} \leftarrow 0, \theta \leftarrow \text{Bernoulli}(p_f)$, $\boldsymbol{s} = (\textit{diff}, \theta)$
for $t = 1, 2 \dots T$ **do**
    Take action $\boldsymbol{a}$ derived from $Q(s, A)$ acc to ϵ greedy
    Observe $\boldsymbol{r}$ and $\boldsymbol{s'}$
    $\boldsymbol{Q(s, a) \leftarrow (1 - \alpha)Q(s, a) + \alpha(r + \sum_a p_{\epsilon g}(a|s') \cdot Q(s', a))}$
    $s \leftarrow s'$
end for

4.4.3 Q learning

In expected sarsa the probabilities $p_{\epsilon g}(a|s)$ are derived from Q according to ϵ greedy. Pure greedy is $p_{pure-g}(a|s) = 1$ for an action maximizing $Q(s, a)$ and 0 for the rest, this gives the famous Q Learning algorithm. Q learning is an **off-policy** algorithm as it uses ϵ greedy for its action selection and pure greedy for the Q table update.

Algorithm 8 Q learning

Algorithm Parameters: learning rate $\alpha \in (0, 1]$, small $\epsilon > 0$
Initialize $Q \sim \text{Uniform}[0, 1]^{S \times A}$
$\textit{diff} \leftarrow 0, \theta \leftarrow \text{Bernoulli}(p_f)$, $\boldsymbol{s} = (\textit{diff}, \theta)$
for $t = 1, 2 \dots T$ **do**
    Take action $\boldsymbol{a}$ derived from $Q(s, A)$ acc to ϵ greedy
    Observe $\boldsymbol{r}$ and $\boldsymbol{s'}$
    $\boldsymbol{Q(s, a) \leftarrow (1 - \alpha)Q(s, a) + \alpha(r + \max_a Q(s', a))}$
    $s \leftarrow s'$
end for

4.4.4 Double Q learning

A known theoretical problem with plain TD(0) updates is maximization bias. A maximum over estimated values(be it using ϵ greedy, pure greedy) is used as an estimate of the max value. If the number of time steps T is not large(see [11]) this leads to a positive bias i.e the estimate $Q(s,a)$ is larger than $Q^*(s,a)$. One solution is to use two Q tables, both estimating $Q^*(s,a)$ but at a given time step only one table will be updated. The action is chosen using an ϵ greedy policy on $Q_1 + Q_2$

Algorithm 9 Double Q learning

Algorithm Parameters: learning rate $\alpha \in (0,1]$, small $\epsilon > 0$
Initialize $Q_1, Q_2 \sim \text{Uniform}[0,1]^{S \times A}$
$diff \leftarrow 0, \theta \leftarrow \text{Bernoulli}(p_f)$, $\boldsymbol{s} = (diff, \theta)$
for $t = 1, 2 \dots T$ **do**
    Take action $\boldsymbol{a}$ derived from $Q_{combined} := (Q_1(s,A) + Q_2(s,A))$ acc to ϵ greedy
    Observe $\boldsymbol{r}$ and $\boldsymbol{s'}$
    if $Bernoulli(0.5) == 1$ **then** ▷ Fair coin flip decides which Q table is to be updated
        $\boldsymbol{Q_1(s,a) \leftarrow (1-\alpha)Q_1(s,a) + \alpha(r + Q_2(s', \arg\max_a Q_1(s',a))}$
    else
        $\boldsymbol{Q_2(s,a) \leftarrow (1-\alpha)Q_2(s,a) + \alpha(r + Q_1(s', \arg\max_a Q_2(s',a))}$
    end if
    $s \leftarrow s'$
end for

5 Experiments

Here we describe the empirical performance of the algorithms mentioned in Section 4.2.

5.1 Simulation parameters

We consider discrete bids $\in \{0, \frac{1}{n}, \frac{2}{n} \dots, 1.0\}$, $n = 100$. The other advertiser's bids follow scaled Beta binomial distributions, the scaling factor being $\frac{1}{n}$. i.e, B_{other} is of the form $\frac{1}{n} BetaBinom(n, \alpha, \beta)$ [1]. This means each of their bids lie $\in [0,1]$ and are discrete. Moreover, the choices of α, β can make the distribution take many shapes.

According to the analysis in section 3.4, what is more important for the bidding dynamics is the distributions for D_m and D_f - the maximum of the other advertiser's bids when the user is male or female. For specific choices of α, β, the distributions for D_m, D_f resemble normal distributions (see Appendix 8.2).

The total number of bidders N is 50 (including the constrained advertiser), probability of male p_m is set to 0.5 , the ad campaign duration T is 10000 rounds and absolute parity constraint $K = 5$. We indeed have results for various other parameters but the above choices are good for understanding the general empirical performance of the various algorithms.

5.2 Full information

We first observe what the cumulative reward in the full information setting is. Broadly speaking we are in 2 cases - 1)The total reward from truthful bidding **is not comparable** to the optimal total reward. 2)The total reward from truthful bidding **is comparable** to the optimal total reward.

[1] Do not confuse this α with the learning rate of TD learning

Scenario	v_m	v_f	$\mathbb{E}(D_m)$	$\mathbb{E}(D_f)$	TR_{opt}	TR_{opt} *simulated*	$TR_{truthful}$	$TR_{truthful}$ *simulated*
1	0.4	0.7	0.146	0.855	471.80	471.29 ± 3.57	1.27	1.27 ± 0.075
2	0.4	0.7	0.265	0.365	2138.56	2136.91 ± 13.08	2131.68	2130.07 ± 13.46
3	0.4	0.7	0.672	0.855	$4.97 \times 10^{-11} \simeq 0$	0	$4.97 \times 10^{-11} \simeq 0$	0

Table 5.1: 3 scenarios

In the table above, TR_{opt} is the optimal expected total reward(precomputed by value iteration). TR_{opt} *simulated* is obtained after 50 simulation runs, in each run the bidding policy is given by (3.7), the ± refers to the standard deviation from the mean of these 50 simulations. $TR_{truthful}$ and $TR_{truthful}$ *simulated* are obtained similarly for truthful bidding.

Explanation for observations: Recall the form of reward (2.1)

- In scenario 1 since $v_m > \mathbb{E}[D_m]$ and $\mathbb{E}[D_f] > v_f$ the optimal policy has an incentive to overbid for females in order to win males for whom it underbids [2]. Whereas truthful bidding performs poorly.
- For scenario 2, $v_m > \mathbb{E}[D_m]$ and $v_f > \mathbb{E}[D_f]$, thus bidding truthfully still wins many auctions with positive reward. Thus optimal and truthful bidding are comparable, both giving a high total reward.
- In scenario 3, $v_m < \mathbb{E}[D_m]$ and $v_f < \mathbb{E}[D_f]$, there is no incentive to overbid for either gender by the advertiser as it cannot obtain the other gender at a positive reward. Here optimal and truthful bidding are comparable and both give $\simeq 0$ total reward. In addition, the simulated results have no deviation from 0, i.e they never won an auction.

