Online Second Price Auction with Semi-bandit Feedback Under the Non-Stationary Setting

Haoyu Zhao,[1] **Wei Chen**[2]**.**
[1]IIIS, Tsinghua University, Beijing, China
zhaohy16@mails.tsinghua.edu.cn
[2]Microsoft Research, Beijing, China
weic@microsoft.com

Abstract

In this paper, we study the non-stationary online second price auction problem. We assume that the seller is selling the same type of items in T rounds by the second price auction, and she can set the reserve price in each round. In each round, the bidders draw their private values from a joint distribution unknown to the seller. Then, the seller announced the reserve price in this round. Next, bidders with private values higher than the announced reserve price in that round will report their values to the seller as their bids. The bidder with the highest bid larger than the reserved price would win the item and she will pay to the seller the price equal to the second-highest bid or the reserve price, whichever is larger. The seller wants to maximize her total revenue during the time horizon T while learning the distribution of private values over time. The problem is more challenging than the standard online learning scenario since the private value distribution is non-stationary, meaning that the distribution of bidders' private values may change over time, and we need to use the *non-stationary regret* to measure the performance of our algorithm. To our knowledge, this paper is the first to study the repeated auction in the non-stationary setting theoretically. Our algorithm achieves the non-stationary regret upper bound $\tilde{\mathcal{O}}(\min\{\sqrt{\mathcal{S}T}, \bar{\mathcal{V}}^{\frac{1}{3}}T^{\frac{2}{3}}\})$, where $\mathcal{S}$ is the number of switches in the distribution, and $\bar{\mathcal{V}}$ is the sum of total variation, and $\mathcal{S}$ and $\bar{\mathcal{V}}$ are not needed to be known by the algorithm. We also prove regret lower bounds $\Omega(\sqrt{\mathcal{S}T})$ in the switching case and $\Omega(\bar{\mathcal{V}}^{\frac{1}{3}}T^{\frac{2}{3}})$ in the dynamic case, showing that our algorithm has nearly optimal *non-stationary regret*.

1 Introduction

As the Internet is rapidly developing, there are more and more online repeated auctions in our daily life, such as the auctions on the e-Bay website and the online advertisement auctions on Google and Facebook. Perhaps the most studied and applied auction mechanism is the online repeated second price auctions with a reserve price. In this auction format, a seller repeatedly sells the same type of items to a group of bidders. In each round t, the seller selects and announces a reserve price $r^{(t)}$ while the bidders draw their private values $\boldsymbol{v}^{(t)}$ on the item from a joint value distribution, which is unknown to the seller. For each bidder i, if its private value $\boldsymbol{v}_i^{(t)}$ is at least the reserve price $r^{(t)}$, she will submit her bid $\boldsymbol{v}_i^{(t)}$ to the seller; otherwise she will not submit her bid since she would not win if her value is less than the announced reserve price. After the seller collects the bids in this round (if any), she will give the item to the highest bidder, and collect from this winner the payment equal to the value of the second-highest bid or the reserve price, whichever is higher. If no bidder submits bids in this round, that means the reserve price the seller announced is too high, and the seller receives no payment. Such repeated auctions are common in online advertising applications on search engine or social network platforms. The seller's objective is to maximize her cumulative revenue, which is the total payment she collects from the bidders over T rounds. Since the seller does not know the private value distribution of the bidders, the seller has to adjust the reserve price over time, hoping to learn the optimal reserve price.

The above setting falls under the multi-armed bandit framework, where reserve prices can be treated as arms and payments as rewards. As in the multi-armed bandit framework, the performance of an online auction algorithm is measured by its *regret*, which is the difference between the optimal reward that always chooses the best reserve price and the expected cumulative reward of the algorithm. When the distribution of private values does not change over time, results from (Cesa-Bianchi et al. 2017; Zhao and Chen 2019b) can be applied to solve the above problem, whereas the work in (Cesa-Bianchi, Gentile, and Mansour 2015) considers a somewhat different setting where the seller only gets the reward as the feedback but does not see the bids (full-bandit feedback) and the private value distribution of each bidder is i.i.d.

In real-world applications, however, the private value distribution of the bidders may likely change over time, e.g., some important events happen, which greatly influence the market perception. When the private value distribution changes over time, the optimal reserve price will also change and there is no single optimal reserve value. None of the above studies would work under this realistic setting, except resetting the algorithms by human intervention. Since

it is difficult to predict distribution changes, we prefer to have algorithms that could automatically detect distribution changes and adjust their actions accordingly, and still provide nearly optimal performance over the long run.

In this paper, we design the first online learning algorithm for online second price auction with non-stationary distributions of private values. We assume that the private values of the bidders at time t follow the joint distribution $\mathcal{D}_t$, and we assume that r_t^* is the best reserve price at time t. We use *non-stationary regret* to measure the performance of the algorithm, which is the difference between the expected cumulative reward of the best reserve prices at each round and the expected cumulative reward of the algorithm. We use two quantities to measure the changing of the distributions $\{\mathcal{D}_t\}_{t\leq T}$: switchings and total variation. The number of switchings is defined as $\mathcal{S} := 1+\sum_{t=2}^{T}\mathbb{I}\{\mathcal{D}_t \neq \mathcal{D}_{t-1}\}$, and the total variation is given as $\bar{\mathcal{V}} := \sum_{t=2}^{T}||\mathcal{D}_t - \mathcal{D}_{t-1}||_{\text{TV}}$, where $||\cdot||_{\text{TV}}$ denotes the total variation of the distribution and T is the total time horizon (Section 2).

In this paper, we provide an elimination-based algorithm that can achieve the *non-stationary regret* of $\tilde{\mathcal{O}}(\min\{\sqrt{\mathcal{S}T}, \bar{\mathcal{V}}^{\frac{1}{3}}T^{\frac{2}{3}}\})$ (Section 3). This regret bound shows that if the switchings or the total variations are not large (sublinear to T in particular), our algorithm can still achieve sublinear *non-stationary regret*. We give a proof sketch in Section 4 to show the main technical ideas of the regret analysis. We further show the non-stationary regret is lower bounded by $\Omega(\sqrt{\mathcal{S}T})$ in the switching case, and lower bounded by $\Omega(\bar{\mathcal{V}}^{\frac{1}{3}}T^{\frac{2}{3}})$ in the dynamic case (Section 5), which means that our Elim-NS algorithm achieves nearly optimal regret in the non-stationary environment. Moreover, our algorithm is parameter-free, which means that we do not need to know the parameters $\mathcal{S}$ and $\bar{\mathcal{V}}$ in advance and the algorithm is self-adaptive. Our main method is to reduce the non-stationary online auction problem into a variant of the non-stationary multi-armed bandit problem called *non-stationary one-sided full information bandit*, and solve this problem with some novel techniques.

