Author Response for Submission 3526

We thank the reviewers for their effort, thoroughness, and constructive comments.

To Reviewer 1

1. We appreciate your suggestions for improving the presentation and will follow them in the final version.
2. Stability-spans have been previously used in the full information setting by Joulani et al. (2016). (Joulani et al. denote the quantity by $\tilde{\tau}$, but do not give it an explicit name.) The use of stability-spans in the analysis of delayed Exp3 is new and generalizes the role of the delay in the fixed-delay setting (Cesa-Bianchi et al., 2016).
3. The stability-span N_t is the amount of feedback that arrives between playing action A_t and observing its feedback. This may include up to $\max_s d_s$ observations from the actions that were played *before* A_t (assuming that their delay is large enough, so that they arrive after time t) *and* up to d_t observations from actions that were played *after* A_t (assuming that their delay is small enough, so that they arrive before $t + d_t$). Together it gives the factor of 2.
4. Regarding bounded losses in Theorem 1: We assume that the losses are in the unit interval $[0, 1]$, which is a customary assumption in many bandit papers. We will make sure to state this explicitly.
5. By throwing away information from observations with excessively large delays we obtain a *simpler analysis* of the algorithm. We do not claim that throwing away information lowers the regret. The analysis of weight updates for observations with large delays requires stability of the weights over the corresponding time span. When the delays are highly unequal, from the analysis perspective it is cheaper to ignore the large delays than to analyze them. As long as the number of skipped observations is comparable with the regret bound for the remaining rounds, we do not lose much in the regret bound (at most a constant factor), but significantly simplify the analysis.
6. The reasoning behind the definition of the epochs is to balance the individual terms of the bound in equation (4). The selection of β_m in equation (5) directly controls the middle term, while the doubling condition in equation (6) makes sure that the sum of the first and the last terms is of the same order as the middle term. We will add the intuition to the final version of the paper.

To Reviewer 2

1. Regarding experiments: We agree that in general experiments are a valuable addition to corroborate theoretical results, however, there are a number of reasons that make it difficult to design comprehensive experiments for our work. First of all, it is impossible to design comprehensive experiments for algorithms for adversarial problems because of the impossibility to cover all possible adversarial scenarios. Second, this is the first work on adversarial bandits with arbitrary delays and we had no natural prior work to compare to. We believe that adding experiments at this stage would constitute an overly major change, but if the reviewer has any particular setups in mind (what kind of loss sequences and delays should we test; what algorithms should we compare to) we will be happy to consider them in potential extensions of the work.

To Reviewer 3

1&2. The proposed ideas for extension of our work are very interesting! In particular, the robustness analysis and the idea of receiving the expected regret at action time and the realized regret at observation time would be an interesting variation of the problem. This would relax the assumption of "observation at action" time significantly. We believe that it should be possible to achieve regret guarantees without prior knowledge of T and D in this setting, something that has not yet succeeded in the harder "delay at observation time" setting.

3. As mentioned in our discussion, refined lower bounds for varied delays would be incredibly interesting. As we have written in the paper, our results match the lower bound up to logarithmic factors in the case of uniform delays. It is also easy to see that we match the lower bound up to logarithmic factors in the other extreme case described in Example 8: when observations for $\mathcal{O}(\sqrt{KT})$ rounds arrive at the end of the game and observations for the remaining rounds arrive with no delay. In this case there are $\Omega(T)$ no-delay rounds and we have the standard $\Omega(\sqrt{KT})$ lower bound for the no-delay rounds by the standard multiarmed bandits analysis (Auer et al., 2002), which implies the same lower bound for the whole game. This lower bound is matched by our algorithm within logarithmic factors, as described in Example 8. It does not seem trivial to obtain lower bounds for intermediate setups between the two extremes and we leave it to future work. We will add the discussion above to the paper.