5.3 Comparing learning algorithms

We now compare the performance by the learning algorithms for the 3 scenarios in table 5.1. Each of the algorithms is run 50 times, $AR_{alg}(t)$ the average reward for an algorithm in slot t is calculated [3].

$$CR_{alg}(t) := \sum_{t'=1}^{t} AR_{alg}(t') \tag{5.1}$$

$$Reg_{\mathcal{L}}(t) := CR_{opt}(t) - CR_{\mathcal{L}}(t) \tag{5.2}$$

Where $CR_{alg}(t)$ is cumulative reward for an algorithm till time t (averaged over 50 runs), $Reg_{\mathcal{L}}(t)$ is the regret of the learning algorithm. Note that regret $\Delta(M, \mathcal{L}, s, T)$ in (4.2) is defined only for T not for $t \in \{1, \ldots, T-1\}$. Thus technically speaking $Reg_{\mathcal{L}}$ is a valid estimator of Δ only at the last slot T.
We display two kinds of plots 1)$CR_{alg}(t)$ vs t. 2) $Reg_{\mathcal{L}}(t)$ vs t.

[2] Underbidding and overbidding refers to bidding under or over the true value
[3] Reward in slot t across all 50 runs is averaged to get $\mathrm{AR}_{alg}(t)$

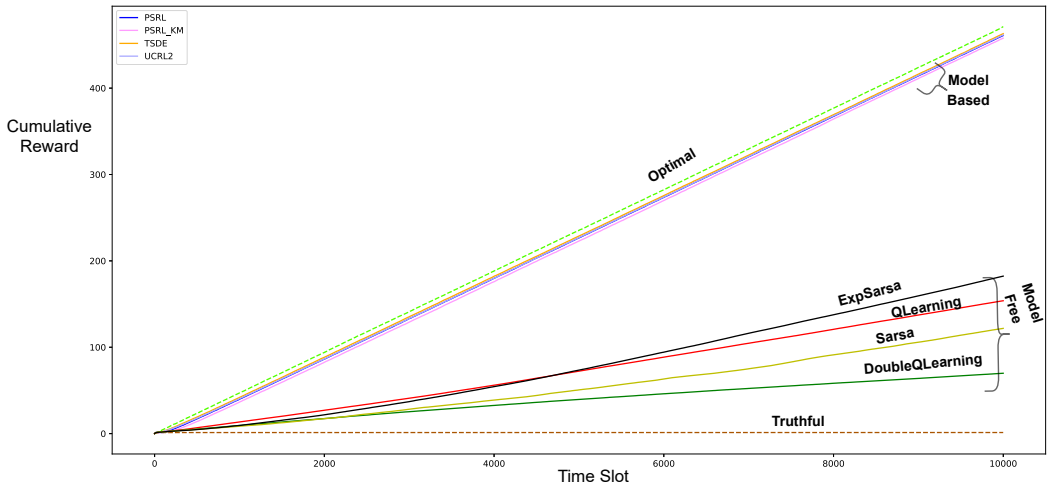


Figure 5.1: Cumulative reward for scenario 1

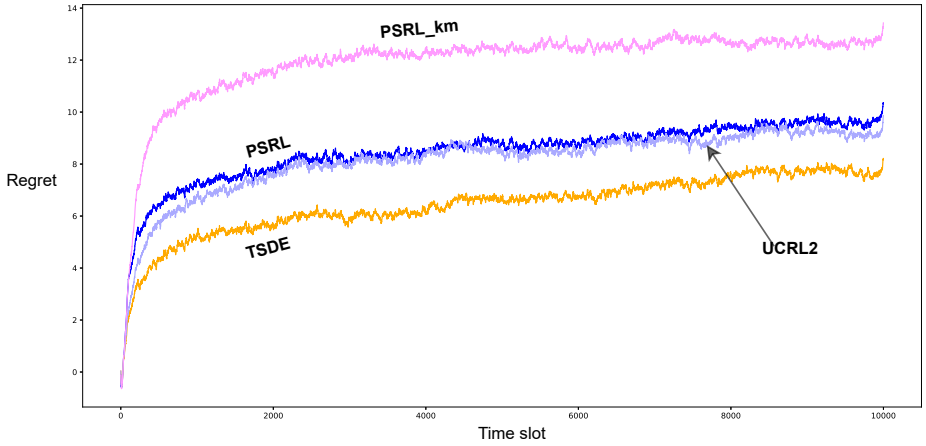


Figure 5.2: Regret for scenario 1

Scenario 1 According to Fig 5.1 the model based algorithms beat the model free algorithms. However the model free algorithms still perform better than the truthful bidding strategy - which gives the lowest cumulative reward. The model based algorithms are close together[4] in Fig 5.1 and can be made apart in Fig 5.2. Among the model-based algorithms *TSDE* gives the lowest regret at time T. In the model free algorithms, Expected sarsa gives the highest cumulative reward at time T.