Due to the space constraint, the detailed technical proofs are included in our full report (Zhao and Chen 2019a), but the proof sketch covering all essential ideas are included in the main text.

1.1 Related Work

Multi-armed bandit: Multi-armed bandit (MAB) problem is first introduced in (Robbins 1952). MAB problems can be classified into stochastic bandits and the adversarial bandits. In the stochastic case, the reward is drawn from an unknown distribution, and in the adversarial case, the reward is determined by an adversary. Our model is a generalization of the stochastic case, as discussed below. The classical MAB algorithms include UCB (Auer, Cesa-Bianchi, and Fischer 2002) and Thompson sampling (Thompson 1933) for the stochastic case and EXP3 (Auer et al. 2002) for the adversarial case. We refer to (Bubeck and Cesa-Bianchi 2012) for comprehensive coverage on the MAB problems.

Non-stationary MAB: Non-stationary MAB can be view as a generalization of the stochastic MAB, where the reward distributions are changing over time. The non-stationary MAB problems are analyzed mainly under two settings: The first considers the switching case, where there are $\mathcal{S}$ number of switchings in the distribution, and derives switching regret in terms of $\mathcal{S}$ and T (Garivier and Moulines 2011; Wei, Hong, and Lu 2016; Liu, Lee, and Shroff 2018); The second considers the dynamic case, where the distribution is changing continuously but the variation $\mathcal{V}$ is bounded, and present dynamic regret in terms of $\mathcal{V}$ and T(Gur, Zeevi, and Besbes 2014; Besbes, Gur, and Zeevi 2015). However, most of the studies need to use $\mathcal{S}$ or $\mathcal{V}$ as algorithm parameters, which may not be easy to obtain in practice. Designing parameter-free algorithms has been studied in the full-information case (Luo and Schapire 2015; Jun et al. 2017; Zhang et al. 2018). There are also several attempts to design parameter-free algorithms in the bandit case (Karnin and Anava 2016; Luo et al. 2018; Cheung, Simchi-Levi, and Zhu 2019), but the regret bound is not optimal. A recent and innovative study (Auer, Gajane, and Ortner 2019) solves the problem in the bandit case and achieves optimal regret. Then, (Chen et al. 2019) significantly generalizes the previous work by extending it into the non-stationary contextual bandit and also achieves optimal regret. Our study is the first one on the non-stationary one-sided full information bandit and its application to the online auction setting.

Online auction: For the online case where the private value distribution is unknown, (Cesa-Bianchi, Gentile, and Mansour 2015; Cesa-Bianchi et al. 2017; Zhao and Chen 2019b) consider different forms of the online second price auction. These studies assume that bidders truthfully follow their private value distributions, the same as we assume in this work. (Mohri and Medina 2015) further considers the online second price auction with strategic bidders, which means that their bidding may not be truthful. (Roughgarden and Wang 2016) studies the online second price auction with bidder specific reserve price. However, they need to use all the bidding information, and they also assume that the bidders are truthful. For the offline case where the private value distribution is known, the classical work by Myerson (Myerson 1981) provides an optimal auction algorithm when the private value distributions of all bidders are independent and known, and the seller could set different reserve prices for different bidders.

2 Preliminary and Model

In this section, we introduce the non-stationary online second price auction with semi-bandit feedback. We will also introduce the non-stationary regret to measure the performance of the algorithm. As mentioned before, we reduce the non-stationary online second price auction problem to a non-stationary bandit problem, which we called non-stationary one-sided full information bandit. We will also give the formal definition of the bandit problem and show the performance measurement for the corresponding bandit problem.

Definition 1 (Non-stationary Online Second Price Auction)**.** *There are a fixed number of n bidders and a seller, and the seller sells the same item in each round $t \in [T]$. In each*

round t, the seller sells the item through second price auction with reserve price $r^{(t)}$, where $r^{(t)}$ is chosen by the seller at the beginning of each round t and is announced to the bidders before the bidders give their private values. The values of the bidders follow a distribution $\mathcal{D}_t$ with support $[0,1]^n$ in round t, and the environment draws a vector of realized values for the bidders $\boldsymbol{v}^{(t)} \sim \mathcal{D}_t$. For each bidder $i \in [n]$, if her value $\boldsymbol{v}_i^{(t)} \geq r^{(t)}$, she will report her value $\boldsymbol{v}_i^{(t)}$ to the seller, otherwise she will not report her value and not attend the auction in this round.[1] *The seller then dispatches the item using the second price auction with reserve price $r^{(t)}$. We assume that the distributions $\mathcal{D}_t$ are generated **obliviously**, i.e., $\mathcal{D}_t$ are generated before our algorithm starts, or equivalently, $\mathcal{D}_t$ are generated independently to the randomness of $\mathcal{D}_s$ for all $s \leq t$ and the randomness of the algorithm.*

The performance of the reserve price in auction is always measured by the revenue: $\mathcal{R}(r^{(t)}, \mathcal{D}_t) := \mathbb{E}_{\boldsymbol{v}\sim\mathcal{D}_t}\left[\sum_{i=1}^n p_i(r^{(t)}, \boldsymbol{v})\right]$, where $p_i(r^{(t)}, \boldsymbol{v})$ denote the money bidder i needs to pay when the reserve price is $r^{(t)}$ and $\boldsymbol{v}$ is the private value vector of the bidders is $\boldsymbol{v}$. In particular, if bidder i has the highest bid among all bidders and its bid is also larger than the reserve price $r^{(t)}$, then i pays the maximum value among all other bids and the reserve price and gets the auction item; otherwise the bidder i pays nothing and does not get the item. Note that if we fix a reserve price r, whether bidders with values less than r report their values or not does not affect the revenue. Given the revenue of a reserve price, we have the following definition for the non-stationary regret in the online second price auction.

Definition 2 (Non-stationary Regret for Online Second Price Auction)**.** *The non-stationary regret of algorithm $\mathcal{A}$ for the online second price auction is defined as follow,*

$$Reg_{\mathcal{A}}^{SP} := \mathbb{E}\left[\sum_{t=1}^{T}(\mathcal{R}(r_t^*, \mathcal{D}_t) - \mathcal{R}(r^{(t)}, \mathcal{D}_t))\right],$$

where $r_t^* := \operatorname{argmax}_r \mathcal{R}(r, \mathcal{D}_t)$ and $r^{(t)}$ is the reserve price algorithm $\mathcal{A}$ chooses in round t, and the expectation $\mathbb{E}[\cdot]$ is taken over all the randomness, including the randomness of the algorithm itself and the randomness of $\boldsymbol{v}^{(1)}, \ldots, \boldsymbol{v}^{(t-1)}$ leading to the randomness in the selection of $r^{(t)}$.

