We thank all the reviewers for their insightful reviews. We address the comments of each reviewer separately.

Reviewer #3: Regarding your comment 'the authors claim "the buyer does not play..." ': as we state in line 211, the goal of that paragraph was to give a high level idea of the construction. Indeed, while elaborating the details later on, we make precise this very point in line 233: there we say that the buyer's probability of playing the 0-th option (which we also call as the high option, as mentioned in line 212) is at most γ. Since $\gamma = o(1)$ for mean-based algorithms, we said in line 214 (the high level paragraph) "the buyer does not play"; indeed, at a high level, playing with probability $o(1)$ is equivalent to not playing at all, and in fact it does not affect the asymptotics of our regret guarantee whether the arm is played with probability $o(1)$ or 0 (as the difference would amount to at most $o(T)$ additive regret). We will ensure that we leave no room for ambiguity even in the high level ideas section in the next version.

The previous paragraph also answers your questions "And how does one guarantee that an arm is not played? In most MAB algorithms, aren't even historically unhelpful arms are played with some low probability?". Indeed, even historically unhelpful arms are played with some low probability, and in our case this low probability is $\gamma = o(1)$, as mentioned in line 233.

Regarding your comment "I was under the impression that the authors assumed the buyer was running a multi-arm bandit algorithm, which doesn't observe all rewards at each time step": yes, the buyer is indeed running a multi-armed bandit algorithm, and, the buyer can only observe the reward of the option that he chose, not of all the options.

Regarding your comment of "utility gain from session 0 is exactly cancelled out with his utility loss from session 2 (line 236)" and the related question of "How then can you guarantee that the observed utilities exactly sum up to some constant?": we believe the ambiguity gets cleared by focusing on the word "observed" in your comment. What we meant by "cancelled out" is that the cumulative utility of playing arm i (*not the cumulative observed utility*) is 0. Formally, $\sigma_{i,t}(1) = 0$ for any $t > \kappa_{i+1}$ where $\sigma_{i,t}(1)$ (formally defined in Definition 2 line 185) is the cumulative utility of playing option i for all of the first t rounds with a buyer value of 1. Clearly, in a multi-armed bandit setting the cumulative utility for playing an arm i (not cumulative observed utility), can be set by the adversary/seller in an arbitrary manner to whatever the adversary/seller wants, regardless of the buyer's strategy. The interesting question then is how can the cumulative utility (instead of cumulative observed utility) have any consequence on the probability with which the player/buyer plays an arm/option? This is where the definition of a mean-based strategy helps (Definition 2, line 185): it has strong consequences on the probability with which a player/buyer plays an arm/option based purely on cumulative utilities. While this may seem very restrictive at first glance, surprisingly several common/natural no-regret algorithms are mean-based: for instance the famous EXP3 algorithm for the multi-armed bandit setting is a mean-based algorithm. We will make sure to emphasize this and make it fully clear.

Given that these were Reviewer 3's main concerns, we sincerely hope Reviewer 3 would consider revising their score.

Reviewer #4: We thank the reviewer for the positive and encouraging review. It is a very interesting and challenging open question to extend both our results and Braverman et al.'s results to the multi-agent setting.

Reviewer #5:

1. Regarding the collection of related works that you mention: in all these works, the *seller is learning* to set prices for a strategic buyer. Whereas in our paper the *buyer is learning* a good bidding strategy using a no-regret learning algorithm. So our work is fundamentally different from all these related works both in subject and in techniques. We will add a short discussion to emphasize this crucial difference from the related works you mention.

2. Regarding the comparison to Heidari et al.: Heidari et al. [ICML16] studies a setting in which the buyers use no-regret algorithms for choosing ad exchanges. In our paper there is only one ad exchange (i.e., one seller), and the buyers use no-regret algorithms to learn a bidding strategy. We believe that our setting is quite realistic: in fact Nekipelov et al. [EC15] (cited in line 112 in our paper) provide a theoretical model to empirically confirm that advertisers (i.e., buyers) indeed behave as if they are running no-regret algorithms to learn their bidding strategies.

3. Regarding proofs being moved to Appendix: we would have loved to discuss more in the main body. Given the technical nature of our proofs and the 8-page limit, we had to move most proofs to Appendix. We will make sure to give proof sketches in the next version.

4. Regarding the text in line 68-70: what we mean here is that if the number of options K we provide to the buyer is too large or infinite, then the no-regret guarantee for bandit algorithms would become very weak. This is because the regret increases with K (apart from increasing with T): for example, the regret bound for the EXP3 bandits algorithm is $O(\sqrt{KT})$; if K gets quite large, say $K = \Theta(T)$, this regret bound becomes trivial/vacuous (i.e., linear in T) like we mention in lines 146-147. Thus it is important to strive to design mechanisms that have a small K (at the very least, K which grows sublinearly in T), and this is something we focus on in our paper.

Given that these were Reviewer 5's main concerns, we sincerely hope Reviewer 5 would consider revising their score.