[4]all are within the curly bracket marker in the figure

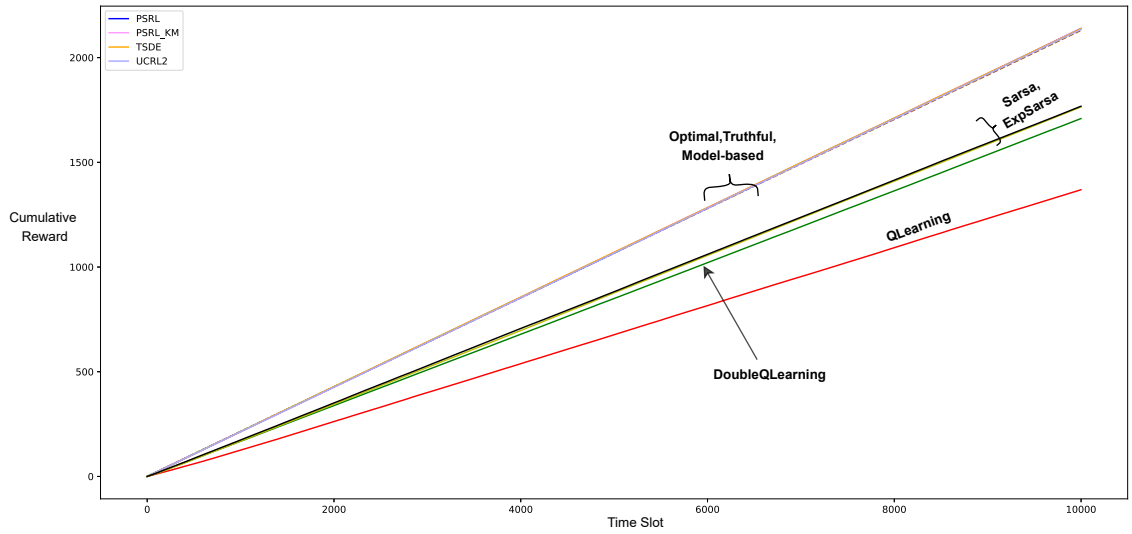


Figure 5.3: Cumulative reward for scenario 2

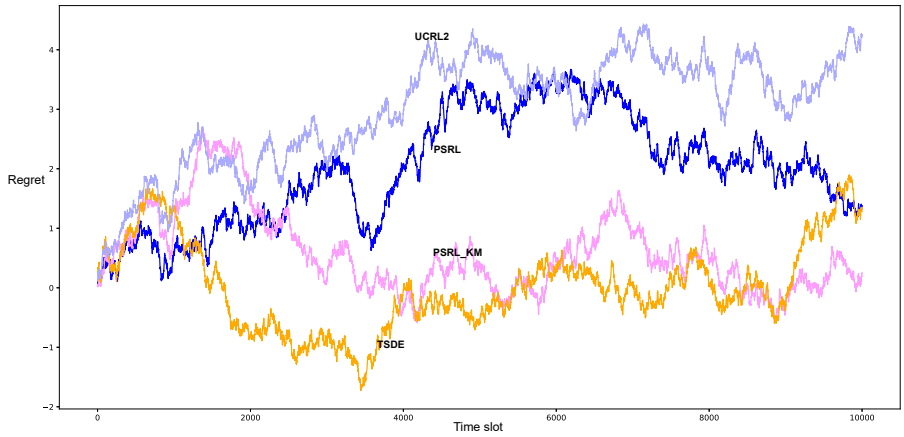


Figure 5.4: Regret for scenario 2

Scenario 2 In Fig 5.3 optimal, truthful bidding and model-based algorithms are all close together and they beat the model free algorithms. The model based algorithms can be made apart in Fig 5.4, but there seems to be no clear trend like in Fig 5.2. *PSRL_km* [5] gives the lowest regret at time T. Among the model free algorithms, expected sarsa and sarsa are close but expected sarsa has the higher cumulative reward at time T.

[5]It runs PSRL when given type 2 feedback

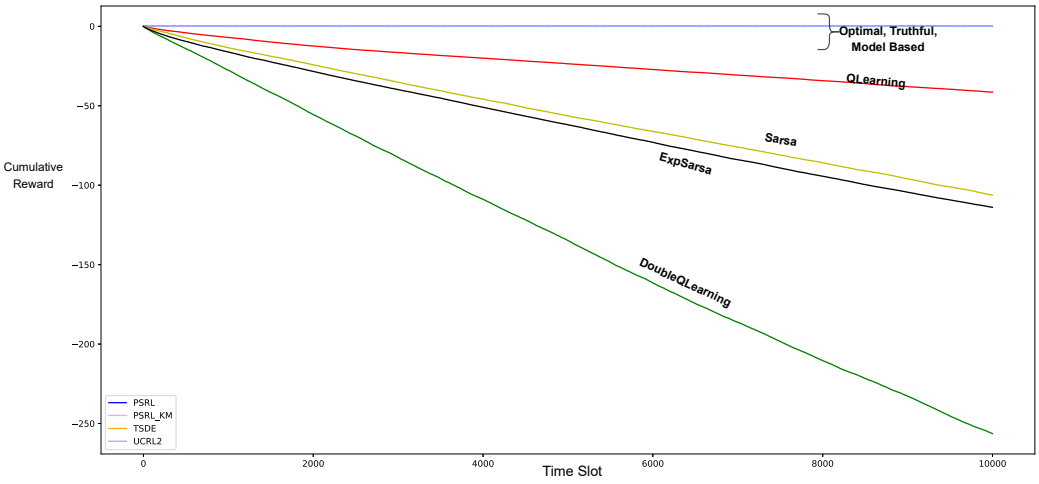


Figure 5.5: Cumulative reward for scenario 3

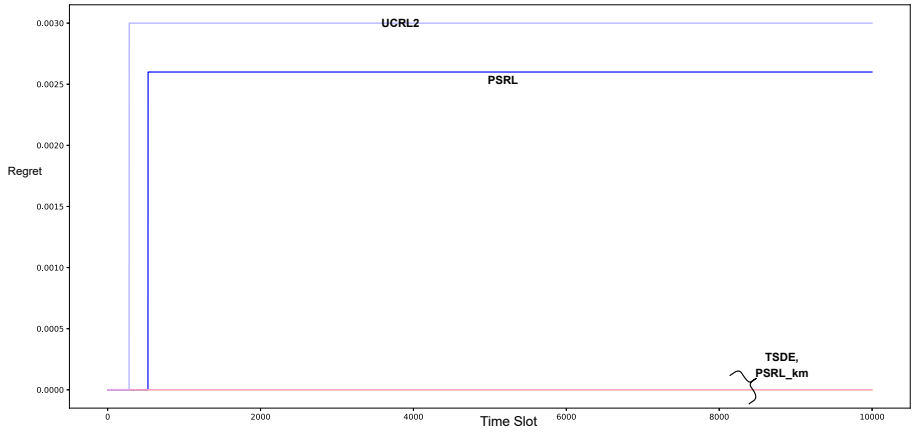


Figure 5.6: Regret for scenario 3

Scenario 3 In Fig 5.5 optimal bidding, truthful bidding, *TSDE* and *PSRL_km* give zero cumulative reward throughout. Thus the regret is zero for *TSDE* and *PSRL_km* as seen in Fig 5.6. *UCRL2* and *PSRL* have a very small non zero regret. All the **model free algorithms fail badly**, giving negative cumulative reward. The best among them, Q learning gives a cumulative reward of $\simeq -40$ at the end of the ad campaign. In scenario 3 the v_m, v_f are such that for any bid that wins an auction the advertiser obtains negative reward with high probability, therefore any exploration step is very costly, this explains the bad performance of the model-free algorithms.

About the regret For the absolute parity MDP with $K = 5$, $|A| = 100$, $T = 10000$, $S := 4K+2$, $D_M = 2K$ we have upper bound on regret $\tilde{O}(DS\sqrt{AT}) = 22 \times 10^4$, which is a trivial upper bound. However the regret computed empirically for the model-based algorithms are much lower than this upper bound,

even close to zero.

6 Conclusion

In conclusion, this thesis described ad auctions with an advertiser side fairness constraint, its modelling as a MDP, the solution of this MDP in the full information setting and dealing with partial information via online learning. Empirically, the model-based algorithms were close to optimal bidding and always beat the model-free algorithms. In addition, the model-free algorithms were unreliable - sometimes performing better than the baseline of truthful bidding and at times much worse.

References

[1] *About Ad Auctions — Facebook Business Help Center*. https://www.facebook.com/business/help/430291176997542?id=561906377587030. (Accessed on 08/24/2020).

[2] *About automated bidding - Google Ads Help*. https://support.google.com/google-ads/answer/2979071?hl=en. (Accessed on 08/24/2020).

[3] Eitan Altman. *Constrained Markov Decision Processes*. 1999.

[4] Peter Auer, Thomas Jaksch, and Ronald Ortner. "Near-optimal Regret Bounds for Reinforcement Learning". In: *Advances in Neural Information Processing Systems 21*. Ed. by D. Koller et al. Curran Associates, Inc., 2009, pp. 89–96. URL: http://papers.nips.cc/paper/3401-near-optimal-regret-bounds-for-reinforcement-learning.pdf.