We now introduce the measurement of the non-stationarity. In general, there are two measurements of the change of the environment: the first is the number of the switchings $\mathcal{S}$, and the second is the total variation $\bar{\mathcal{V}}$. For any interval $\mathcal{I} = [s, s']$, we define the number of switchings on $\mathcal{I}$ to be $\mathcal{S}_{\mathcal{I}} := 1 + \sum_{t=s+1}^{s'} \mathbb{I}\{\mathcal{D}_t \neq \mathcal{D}_{t-1}\}$. As for the total variation, the formal defintion is given as $\bar{\mathcal{V}}_{\mathcal{I}} := \sum_{t=s+1}^{s'} ||\mathcal{D}_t - \mathcal{D}_{t-1}||_{\text{TV}}$, where $||\cdot||_{\text{TV}}$ denotes the total variation of the distribution. For convenience, we use $\mathcal{S}$ and $\bar{\mathcal{V}}$ to denote $\mathcal{S}_{[1,T]}$ and $\bar{\mathcal{V}}_{[1,T]}$.

[1]We fully understand that in the repeated online second price auction, the bidder may not be truthful since she may participate in the auction in several rounds. However, this is out of the scope of the current paper. We will assume that the bidders are truthful in each round, and it is a good approximation in some cases.

Next, we briefly discuss how to reduce the online second price auction to the one-sided full-information bandit: 1) We can discretize the reserve price into $r_1, \ldots, r_K$. Because the revenue of the second price auction is one-sided Lipschitz, when K is large enough, the revenue of the best discretized reserve price should not make so much difference to that of the best reserve price on the whole domain. 2) The distribution of the value $\mathcal{D}_t$ will induce a distribution of reward on $(r_1, \ldots, r_K)$. More specifically, any private value vector $\boldsymbol{v}^{(t)} \sim \mathcal{D}_t$ will induce a reward vector $X^{(t)} = (X_1^{(t)}, \ldots, X_K^{(t)})$ for the discretized reserve price $r_1, \ldots, r_K$, and the reward vector $X^{(t)}$ follows a distribution ν_t. 3) At time t, because all bidders with values at least $r^{(t)}$ will report their values, we can compute the rewards for all $r \geq r^{(t)}$ given the specific private values larger than or equal to $r^{(t)}$. This gives us the following definition of the non-stationary one-sided full-information bandit. The formal reduction from the online auction to the bandit problem will be given in the proof of the Theorem 3.

Definition 3 (Non-stationary One-sided Full Information Bandit)**.** *There is a set of arms $\{1, 2, \ldots, K\}$, and for each arm $a \in [K]$ at time t, it corresponds to an unknown distribution $\nu_{a,t}$ with support $[0,1]$, where $\nu_{i,t}$ is the marginal distribution of ν_t with support $[0,1]^K$. In each round t, the environment draws a reward vector $X^{(t)} = (X_1^{(t)}, \ldots, X_K^{(t)})$, where $X^{(t)}$ is drawn from distribution ν_t. The player then chooses an arm A_t to play, gains the reward $X_{A_t}^{(t)}$ and observes the reward of arms $A_t, A_t + 1, \ldots, K$, i.e. observes $X_i^{(t)}, \forall i \geq A_t$. We assume that the distribution ν_t at each round t is generated **obliviously**, i.e. ν_t are generated before the algorithm starts.*

We use $\mu_{a,t}$ to denote the mean of $X_a^{(t)}$, i.e. $\mu_{a,t} = \mathbb{E}[X_a^{(t)}]$. We also use $\mu_t^* = \max_a \mu_{a,t}$ to denote the mean of the best arm at time t. Then we have the following definition of the non-stationary regret.

Definition 4 (Non-stationary Regret)**.** *We use the following to denote the non-stationary regret of algorithm $\mathcal{A}$.*

$$Reg_{\mathcal{A}} := \mathbb{E}\left[\sum_{t=1}^{T}(\mu_t^* - \mu_{A_t,t})\right].$$

For convenience, we will simply use regret to denote the non-stationary regret. We now introduce the measurements for the non-stationarity for the one-sided bandit case. Similar to the auction case, we have switchings $\mathcal{S}$ and variation $\mathcal{V}$. For any interval $\mathcal{I} = [s, s']$, we define the number of switchings on $\mathcal{I}$ to be $\mathcal{S}_{\mathcal{I}} := 1 + \sum_{t=s+1}^{s'} \mathbb{I}\{\nu_t \neq \nu_{t-1}\}$. As for the sum of variation, the formal definition is given as $\mathcal{V}_{\mathcal{I}} := \sum_{t=s+1}^{s'} \max_a |\mu_{a,t} - \mu_{a,t-1}|$, which sums up the max difference of mean in each round. For convenience, we use $\mathcal{S}$ and $\mathcal{V}$ to denote $\mathcal{S}_{[1,T]}$ and $\mathcal{V}_{[1,T]}$. Note that the number of switchings in the bandit case is the same as that of the auction case, so we reuse the notations, and the variation definition in the bandit case uses the sum of the maximal differences in the consecutive mean vectors instead of the sum

of total variations in the auction case, so we use the notation $\mathcal{V}$ instead of $\bar{\mathcal{V}}$ for differentiation. The variation $\mathcal{V}$ defined for the bandit case is consistent with the variation defined in other non-stationary bandit papers.

We will use Switching Regret to denote the non-stationary regret in the switching case, and Dynamic Regret to denote the non-stationary regret in the dynamic case.

3 Algorithm

In this section, we present our algorithm Elim-NS for the non-stationary one-sided full-information bandit problem and its regret bounds. The algorithm for the online auction problem can be easily derived from Elim-NS, as outlined in Section 2, and we present its regret bound in Theorem 3.

Our algorithm Elim-NS borrows ideas from (Zhao and Chen 2019b) and (Auer, Gajane, and Ortner 2019). (Zhao and Chen 2019b) introduce an elimination-based algorithm for the one-sided full-information bandit, and (Auer, Gajane, and Ortner 2019) presents an elimination-based algorithm to adaptively follow the best arm in the switching case without knowing the number of switches $\mathcal{S}$. Our algorithm is a non-trivial combination of these ideas, and our innovation highly depends on the feedback structure of the one-sided bandit problem. The algorithm is given in Algorithm 1.

Generally speaking, our algorithm maintains a set $\mathcal{E}$ to record the exploration phases for the adaptive detection of the dynamic changes in the distribution, and a set $\mathcal{M}$ to record the information when an arm is eliminated. If we were dealing with the stationary case where the distribution of arms does not change, after observing arms for enough times, we can eliminate an empirically sub-optimal arm, and with high probability, the eliminated arm is indeed sub-optimal. However, in the non-stationary case, the optimal arm is changing, and thus we need to properly add exploration phases to observe the eliminated arms with some probability. When we detect that the distribution indeed has changed from these exploration phases, the algorithm starts a new epoch and resets $\mathcal{E}$ and $\mathcal{M}$ to empty sets. [2]

Set $\mathcal{M}$ records the information at the time when an arm is eliminated. Each element $(g, e, \boldsymbol{v}) \in \mathcal{M}$ is a tuple, where $g \in \mathbb{R}$ records the empirical gap, which is the difference of the empirical means of the empirically optimal arm and that of the eliminated arm $a_{\min}$; $e = a_{\min}$ records the index of the eliminated arm; and $\boldsymbol{v}_k$ for $k \geq a_{\min}$ records the empirical mean of arm k when the arm e is eliminated ($\boldsymbol{v} \in \mathbb{R}^K$). An exploration phase is a pair $(d, \mathcal{I})$ where $d = 2^{-k}$ and interval $\mathcal{I} \subseteq [T], |\mathcal{I}| = \Theta(\frac{1}{d^2})$. Each such phase is stored independently into $\mathcal{E}$ with a probability (in line 8 of Step 1). The purpose of these exploration phases is to re-examine arms that have been eliminated to detect possible changes in the distribution, with $\mathcal{I}$ indicating the range of rounds for an exploration. Intuitively, if there is no change in the distribution, such an exploration would pay an extra regret. To control this extra regret, we use d to indicate the per-round regret that such an exploration could tolerate, and the length of $\mathcal{I}$ is controlled to be $\tilde{\mathcal{O}}(1/d^2)$ to bound the total regret.

[2] We mark the actual values of $\mathcal{E}$ and $\mathcal{M}$ in each round as $\mathcal{E}_t$ and $\mathcal{M}_t$ in the algorithm, to be used in our analysis.

At each round, Our algorithm Elim-NS has the following four steps. In Step 1, we randomly add exploration phases into the set $\mathcal{E}$. We set $p_{\ell,i} = d_i\sqrt{(\ell+1)/T}$ to be the probability to add an exploration phase $(d_i, [t, t + \lceil C_2 \ln(KT^3)/d_i^2 \rceil))$ into $\mathcal{E}$ in epoch ℓ at time t. This probability is chosen carefully, not too small to omit the non-stationarity, and not too big to induce large regret.

In Step 2, we choose the action to play. If the current round t is not in any exploration phase, then we will play the arm that is not eliminated and has the smallest index. If t is in an exploration phase $(d, \mathcal{I})$, we will find the maximum value $d_{\max,t} = \max_{(d,\mathcal{I})\in\mathcal{E}, t\in\mathcal{I}} d$. We will play arm $A_t \leftarrow a_{\text{exp}} = \min\{k : \exists(g, e, \boldsymbol{v}) \in \mathcal{M}, k = e, g \leq 8d_{\max,t}\}$ and observe the reward $X_a^{(t)}$ for all $a \geq a_{\text{exp}}$. This arm selection in the exploration phase guarantees that the arm we play would induce the regret of at most $\mathcal{O}(d_{\max,t})$ per round if the distribution has not changed.

In Step 3, we perform arm elimination when the proper condition holds. In particular, when we find an arm is empirically sub-optimal among the remaining arms, we eliminate this arm in this epoch. When an arm is eliminated, the algorithm will add an tuple $(g, e, \boldsymbol{v})$ into the set $\mathcal{M}$ to store the information at this point, where g stores the empirical gap with the best arm, e stores the index of the eliminated arm, and for $k \geq e$ $\boldsymbol{v}_k$ stores the empirical mean of arm k.

In Step 4, we apply the non-stationarity check. At the end of an exploration phase, we check that if there is a tuple $(g, e, \boldsymbol{v}) \in \mathcal{M}$ and an arm $a \geq e$, such that the gap between the current empirical mean of arm a during the exploration phase and the stored empirical mean $\boldsymbol{v}_a$ is $\Omega(g)$. If so, it means that the empirical mean has a significant deviation indicating a change in distribution, and thus we will start a new epoch to redo the entire process from scratch.

The algorithm incorporates ideas from (Auer, Gajane, and Ortner 2019; Zhao and Chen 2019b), and its main novelty is related to the maintenance and use of set $\mathcal{M}$ in arm selection (Step 2), arm elimination (Step 3) and stationarity check (Step 4), which make use of the feedback observation to balance exploration and exploitation.

Now, we use a simple example to illustrate how the Elim-NS algorithm detects the distribution changes in the switching case. Suppose that we have three arms. At first, arm 1 always outputs 0, arm 2 always outputs 0.45, and arm 3 always outputs 0.5. Then arm 1 will be eliminated first, and the tuple $(g, e, \boldsymbol{v}) = (0.5, 1, (0, 0.45, 0.5))$ will be stored in $\mathcal{M}$, where $g = 0.5$ is the empirical gap between the means of arm 1 and the empirically best arm 3. Next, arm 2 will be eliminated, and the algorithm will store $(0.05, 2, (?, 0.45, 0.5))$ in $\mathcal{M}$, where ? means that the value at that position has no meaning. At this point, the algorithm may have randomly selected many exploration phases, but they all fail to start a new epoch since the distribution does not change and non-stationarity would not be detected. Then suppose that at round t, the distribution changes, and arm 1 will output 1 from now on and thus becomes the best arm. Suppose that after round t, we randomly select an exploration phase with $d = 2^{-5}$, and in this exploration phase, we will play arm 2 but not arm 1 (since $0.05 \leq 8 * 2^{-5} < 0.5$),

Algorithm 1: Elim-NS

Data: Total time horizon T, total number of arms K. Parameters C_1, C_2.

$t \leftarrow 1, \ell \leftarrow 1, \tau_\ell \leftarrow t$.

$\mathcal{M} \leftarrow \phi, a_{\min} \leftarrow 1, \mathcal{E} \leftarrow \phi$.

Let $\hat{\mu}_a[t_1, t_2)$ denote the empirical mean of arm a in the time interval $[t_1, t_2)$.

while $t \leq T$ **do**

Step 1. Randomly select the exploration phases

if $\mathcal{M} \neq \phi$ **then** $\Delta_{t,\min} \leftarrow \min_{(g,e,\boldsymbol{v})\in\mathcal{M}} g$.

Let $d_i \leftarrow 2^{-i}$ for every $i \in \mathbb{N}$, and $I_t \leftarrow \max\{i : 8d_i \geq \Delta_{t,\min}\}$.