[5] Peter Auer and Ronald Ortner. "Logarithmic Online Regret Bounds for Undiscounted Reinforcement Learning". In: *Advances in Neural Information Processing Systems 19*. Ed. by B. Schölkopf, J. C. Platt, and T. Hoffman. MIT Press, 2007, pp. 49–56. URL: http://papers.nips.cc/paper/3052-logarithmic-online-regret-bounds-for-undiscounted-reinforcement-learning.pdf.

[6] Santiago R. Balseiro and Yonatan Gur. "Learning in Repeated Auctions with Budgets: Regret Minimization and Equilibrium". In: *Proceedings of the 2017 ACM Conference on Economics and Computation*. EC '17. Cambridge, Massachusetts, USA: Association for Computing Machinery, 2017, p. 609. ISBN: 9781450345279. DOI: 10.1145/3033274.3084088. URL: https://doi.org/10.1145/3033274.3084088.

[7] Elisa Celis, Anay Mehrotra, and Nisheeth Vishnoi. "Toward Controlling Discrimination in Online Ad Auctions". In: ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. Long Beach, California, USA: PMLR, Sept. 2019, pp. 4456–4465. URL: http://proceedings.mlr.press/v97/mehrotra19a.html.

[8] Amit Datta et al. "Discrimination in Online Advertising: A Multidisciplinary Inquiry". In: ed. by Sorelle A. Friedler and Christo Wilson. Vol. 81. Proceedings of Machine Learning Research. New York, NY, USA: PMLR, 23–24 Feb 2018, pp. 20–34. URL: http://proceedings.mlr.press/v81/datta18a.html.

[9] Ronan Fruit et al. "Efficient Bias-Span-Constrained Exploration-Exploitation in Reinforcement Learning". In: *ICML 2018 - The 35th International Conference on Machine Learning*. Vol. 80. Proceedings of Machine Learning Research. Stockholm, Sweden, July 2018, pp. 1578–1586. URL: https://hal.inria.fr/hal-01941206.

[10] R. Gummadi, Peter Key, and Alexandre Proutière. "Repeated Auctions under Budget Constraints : Optimal bidding strategies and Equilibria". In: 2012.

[11] Hado V. Hasselt. "Double Q-learning". In: *Advances in Neural Information Processing Systems 23*. Ed. by J. D. Lafferty et al. Curran Associates, Inc., 2010, pp. 2613–2621. URL: http://papers.nips.cc/paper/3964-double-q-learning.pdf.

[12] E. L. Kaplan and Paul Meier. "Nonparametric Estimation from Incomplete Observations". In: *Journal of the American Statistical Association* 53.282 (1958), pp. 457–481. ISSN: 01621459. URL: http://www.jstor.org/stable/2281868.

[13] Vijay Krishna. *Auction Theory*. 1st ed. Elsevier, 2002. URL: https://EconPapers.repec.org/RePEc:eee:monogr:9780124262973.

[14] Milad Nasr and Michael Carl Tschantz. "Bidding Strategies with Gender Nondiscrimination Constraints for Online Ad Auctions". In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. FAT* '20. Barcelona, Spain: Association for Computing Machinery, 2020, pp. 337–347. ISBN: 9781450369367. DOI: 10.1145/3351095.3375783. URL: https://doi.org/10.1145/3351095.3375783.

[15] Ian Osband and Benjamin Van Roy. "Posterior Sampling for Reinforcement Learning Without Episodes". In: *ArXiv* abs/1608.02731 (2016).

[16] Ian Osband, D. Russo, and B. Roy. "(More) Efficient Reinforcement Learning via Posterior Sampling". In: *ArXiv* abs/1306.0940 (2013).

[17] Yi Ouyang et al. "Learning Unknown Markov Decision Processes: A Thompson Sampling Approach". In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS'17. Long Beach, California, USA: Curran Associates Inc., 2017, pp. 1333–1342. ISBN: 9781510860964.

[18] Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. 1st. USA: John Wiley amp; Sons, Inc., 1994. ISBN: 0471619779.

[19] Richard S. Sutton and Andrew G. Barto. *Introduction to Reinforcement Learning*. 1st. Cambridge, MA, USA: MIT Press, 1998. ISBN: 0262193981.

7 Appendix A - Upper bounding the regret

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**

Our goal is to obtain upper bounds for regret in both settings

- First we show that for any deterministic "continous-bid" policy $\pi : S \to \mathbb{R}^+$ for the MDP $M = (S, A = [0, 1], P, R)$ we can construct a stochastic "discrete-bid" policy $\pi_\varepsilon : S \to \Delta(N')$ [1] for the MDP $M' = (S, A = N', P, R)$ such that π_ε performs close to π.
- An upper bound for regret in the discrete bids settings is obtained by an analysis similar to UCRL2 but the structure of our MDP allows us to obtain a tighter regret bound.
- We use the two above results to obtain an upper bound for regret in 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.

[1] N' is the set $\{n\varepsilon | n \in \mathbb{N} \text{ and } n\varepsilon \in [0, 1]\}$, $\Delta(N')$ indicates we are considering probability distributions over N'

7.1 Continous v/s discrete bids

7.1.1 Bidding policies

Let $\pi : S \to \mathbb{R}^+$ be a deterministic continous bid policy and $s = (k, \theta) \in \mathcal{S}$ be a state. We denote by $r_\pi(s)$ the expected immediate reward of policy π at state s and by $W_\pi(s)$ the probability of winning the auction while in state s (by bidding $\pi(s)$).

$$r_\pi(k,\theta) \coloneqq \bar{R}(\theta, \pi(k,\theta)) = (v_\theta - \pi(k,\theta))P_{\text{win},\theta}(\pi(k,\theta)) + \int_0^{\pi(k,\theta)} P_{win,\theta}(u)du$$
$$W_\pi(k,\theta) \coloneqq P_{\text{win},\theta}(\pi(k,\theta))$$

7.1.2 Discrete bidding policy

Let π be a policy and let $\varepsilon > 0$. We will construct a policy π_ε whose bids are multiple of ε and whose performance is close to π. We do it by using the following lemma:

Lemma 7. *Let $\pi : S \to \mathbb{R}^+$ be a (deterministic) continous bid policy. Then there exists a (randomized) discrete bidding policy π_ε whose bids are restricted to $\varepsilon\mathbb{N}$ such that:*

$$r_{\pi_\varepsilon}(k,\theta) \geq r_\pi(k,\theta) - 2\varepsilon$$
$$W_{\pi_\varepsilon}(k,\theta) = W_\pi(k,\theta)$$

Note that for randomized policy $\pi_\varepsilon : S \to \Delta(N')$

$$r_{\pi_\varepsilon}(k,\theta) \coloneqq \sum_{\forall i} P(\pi_\varepsilon(k,\theta) = b_i)\bar{R}(\theta, b_i)$$
$$W_{\pi_\varepsilon}(k,\theta) \coloneqq \sum_{\forall i} P(\pi_\varepsilon(k,\theta) = b_i)P_{win,\theta}(b_i)$$

Proof. Let $(k,\theta) \in S$ and let $n \in \mathbb{N}$ be such that $n\varepsilon \leq \pi(k,\theta) < (n+1)\varepsilon$. Consider a policy that, in this state (k,θ) bids $n\varepsilon$ with probability p and $(n+1)\varepsilon$ with probability $1-p$. Consider:

$$pP_{\text{win},\theta}(n\varepsilon) + (1-p)P_{\text{win},\theta}((n+1)\varepsilon).$$

As $b \mapsto P_{\text{win},\theta}(b)$ is a non-decreasing function, there exists p such that the above expression equals $P_{\text{win},\theta}(\pi(k,\theta))$. This defines the policy π_ε.