For every $i \leq I_t$, independently add pair $\left(d_i, \left[t, t + \lceil C_2 \ln(KT^3)/d_i^2 \rceil\right)\right)$ into $\mathcal{E}$ with probability $p_{\ell,i} = d_i\sqrt{\frac{\ell+1}{T}}$.

(Let $\mathcal{E}_t$ and $\mathcal{M}_t$ be the values of $\mathcal{E}$ and $\mathcal{M}$ respectively at this point, to be used in the proof)

Step 2. Choose an action to play

if $\exists(d, \mathcal{I}) \in \mathcal{E}$ *such that* $t \in \mathcal{I}$ **then**

$d_{\max,t} \leftarrow \max_{(d,\mathcal{I})\in\mathcal{E}, t\in\mathcal{I}} d$.

Play arm $A_t \leftarrow a_{\text{exp}} = \min\{k : \exists(g, e, \boldsymbol{v}) \in \mathcal{M}, k = e, g \leq 8d_{\max,t}\}$ and observe the reward $X_a^{(t)}$ for all $a \geq a_{\text{exp}}$.

else

Play arm $A_t \leftarrow a_{\min}$ and observe the reward $X_a^{(t)}$ for all $a \geq a_{\min}$

Step 3. Perform the elimination process

while $\exists\sigma \geq \tau_\ell, a > a_{min}$ *such that* $\hat{\mu}_a[\sigma, t+1) - \hat{\mu}_{a_{min}}[\sigma, t+1) > \sqrt{\frac{C_1 \ln(KT^3)}{t+1-\sigma}}$ **do**

Let $\boldsymbol{v}$ be a vector with length K.

Let b be the arm such that $\hat{\mu}_b[\sigma, t+1) - \hat{\mu}_{a_{\min}}[\sigma, t+1)$ is maximized.

$g \leftarrow \hat{\mu}_b[\sigma, t+1) - \hat{\mu}_{a_{\min}}[\sigma, t+1)$, $e \leftarrow a_{\min}$, and $\boldsymbol{v}_i \leftarrow \hat{\mu}_i[\sigma, t+1)$ for all $i \geq a_{\min}$.

$\mathcal{M} \leftarrow \mathcal{M} \cup \{(g, e, \boldsymbol{v})\}$, $a_{\min} \leftarrow a_{\min} + 1$.

Step 4. Perform the non-stationarity check

if $\exists(d, [t', t+1)) \in \mathcal{E}, (g, e, \boldsymbol{v}) \in \mathcal{M}, a \geq e$ *such that* $g \leq 8d$ *and* $|\hat{\mu}_a[t', t+1) - \boldsymbol{v}_a| > \frac{d}{4}$ **then**

$\ell \leftarrow \ell + 1, \mathcal{M} \leftarrow \phi, \mathcal{E} \leftarrow \phi, a_{\min} \leftarrow 1, \tau_\ell \leftarrow t + 1$.

$t \leftarrow t + 1$.

and thus we will still not detect the non-stationarity of arm 1. However, when we randomly select an exploration phase with $d = 0.5$ in step 1 (perhaps in a later round), we will play arm 1 according to the key selection criteria for arm exploration in line 13 of step 2. This would allow us to observe the distribution change on arm 1 in the exploration phase and then start a new epoch, which will restart the algorithm from scratch by playing arm 1 again.

The following two theorems summarize the regret bounds of algorithm Elim-NS in the switching case and the variation case for the one-sided full-information bandit.

Theorem 1 (Switching Regret)**.** *Suppose that we choose parameters $C_1 \geq 2048, C_2 \geq 32$, then the algorithm* Elim-NS *has regret in the switching case bounded by $\tilde{\mathcal{O}}(\sqrt{\mathcal{S}T})$, where $\tilde{\mathcal{O}}(\cdot)$ hides the polynomial factor of $\log K$ and $\log T$.*

Theorem 2 (Dynamic Regret)**.** *Suppose that we $C_1 \geq 8192, C_2 \geq 128$, and suppose that the variation is not too small ($\mathcal{V} = \Omega(1)$). Then the algorithm* Elim-NS *has regret in the dyanmic case bounded by $\tilde{\mathcal{O}}(\mathcal{V}^{\frac{1}{3}}T^{\frac{2}{3}})$, where $\tilde{\mathcal{O}}(\cdot)$ hide the polynomial factor of $\log K$ and $\log T$.*

As outlined in Section 2, Elim-NS can be easily adapted to solve the online second price auction problem by discretizing the reserve price. The following theorem provides the regret bound of Elim-NS on solving the online second price auction problem.

Theorem 3 (Regret for Online Second Price Auction)**.** *For every $0 \leq k \leq \lceil\sqrt{T}\rceil$, let $r_k = k/\lceil\sqrt{T}\rceil$, and we only set reserve price $r^{(t)} \in \{r_1, \ldots, r_{\lceil\sqrt{T}\rceil}\}$. Each time we set reserve price $r^{(t)} = r_{A_t}$ and get all the private value $v_i^{(t)} \geq r^{(t)}$, we compute the reward $X_k^{(t)}$ for all $k \geq A_t$ and receive the reward $X_{A_t}^{(t)}$. Then we apply our algorithm* Elim-NS *and set C_1, C_2 appropriately, and the regret is bounded by*

$$Reg_{\mathcal{A}}^{SP} \leq \tilde{\mathcal{O}}(\min\{\sqrt{\mathcal{S}T}, \bar{\mathcal{V}}^{\frac{1}{3}}T^{\frac{2}{3}}\}),$$

where we assume that $\bar{\mathcal{V}} = \Omega(1)$ is not too small.

4 Proof Sketch for the Regret Analysis

In this section, we will give a proof sketch of the regret analysis in the switching case (Theorem 1) and the dynamic case (Theorem 2). In general, we first give a proof in the

switching case, and then we reduce the dynamic case into the switching case. The proof strategy in the dynamic case is nearly the same as that in the switching case, and we will briefly discuss how to do the reduction.

4.1 Proof Sketch of Theorem 1

Generally speaking, our proof strategy for Theorem 1 is to define several events (Definitions 5,6,7,8), and decompose the regret by these events. We show that each term in the decomposition is bounded by $\tilde{\mathcal{O}}(\sqrt{\mathcal{S}T})$.

Definition 5 (Sampling is nice)**.** *We say that the sampling is nice if for every interval $\mathcal{I} \subseteq [T]$ and every arm a, we have*

$$\frac{1}{|\mathcal{I}|}\left|\sum_{t\in\mathcal{I}} X_a^{(t)} - \sum_{t\in\mathcal{I}} \mu_{a,t}\right| < \sqrt{\frac{\ln(KT^3)}{2|\mathcal{I}|}},$$

where $|\mathcal{I}|$ is the length of interval $\mathcal{I}$. We use $\mathcal{N}^s$ to denote this event. We use $\mathcal{N}_t^s$ to denote the event when the above inequality holds for all $\mathcal{I} \subseteq [t]$.

Definition 6. *We use $\mathcal{P}_t$ to denote the event such that t is in an exploration phase, i.e. $\exists(d, \mathcal{I}) \in \mathcal{E}_t$ such that $t \in \mathcal{I}$.*

Definition 7 (Records are consistent)**.** *We say that the records are consistent at time t if for every $(g, e, \boldsymbol{v}) \in \mathcal{M}_t$, for every arm $a \geq e$, we have $|\mu_{a,t} - \boldsymbol{v}_a| \leq g/4$. We use $\mathcal{C}_t$ to denote this event.*

We have the following definition when $\mathcal{C}_t$ does not happen.

Definition 8 (Playing bad arm)**.** *Let b_t denote the smallest index of an arm such that $\exists(g, e, \boldsymbol{v}) \in \mathcal{M}_t$, $e = b_t$ and there exists $a \geq e, |\boldsymbol{v}_a - \mu_{a,t}| > g/4$, i.e.*

$$b_t = \min\left\{e : (g, e, \boldsymbol{v}) \in \mathcal{M}_t, \exists a \geq e, |\boldsymbol{v}_a - \mu_{a,t}| > \frac{g}{4}\right\}.$$

We use $\mathcal{B}_t$ to denote the event $\{A_t \geq b_t\}$.

Generally speaking, b_t is the smallest index of an eliminated arm such that the recorded mean when b_t is eliminated induces the event $\neg\mathcal{C}_t$.

Based on the above definitions, we decompose the regret into four mutually exclusive events and bound the regret for each event in the order of $\tilde{\mathcal{O}}(\sqrt{\mathcal{S}T})$. These four event cases are listed below, where the first three are when the sampling is nice, and the last case is when sampling is not nice.

Case 1: $\mathcal{N}^s \wedge \mathcal{C}_t \wedge \neg\mathcal{P}_t$. This means that the sampling is nice, the records are consistent at time t, and round t is not in an exploration phase. The regret should be bounded in this case since when $\mathcal{C}_t$ happens, the distribution does not change much, and it is also not in an exploration phase (Lemma 1).

Case 2: $\mathcal{N}^s \wedge \mathcal{C}_t \wedge \mathcal{P}_t$ or $\mathcal{N}^s \wedge \neg\mathcal{C}_t \wedge \neg\mathcal{B}_t$. The sampling is still nice. When $\mathcal{C}_t \wedge \mathcal{P}_t$ is true, round t is in an exploration phase and the records are consistent, meaning that the current arm means have not deviated much from the records. In this case, similar as discussed before, the definition of the exploration phase $(d, \mathcal{I})$ and the setting in line 13 guarantee that the arm explored would not have a large regret. When $\neg\mathcal{C}_t \wedge \neg\mathcal{B}_t$ is true, we first claim that $\neg\mathcal{C}_t \wedge \neg\mathcal{B}_t$ implies $\mathcal{P}_t$. This is because if the records are not consistent (i.e. $\neg\mathcal{C}_t$) but $A_t < b_t$ (i.e. $\neg\mathcal{B}_t$), it means A_t played in round t has smaller index than b_t, but b_t is an eliminated arm according to Definition 8, and thus arm A_t must be played due to exploration. Next, since $A_t < b_t$, the arm played is not a bad arm with a large gap, so its regret is still bounded (Lemma 2).

Case 3: $\mathcal{N}^s \wedge \neg\mathcal{C}_t \wedge \mathcal{B}_t$. The sampling is nice, the records are not consistent, and in round t, we play a bad arm with a large gap between the current mean and the recorded mean. Although the regret in this case cannot be bounded by $\mathcal{O}(g)$ where $(g, A_t, \boldsymbol{v}) \in \mathcal{M}_t$, the key observation is that, due to the random selection of the exploration phase, we will observe the non-stationarity (since $\mathcal{C}_t$ does not happen and $\mathcal{B}_t$ happens) with some probability, and the expected regret can be bounded (Lemma 3).

Case 4: $\neg\mathcal{N}^s$. The sampling is not nice, which is a low probability event, and its regret can be easily bounded by a constant (Lemma 4).

Lemma 1.

$$\begin{aligned}&\mathbb{E}\left[\sum_{t=1}^{T}(\mu_t^* - \mu_{A_t,t})\cdot\mathbb{I}\left\{\mathcal{N}^s \wedge \mathcal{C}_t \wedge \neg\mathcal{P}_t\right\}\right]\\ \leq& 2\mathcal{S} + 2(\sqrt{C_1}+\sqrt{2})\sqrt{\ln(KT^3)}\sqrt{2\mathcal{S}T}.\end{aligned}$$

The proof of Lemma 1 is similar to the analysis in (Zhao and Chen 2019b) and can be viewed as a generalization of the original proof. The key difference is that in the proof of Lemma 1, we divide the interval into

$$[1, T] = [s_1, e_1] \cup [s_2, e_2] \cup \cdots \cup [s_{\mathcal{S}}, e_{\mathcal{S}}],$$

and we sum the regret in each interval first, and get the regret in each interval to be $\tilde{\mathcal{O}}(\sqrt{e_i - s_i + 1})$. Then we sum them up and show that the regret is in the order of $\tilde{\mathcal{O}}(\sqrt{\mathcal{S}T})$.

Lemma 2.

$$\begin{aligned}&\mathbb{E}\left[\sum_{t=1}^{T}(\mu_t^* - \mu_{A_t,t})\cdot\mathbb{I}\left\{\mathcal{N}^s \wedge \mathcal{C}_t \wedge \mathcal{P}_t\right\}\right]\\ &+\mathbb{E}\left[\sum_{t=1}^{T}(\mu_t^* - \mu_{A_t,t})\cdot\mathbb{I}\left\{\mathcal{N}^s \wedge \neg\mathcal{C}_t \wedge \neg\mathcal{B}_t\right\}\right]\\ \leq&\left(C_2\ln(KT^3)\sqrt{(\mathcal{S}+1)T} + 2\sqrt{\frac{\mathcal{S}+1}{T}}\right)\\ &\times\left(3 - \log_2\sqrt{\frac{C_1\ln(KT^3)}{T}}\right).\end{aligned}$$

This lemma bounds the regret when $\mathcal{B}_t$ does not happen and t is in an exploration phase. In this case, we show that the number of different lengths d of exploration phases $(d, \mathcal{I})$ can be bounded by $polylog(K, T)$. Then, we show that the regret induced by the specific length exploration phase is bounded by $\tilde{\mathcal{O}}(\sqrt{\mathcal{S}T})$. Finally, we combine the previous argument and apply the union bound to show that the total regret considered is bounded by $\tilde{\mathcal{O}}(\sqrt{\mathcal{S}T})$.