By definition, for all states, we have $W_{\pi_\varepsilon}(k,\theta) = W_\pi(k,\theta)$. Moreover, the expected immediate reward of this policy is

$$
\begin{aligned}
r_{\pi_\varepsilon}(k,\theta) &= p\Big[(v_\theta - n\varepsilon)P_{win,\theta}(n\varepsilon) + \int_0^{n\varepsilon} P_{win,\theta}(u)du\Big] \\
&\quad + (1-p)\Big[(v_\theta - (n+1)\varepsilon)P_{win,\theta}((n+1)\varepsilon) + \int_0^{(n+1)\varepsilon} P_{win,\theta}(u)du\Big] \\
&= (v_\theta - n\varepsilon)[pP_{win,\theta}(n\varepsilon) + (1-p)P_{win,\theta}((n+1)\varepsilon)] - (1-p)P_{win,\theta}((n+1)\varepsilon)\varepsilon \\
&\quad + \overbrace{p[\int_0^{n\varepsilon} P_{win,\theta}(u)du] + (1-p)[\int_0^{(n+1)\varepsilon} P_{win,\theta}(u)du]}^{\text{Integral term}} \\
&= (v_\theta - n\varepsilon)P_{win,\theta}(\pi(k,\theta)) - (1-p)P_{win,\theta}((n+1)\varepsilon)\varepsilon + \text{Integral term} \\
&\geq (v_\theta - \pi(k,\theta))P_{win,\theta}(\pi(k,\theta)) - (1-p)P_{win,\theta}((n+1)\varepsilon)\varepsilon + \text{Integral term} \\
&\geq (v_\theta - \pi(k,\theta))P_{win,\theta}(\pi(k,\theta)) - \varepsilon + \text{Integral term} \\
&\geq (v_\theta - \pi(k,\theta))P_{win,\theta}(\pi(k,\theta)) - \varepsilon + \int_0^{\pi(k,\theta)} P_{win,\theta}(u)du - \varepsilon \\
&\geq r_\pi(k,\theta) - 2\varepsilon
\end{aligned}
$$

where the first inequality holds because $\pi(k,\theta) \geq n\varepsilon$. The second last inequality is obtained using the following observation

$$
\begin{aligned}
&p[\int_0^{n\varepsilon} P_{win,\theta}(u)du] + (1-p)[\int_0^{(n+1)\varepsilon} P_{win,\theta}(u)du] \\
&= \Big[\overbrace{\int_0^{(n+1)\varepsilon} P_{win,\theta}(u)du}^{\geq \int_0^{\pi(k,\theta)} P_{win,\theta}(u)du}\Big] + \Big[\overbrace{-p(\int_{n\varepsilon}^{(n+1)\varepsilon} P_{win,\theta}(u)du)}^{\geq -\varepsilon}\Big]
\end{aligned}
$$

□

7.1.3 Discrete-bids policies are almost optimal

Lemma 8. *For the continous bids MDP $M = (S, A = [0,1], P, R)$, let $\pi : S \to \mathbb{R}^+$ be an optimal policy for the average reward criteria and ρ^*_M be its optimal average reward. Then for the discrete bids MDP $M' = (S, A = N', P, R)$, there exists an optimal (deterministic) discrete-bids policy $\pi' : S \to N'$ with optimal average reward $\rho^*_{M'}$ such that $\rho^*_{M'} \geq \rho^*_M - 2\varepsilon$.*

Proof. We are comparing optimality results for two MDPs $M : (S, A = [0,1], R)$, $M' : (S, A = N', R)$. We know by Lemma 7 that for any continous bid policy in M we can construct a randomized discrete bidding policy π_ε in M' with bids randomized over $N' = \{n\varepsilon | n \in \mathbb{N} \text{ and } n\varepsilon \in [0,1]\}$, so using $r_{\pi_\varepsilon}(k,\theta) \geq r_\pi(k,\theta) - 2\varepsilon$, definition 4.1 $\implies$

$$
\begin{aligned}
\overbrace{\sum_{s=(k,\theta)\in S} \mu_{\pi_\varepsilon}(s) r_{\pi_\varepsilon}(k,\theta)}^{\rho_\varepsilon} &= \sum_{s=(k,\theta)\in S} \mu_\pi(s) r_{\pi_\varepsilon}(k,\theta) \\
&\geq \sum_{s=(k,\theta)\in S} \mu_\pi(s)[r_\pi(k,\theta) - 2\varepsilon] \qquad \geq \overbrace{\sum_{s=(k,\theta)\in S} \mu_\pi(s) r_\pi(k,\theta)}^{\rho^*_M} - 2\varepsilon
\end{aligned}
$$

ρ_ε refers to the average reward for π_ε. For M' it is also known that there exists a deterministic discrete bidding policy $\pi' : S \to N'$ that gives optimal average reward $\rho^*_{M'}$.

$$\rho^*_{M'} \geq \rho_\varepsilon$$
$$\rho_\varepsilon \geq \rho^*_M - 2\varepsilon$$

□

7.2 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 [2]. Here D refers to the diameter of the MDP M' i.e $D = D_{M'}$(see 3.3).

7.2.1 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)[3]
the number of times (gender= θ and action $= a$) in episode k is denoted by $v_k(\theta, a)$

Algorithm 10 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 $\forall\theta, \forall a$:

[2] which has a regret bound of $\tilde{O}(DS\sqrt{AT})$
[3] $\theta = 1$ with probability $p_f, \theta = 0$ with probability $p_m, p_m + p_f = 1$

$$|\tilde{r}(\theta,a) - \hat{r}_k(\theta,a)| \le d'(\theta,a) = \sqrt{\frac{c_1 \log(c_1' A t_k/\delta)}{\max\{1, N_k(\theta,a)\}}} \tag{7.1}$$

$$|\tilde{p}_{win}(\theta,a) - \hat{p}_{win,k}(\theta,a)| \le d(\theta,a) = \sqrt{\frac{c_2 \log(c_2' A t_k/\delta)}{\max\{1, N_k(\theta,a)\}}} \quad [4] \tag{7.2}$$

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 11.