Lemma 3.

$$\mathbb{E}\left[\sum_{t=1}^{T}(\mu_t^* - \mu_{A_t,t})\cdot\mathbb{I}\left\{\mathcal{N}^s \wedge \neg\mathcal{C}_t \wedge \mathcal{B}_t\right\}\right]$$

$$\leq 24\sqrt{(\mathcal{S}+1)T} + 24\sqrt{C_2 \ln(KT^3)\mathcal{S}T}.$$

This lemma bound the regret when $\mathcal{B}_t$ happens, and this lemma is the most technical one. The proof strategy is similar to (Auer, Gajane, and Ortner 2019), which partitions the entire time horizon into several intervals with identical distribution, and applies a two-dimensional induction from back to front. As discussed before, the regret in this case in each round cannot be bounded by $\mathcal{O}(g)$ where $(g, A_t, \boldsymbol{v}) \in \mathcal{M}_t$. However due to the random selection of the exploration phases, with some probability, we will observe the non-stationarity (since $\mathcal{C}_t$ does not happen and $\mathcal{B}_t$ happens), and the expected regret can be bounded.

Finally, by a simple application of the high probability result on $\mathcal{N}^s$, we can get the following lemma.

Lemma 4. $\mathbb{E}\left[\sum_{t=1}^{T}\left(\mu_t^* - \mu_{A_t,t}\right)\cdot \mathbb{I}\left\{\neg\mathcal{N}^s\right\}\right] \leq 2.$

Combining these lemmas together, we complete the proof of Theorem 1.

4.2 Proof Sketch of Theorem 2

In this part, we briefly introduce how to reduce the dynamic case to the switching case. The proof is an imitation of the proof strategy of Theorem 1. Although the means can be changing at every time $t \in [1, T]$, we can approximately divide them into several sub-intervals such that in each interval, the change of means is not large. Recall that for interval $\mathcal{I} = [s, s']$, $\mathcal{V}_{\mathcal{I}} := \sum_{t=s+1}^{s'} \max_a |\mu_{a,t} - \mu_{a,t-1}|$ and we use $\mathcal{V} := \mathcal{V}_{[1,T]}$. We have the following lemma,

Lemma 5 (Interval Partition (Chen et al. 2019))**.** *There is a way to partition the interval $[1, T]$ into $\mathcal{I}_1 \cup \mathcal{I}_2 \cup \cdots \cup \mathcal{I}_\Gamma$ such that $\mathcal{I}_i \cap \mathcal{I}_j = \phi$, and for any $i \leq \Gamma$, $\mathcal{V}_{\mathcal{I}} \leq \sqrt{C_3/|\mathcal{I}_i|}$ and $\Gamma \leq (2T/C_3)^{1/3}\mathcal{V}^{\frac{2}{3}} + 1$.*

Suppose that we have a partition shown in the above lemma. We construct a new instance such that $\mu'_{a,t} = \frac{1}{|\mathcal{I}_j|}\sum_{s\in\mathcal{I}_j}\mu_{a,s}$ for all $j \leq \Gamma$ and all $t \in \mathcal{I}_j$, i.e. we take the average mean of each interval and make them all the same.

Generally speaking, the dynamic regret can be bounded by the sum of 2 parts: the switching regret of the new instance and the difference between the switching regret of the new instance and the dynamic regret. As for the first part, since $\Gamma \leq (2T/C_3)^{1/3}\mathcal{V}^{\frac{2}{3}} + 1$, we know that the switching regret can be bounded by $\tilde{\mathcal{O}}(\sqrt{\Gamma T}) = \tilde{\mathcal{O}}(\mathcal{V}^{\frac{1}{3}}T^{\frac{2}{3}})$. As for the difference between the 2 regret, since $|\mu_{a,t} - \mu'_{a,t}| \leq \mathcal{V}_{\mathcal{I}_j}$ for $t \in \mathcal{I}_j$, we sum up all t, we know that the difference is bounded by $\mathcal{O}(\sum_j \sqrt{|\mathcal{I}_j|}) = \mathcal{O}(\sqrt{\Gamma T}) = \mathcal{O}(\mathcal{V}^{\frac{1}{3}}T^{\frac{2}{3}})$. Combine them together we complete the proof.

4.3 Proof Sketch of Theorem 3

In the proof of Theorem 3, we first show that the online second price auction has one-sided Lipschitz property, and thus discretizing the reserve price will not lead to a large regret. Next, we briefly discuss why discretizing the reserve price can lead to a one-sided full information bandit instance, and then it is easy to show that the regret can be bounded by $\tilde{\mathcal{O}}(\sqrt{\mathcal{S}T})$ in the switching case. To bound the regret in the dynamic case, we only have to set up the connection between the total variation $\bar{\mathcal{V}}$ in the online auction and the variation $\mathcal{V}$ in the bandit problem. The bridge between these two variables can be set up easily by the definition and property of total variation $||\cdot||_{\text{TV}}$.

5 Lower Bounds for Online Second Price Auction in Non-stationary Environment

In this section, we show that for the online second price auction problem, the regret upper bounds achieved by Elim-NS is almost tight, by giving a regret lower bound of $\Omega(\sqrt{\mathcal{S}T})$ for the switching case, and a lower bound of $\Omega(\bar{\mathcal{V}}^{\frac{1}{3}}T^{\frac{2}{3}})$ for the dynamic case.

Theorem 4. *For any algorithm, and any $\mathcal{S} > 0$, there exists a set distributions of bids $\mathcal{D}_1, \ldots, \mathcal{D}_T$ where $\mathcal{S} = 1 + \sum_{t=1}^{T-1}\mathbb{I}\{\mathcal{D}_t \neq \mathcal{D}_{t+1}\}$ is the number of switchings of the distribution and the non-stationary regret is at least $\Omega(\sqrt{\mathcal{S}T})$. Moreover for any algorithm and any $\bar{\mathcal{V}} \geq 1$, there exists $\mathcal{D}_1, \ldots, \mathcal{D}_T$ where $\sum_{t=2}^{T}||\mathcal{D}_t - \mathcal{D}_{t-1}||_{TV} \leq \mathcal{O}(\bar{\mathcal{V}})$, such that the regret is at least $\Omega(\bar{\mathcal{V}}^{\frac{1}{3}}T^{\frac{2}{3}})$.*

Our theorem is based on the following result in (Cesa-Bianchi, Gentile, and Mansour 2015).