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

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

$$\tilde{p}(s'|s = (\mathit{diff}, \theta), a) \coloneqq \begin{cases} p_m \tilde{p}_{win}(\theta,a) & \text{if } s' = (\mathit{diff} + (-1)^\theta, m) \\ p_f \tilde{p}_{win}(\theta,a) & \text{if } s' = (\mathit{diff} + (-1)^\theta, f) \\ p_m(1 - \tilde{p}_{win}(\theta,a)) & \text{if } s' = (\mathit{diff}, m) \\ p_f(1 - \tilde{p}_{win}(\theta,a)) & \text{if } s' = (\mathit{diff}, f) \\ 0 & \text{for any other } s' \end{cases} \tag{7.3}$$

Lemma 9. *If* $\forall(\theta,a)|\hat{p}_{win,k}(\theta,a) - p_{win}(\theta,a)| \le \epsilon$ *then* $\forall(s,a)\, \|\hat{p}_k(.|s,a) - p(.|s,a)\|_1 \le 2\epsilon$ [5]
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 [4]) is the following maximization, the second equation(7.4) is what it looks like for our MDP

[4] See the relation between p_{win} and $p(.|s,a)$ in Lemma 9, note this gets rid of the explicit l^1 norm condition in UCRL2
[5] For the edge states the difference is exactly zero, since we know p_m and p_f

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

$$u_{i+1}(s) = \max_{a \in A} \Big\{ \tilde{r}(s,a) + \max_{\tilde{p}_{win}(\theta,a) \in \text{Polytope}} \Big\{ \tilde{p}_{win}(\theta,a)\phi(s) + c(s) \Big\} \Big\} \tag{7.4}$$

Where the polytope is given by Eq(7.2)(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 $\boxed{\tilde{p}_{win}(\theta,a)\psi(s) + c(s)}$
$c(s), \psi(s)$ are terms we get by collecting $u_i(s')$ i.e the u_i values of the next 4 states from s(like in Eq(3.5)).
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([4])

Theorem 10. *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

7.2.2 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 $X_i \in [a, b]$
$P(S_n \leq \mathbb{E}[S_n] - t) \leq \exp(-\frac{2t^2}{n(b-a)^2})$

$$P\left(\sum_{t=1}^{T} R(\theta_t, a_t) \leq \sum_{\theta,a} N(\theta, a)\bar{R}(\theta, a) - \sqrt{z_1 T log(\frac{z_2 T}{\delta})}\right)$$
$$\leq \exp\left(-\frac{z_1 log(z_2 T/\delta)}{2}\right) \tag{7.5}$$

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

$$\boxed{\sum_{k=1}^{m} \Delta_k \mathbb{1}_{M \notin \mathcal{M}_k} + \sum_{k=1}^{m} \Delta_k \mathbb{1}_{M \in \mathcal{M}_k} + \sqrt{\frac{5}{2}T \log(\frac{8T}{\delta})}} \tag{7.6}$$

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

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 (7.2) and (7.1) (recall $c_1 = 14, c_1' = 2$). $M \notin M(t)$ if $\bar{R}(\theta,a), p_{win}(\theta,a)$ do not lie in (7.1), (7.2) 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 (7.2), (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 $P(M \notin M(t)) \leq \frac{\delta}{30t^6} \leq \frac{\delta}{15t^6}$ □

7.2.4 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 10 was in state s and took action a during episode k. [6]

Theorem 10 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 $\rho^* \leq \tilde{\rho}_k + \frac{1}{\sqrt{t_k}}$

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

[6] 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

The boxed term can be rewritten [7] 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

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

$$\Delta_k = \sum_{s,a} v_k(s,a)(\tilde{\rho}_k - \tilde{r}_k(s,a)) + \sum_{s,a} v_k(s,a)(\tilde{r}_k(s,a) - \bar{R}(s,a)) + \sum_{s,a} \frac{v_k(s,a)}{\sqrt{t_k}} \tag{7.7}$$

$$\Delta_k \leq \overbrace{\boldsymbol{v_k(\tilde{P}_k - I)u_i}}^{\text{term 1}} + \overbrace{\sum_{s,a} v_k(s,a)(\tilde{r}_k(s,a) - \bar{R}(s,a))}^{\text{term 2}} + \overbrace{2\sum_{s,a} \frac{v_k(s,a)}{\sqrt{t_k}}}^{\text{term 3}} \tag{7.8}$$

$\boldsymbol{v_k} \coloneqq v_k\big(\big(s, \tilde{\pi}_k(s)\big)\big)_s$ 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 (7.3)).

term $3 = 2\sum_{\theta,a} \frac{v_k(\theta,a)}{\sqrt{t_k}}$, term 2 [8] $\leq 2\sum_{s,a} v_k(s,a)\sqrt{\frac{c_1 \log(c_1' A t_k/\delta)}{\max\{1, N_k(\theta_s,a)\}}} = 2\sum_{\theta,a} v_k(\theta,a)\sqrt{\frac{c_1 \log(c_1' A t_k/\delta)}{\max\{1, N_k(\theta,a)\}}}$, also $\max\{1, N_k(\theta,a)\} \leq t_k \leq T$

$$\text{term2+term3} \leq \Big(2\sqrt{c_1 \log\Big(\frac{c_1' AT}{\delta}\Big)} + 2\Big) \sum_{\theta,a} \frac{v_k(\theta,a)}{\sqrt{\max\{1, N_k(\theta,a)\}}} \tag{7.9}$$

$w_k(s) \coloneqq u_i(s) - \frac{\min_s u_i(s) + \max_s u_i(s)}{2}$
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 $\boldsymbol{v_k(\tilde{P}_k - I)w_k} = \boldsymbol{v_k(\tilde{P}_k - P_k)w_k} + \boldsymbol{v_k(P_k - I)w_k}$. Here $\boldsymbol{P_k} \coloneqq \big(p(s'|s, \tilde{\pi}_k(s))\big)_{s,s'}$
First we bound $\boldsymbol{v_k(\tilde{P}_k - P_k)w_k}$

$$\begin{aligned}
v_k(\tilde{P}_k - P_k)w_k &\leq \sum_s v_k(s, \tilde{\pi}_k(s)) \cdot \|\tilde{p}_k(.|s, \tilde{\pi}_k(s)) - p(.|s, \tilde{\pi}_k(s))\|_1 \cdot \|w_k\|_\infty \\
&\leq \sum_s v_k(s, \tilde{\pi}_k(s)) \cdot 4\sqrt{\frac{c_2 \log(c_2' A t_k/\delta)}{\max\{1, N_k(\theta_s, \tilde{\pi}_k(s))\}}} \cdot \frac{D}{2} \\
&\leq 2D\sqrt{c_2 \log(c_2' A t_k/\delta)} \boxed{\sum_s \frac{v_k(s, \tilde{\pi}_k(s))}{\sqrt{\max\{1, N_k(\theta_s, \tilde{\pi}_k(s))\}}}} = 2D\sqrt{c_2 \log(c_2' A t_k/\delta)} \sum_{\theta,a} \frac{v_k(\theta,a)}{\sqrt{\max\{1, N_k(\theta,a)\}}} \\
&\leq 2D\sqrt{c_2 \log(c_2' AT/\delta)} \sum_{\theta,a} \frac{v_k(\theta,a)}{\sqrt{\max\{1, N_k(\theta,a)\}}}
\end{aligned} \tag{7.10}$$