Proposition 1 (Theorem 2 of (Cesa-Bianchi, Gentile, and Mansour 2015))**.** *For any deterministic algorithm, there exists a distribution of bids operating with two bidders and the stationary regret is at least $\Omega(\sqrt{T})$.*

The above proposition shows that in the full-information case, any deterministic algorithm will have stationary regret lower bounded by $\Omega(\sqrt{T})$ for the online second price auction problem. Generally speaking, we divide the time interval into $\mathcal{S}$ segments, each with length $\frac{T}{\mathcal{S}}$. We construct an instance such that the regret in each segment is $\Omega(\sqrt{T/\mathcal{S}})$, and the total non-stationary regret sums up to be $\Omega(\sqrt{\mathcal{S}T})$.

As for the regret in the dynamic case, the proof is very similar. We also divide the time horizon into $\Theta(\bar{\mathcal{V}}^{\frac{2}{3}}T^{\frac{1}{3}})$ segments, and the total variation between the distribution of adjacent segments is bounded by $\left(\bar{\mathcal{V}}/T\right)^{\frac{1}{3}}$.

6 Conclusion and Further Work

We study the non-stationary online second price auction with the "semi-bandit" feedback structure in this paper. We reduce it into the non-stationary one-sided full-information bandit and show an algorithm Elim-NS that solves the problem. Our algorithm is parameter-free, which means that we do not have to know the switchings $\mathcal{S}$ and the variation $\mathcal{V}$ in advance. Our algorithm is also nearly optimal in both cases. There are also some future directions to explore:

First, in this work, we consider the online auction with "semi-bandit" feedback, where all the bidders with private values exceeding or equaling the reserve price will report their private values. We can also consider the "full-bandit" feedback where the seller only gets the reward in each round but does not observe the private values and design parameter-free algorithms to solve it in the non-stationary

case. Second, in this work we use the second price auction and assume that the bidders are truthful. We can also study how to generalize this non-stationary result into the strategic bidders' case or the other auction formats such as the generalized second price auction.

References

Auer, P.; Cesa-Bianchi, N.; Freund, Y.; and Schapire, R. E. 2002. The nonstochastic multiarmed bandit problem. *SIAM J. Comput.* 32(1):48–77.

Auer, P.; Cesa-Bianchi, N.; and Fischer, P. 2002. Finite-time analysis of the multiarmed bandit problem. *Machine Learning* 47(2-3):235–256.

Auer, P.; Gajane, P.; and Ortner, R. 2019. Adaptively tracking the best bandit arm with an unknown number of distribution changes. In *Conference on Learning Theory, COLT 2019, 25-28 June 2019, Phoenix, AZ, USA*, 138–158.

Besbes, O.; Gur, Y.; and Zeevi, A. J. 2015. Non-stationary stochastic optimization. *Operations Research* 63(5):1227–1244.

Bubeck, S., and Cesa-Bianchi, N. 2012. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning* 5(1):1–122.

Cesa-Bianchi, N.; Gaillard, P.; Gentile, C.; and Gerchinovitz, S. 2017. Algorithmic chaining and the role of partial feedback in online nonparametric learning. In *Proceedings of the 30th Conference on Learning Theory, COLT 2017, Amsterdam, The Netherlands, 7-10 July 2017*, 465–481.

Cesa-Bianchi, N.; Gentile, C.; and Mansour, Y. 2015. Regret minimization for reserve prices in second-price auctions. *IEEE Transactions on Information Theory* 61(1):549–564.

Chen, Y.; Lee, C.; Luo, H.; and Wei, C. 2019. A new algorithm for non-stationary contextual bandits: Efficient, optimal and parameter-free. In *Conference on Learning Theory, COLT 2019, 25-28 June 2019, Phoenix, AZ, USA*, 696–726.

Cheung, W. C.; Simchi-Levi, D.; and Zhu, R. 2019. Learning to optimize under non-stationarity. In Chaudhuri, K., and Sugiyama, M., eds., *Proceedings of Machine Learning Research*, volume 89, 1079–1087.

Garivier, A., and Moulines, E. 2011. On upper-confidence bound policies for switching bandit problems. In *Algorithmic Learning Theory - 22nd International Conference, ALT 2011, Espoo, Finland, October 5-7, 2011. Proceedings*, 174–188.

Gur, Y.; Zeevi, A. J.; and Besbes, O. 2014. Stochastic multi-armed-bandit problem with non-stationary rewards. In *Annual Conference on Neural Information Processing Systems*, 199–207.

Jun, K.; Orabona, F.; Wright, S.; and Willett, R. 2017. Online learning for changing environments using coin betting. *CoRR* abs/1711.02545.

Karnin, Z. S., and Anava, O. 2016. Multi-armed bandits: Competing with optimal sequences. In *Annual Conference on Neural Information Processing Systems*, 199–207.

Liu, F.; Lee, J.; and Shroff, N. B. 2018. A change-detection based framework for piecewise-stationary multi-armed bandit problem. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18)*, 3651–3658.

Luo, H., and Schapire, R. E. 2015. Achieving all with no parameters: Adanormalhedge. In *Proceedings of The 28th Conference on Learning Theory, COLT 2015, Paris, France, July 3-6, 2015*, 1286–1304.

Luo, H.; Wei, C.; Agarwal, A.; and Langford, J. 2018. Efficient contextual bandits in non-stationary worlds. In *Conference On Learning Theory, COLT 2018, Stockholm, Sweden, 6-9 July 2018.*, 1739–1776.

Mohri, M., and Medina, A. M. 2015. Revenue optimization against strategic buyers. In *Annual Conference on Neural Information Processing Systems*, 2530–2538.

Myerson, R. B. 1981. Optimal auction design. *Mathematics of Operations Research* 6(1).

Robbins, H. 1952. Some aspects of the sequential design of experiments. *Bull. Amer. Math. Soc.* 58(5):527–535.

Roughgarden, T., and Wang, J. R. 2016. Minimizing regret with multiple reserves. In *Proceedings of the 2016 ACM Conference on Economics and Computation, EC '16, Maastricht, The Netherlands, July 24-28, 2016*, 601–616.

Thompson, W. R. 1933. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika* 25(3/4):285–294.

Wei, C.; Hong, Y.; and Lu, C. 2016. Tracking the best expert in non-stationary stochastic environments. In *Annual Conference on Neural Information Processing Systems*, 3972–3980.

Zhang, L.; Yang, T.; Jin, R.; and Zhou, Z. 2018. Dynamic regret of strongly adaptive methods. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, 5877–5886.

Zhao, H., and Chen, W. 2019a. Online second price auction with semi-bandit feedback under the non-stationary setting. Technical report. arXiv preprint, 1911.05949.

Zhao, H., and Chen, W. 2019b. Stochastic one-sided full-information bandit. In *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD'2019)*.