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 [4])

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

[8] 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 $\|\tilde{p}_k(.|s, \tilde{\pi}_k(s)) - p(.|s, \tilde{\pi}_k(s))\|_1 \leq \|\tilde{p}_k(.|s, \tilde{\pi}_k(s)) - \hat{p}_k(.|s, \tilde{\pi}_k(s))\|_1 + \|\hat{p}_k(.|s, \tilde{\pi}_k(s)) - p(.|s, \tilde{\pi}_k(s))\|_1 \leq 4d(\theta_s, \tilde{\pi}_k(s))$

$$v_k(P_k - I)w_k \leq D + \sum_{t=t_k}^{t_{k+1}-1} X_t$$

$$\sum_{k=1}^{m} v_k(P_k - I)w_k \leq mD + \sum_{t=1}^{T} X_t \tag{7.11}$$

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

Lemma 12. *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

$$N(\theta, a) = \sum_{k=1}^{m} v_k(\theta, a) \geq 2^{K(\theta,a)} - 1$$

$$T = \sum_{\theta,a} N(\theta, a) \geq \sum_{\theta,a} \Big(2^{K(\theta,a)} - 1\Big) \tag{7.12}$$

Also $\sum_{\theta,a} K(\theta, a) \geq m - 1 - |\theta| \cdot A \geq m - 1 - 2A$. [9]
$\sum_{\theta,a} 2^{K(\theta,a)} \geq 2A\Big(\prod_{\theta,a} 2^{K(\theta,a)}\Big)^{1/2A} = 2A \cdot 2^{\sum_{\theta,a} 2^{K(\theta,a)/2A}} \geq 2A2^{\frac{m-1}{2A}-1}$ [10]

Finally from (7.12) 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 $m \leq 1 + 2A + 2A\log_2(\frac{T}{A}) \leq 2A(2 + \log_2(\frac{T}{A})) \leq \boxed{2A(\log_2(\frac{4T}{A}))}$. □

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

Returning to Eq(7.8) the sum of term1,term2,term3 is upper bounded(see below) with probability atleast $1 - \frac{\delta}{12T^{5/4}}$
We use (7.10), (7.11), (7.9). [11]

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

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

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 $1 - \frac{\delta}{12T^{5/4}}$

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

7.2.6 Sum of Δ_k over all episodes

Recall Eq (7.6) and the previous results, Thus with probability atleast
$1 - \frac{\delta}{12T^{5/4}} - \frac{\delta}{12T^{5/4}} - \frac{\delta}{12T^{5/4}} = 1 - \frac{\delta}{4T^{5/4}} \geq 1 - \delta$

$$\Delta(s_1, T) \leq \sum_{k=1}^{m} \Delta_k \mathbb{1}_{M \notin M_k} + \sum_{k=1}^{m} \Delta_k \mathbb{1}_{M \in M_k} + \sqrt{\frac{5}{2} T \log(\frac{8T}{\delta})} \tag{7.13}$$
$$\leq \sqrt{\frac{5}{2} T \log(\frac{8T}{\delta})} + 2\sqrt{T} + D\sqrt{\frac{5}{2} T \log(\frac{8T}{\delta})} + 2DA \log_2(\frac{4T}{A})$$
$$+ \Big(2D\sqrt{14 \log(\frac{2AT}{\delta})} + 2\Big)(\sqrt{2}+1)\sqrt{2AT}$$

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 (7.5)
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 $T > 25^2 A \log(\frac{T}{\delta}) \iff A < \frac{1}{25 \log(\frac{T}{\delta})} \sqrt{AT \log(\frac{T}{\delta})}$ [12], also $log_2(4T) \leq 2\log(T)$ therefore $2DA \log_2(\frac{4T}{A}) \leq$

[12] Also notice how T has to be > 100 to satisfy $T > 625A \log(\frac{T}{\delta}) > 1250 \log(T)$

$\frac{4}{25}D\sqrt{AT\log(\frac{T}{\delta})}$
Notice that for $T > 25^2 A\log(\frac{T}{\delta})$ [13], $\log(\frac{2AT}{\delta}) \le 2\log(\frac{T}{\delta})$ and $\log(\frac{8T}{\delta}) \le 2\log(\frac{T}{\delta})$
Also $A \ge 2$, $\frac{1}{\sqrt{A}} \le \frac{1}{\sqrt{2}}$, Thus using Eq (7.13), we have for $T > 1$ with probability atleast $1-\delta$

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

7.3 Regret bound for continous bids

What we have now:

1. For the discrete bids mdp $M' = (S, A = N', P, R)$ a $\tilde{O}(D\sqrt{AT})$ bound[14] for $T\rho^*_{M'} - \sum_{t=1}^{T} R(\theta_t, a_t)$ $\rho^*_{M'}$ is the optimal average reward obtained by a deterministic discrete bids policy $\pi : S \to N'$.
2. By Lemma 8 $T\rho^*_{M'} \ge T\rho^*_M - 2\varepsilon T$. Where ρ^*_M is the optimal average reward for the MDP $M = (S, A = [0,1], P, R)$

Notice that for $A = N', |A| = 1/\varepsilon$

$$\begin{aligned}
&T\rho^*_{M'} - \sum_{t=1}^{T} R(\theta_t, a_t) \le 50D\sqrt{AT\log(\frac{T}{\delta})} \\
&T\rho^*_M - T\rho^*_{M'} \le 2\varepsilon T \\
&\implies T\rho^*_M - \sum_{t=1}^{T} R(\theta_t, a_t) \le 2\varepsilon T + 50D\sqrt{\frac{1}{\varepsilon}T\log(\frac{T}{\delta})} \qquad (7.14)
\end{aligned}$$

Setting ε to $T^{-1/3}$ Eq(7.14) [15] can be written as

$$\begin{aligned}
T\rho^*_M - \sum_{t=1}^{T} R(\theta_t, a_t) &\le 2T\cdot T^{-1/3} + 50D\sqrt{T\cdot T^{1/3}\log(\frac{T}{\delta})} \\
&\le 2T^{2/3} + 50DT^{2/3}\sqrt{\log(\frac{T}{\delta})} \\
&\le 51DT^{2/3}\sqrt{\log(\frac{T}{\delta})}
\end{aligned}$$

[13] the constraint implies $T > 2A$, why? simple proof by contradiction
[14] Nowhere in the analysis for UCRL2 adapted did we use that $A = N'$ specifically, so this regret bound is valid for any mdp $M : (S, A, P, R)$ long as A is finite and has bounded diameter
[15] How did we get this exponent?, Consider $\varepsilon = T^x$. Solution of $1 + x = \frac{1-x}{2}$, Notice how a lower/higher x value will dominate the other term(in the rhs of (7.14))

The above is a $\tilde{O}(DT^{2/3})$ upper bound.

8 Appendix B

8.1 Important results for 2^{nd} price auctions

Weak dominance of truthful bidding in a 2^{nd} Price auction Suppose advertiser j 's true value is v_j, and it considers bidding $b_j > v_j$. Let d denote the highest bid of the other bidders $i \neq j$. There are three possible outcomes from j's perspective: (i) $d > b_j, v_j$;(ii) $b_j > d > v_j$; or(iii) $b_j, v_j > d$. In the event of the first or third outcome, j would have done equally well to bid v_j rather than $b_j > v_j$. However, in case (ii), j will win and pay more than its value if it bids b_j(thereby obtaining negative reward), something that won't happen if it bids v_j. Thus, j does better to bid v_j than $b_j > v_j$. A similar argument shows that j also does better to bid v_j than to bid $b_j < v_j$

Maximization lemma for bids Consider the function $f_\phi(x) = (\phi - x)g(x) + \int_0^x g(u)du$ where $g(x)$ is a cumulative distribution function with support $\in [0, 1]$. Its derivative wrt x, $f'_\phi(x)$

$$\frac{d}{dx}\Big((\phi - x)g(x)\Big) + \frac{d}{dx}\Big(\int_0^x g(u)du\Big) = -g(x) + (\phi - x)g'(x) + g(x) = (\phi - x)g'(x) \tag{8.1}$$

Note $g'(x)$ is always ≥ 0. Thus $f_\phi(x)$ for $x < \phi$ is non-decreasing as $(\phi - x)g'(x) \geq 0$ and for $x > \phi$ is non-increasing as $(\phi - x)g'(x) \leq 0$. At $x = \phi$, $f_\phi(x) = \int_0^\phi g(u)du$.
Thus for continuous x $f_\phi(x)$ is maximized at ϕ, and for discrete x it is the point on either side of ϕ i.e $\lceil\phi\rceil_{closest}$, $\lfloor\phi\rfloor_{closest}$.

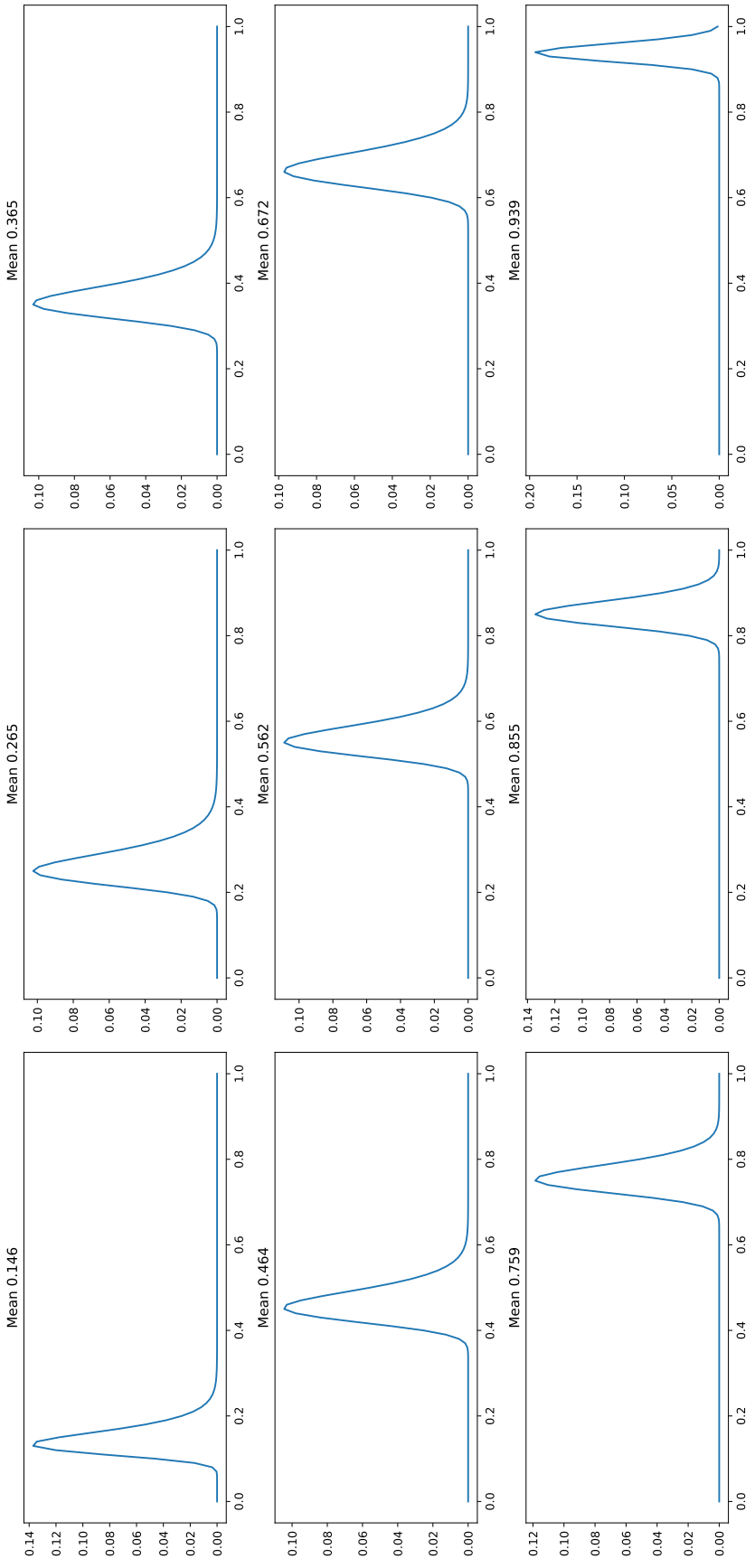
8.2 Pmf parameters

Alpha Beta table The following table gives the mean and standard deviation of $D = \max\{B_1, \ldots, B_{49}\}$ where each $B_i \sim \frac{1}{100} BetaBinom(100, \alpha, \beta)$.

Set	alpha	beta	mean	standard deviation
1	2	47	0.14627666398598949	0.03250368101917681
2	4	34	0.2651640367653201	0.04253368856779681
3	9	38	0.3655329680234962	0.041353136357949356
4	15	38	0.4648532276760166	0.04030517882433639
5	22	36	0.5624004400909328	0.03843523136449991
6	16	19	0.6729414850007737	0.0422982170781183
7	27	20	0.7598823608667473	0.03444315362508511
8	25	12	0.8553404687956032	0.02987724338602748
9	27	7	0.9394899406797824	0.020121239642026383

The plots for all the parameter sets can be seen on the next page. In particular for distributions D_m, D_f for the experiments:

- In scenario 1 - Set 1 and 8 were chosen
- In scenario 2 - Set 2 and 3 were chosen
- In scenario 3 - Set 6 and 8 were chosen


Mean 0.146
Mean 0.265
Mean 0.365
Mean 0.464
Mean 0.562
Mean 0.672
Mean 0.759
Mean 0.855
Mean 0.939