We would like to thank all the reviewers for their constructive feedback. In the following, we respond (**R**) to individual concerns (**C**) summarized in italic. Citations refer to references in the paper and to the additional ones provided below.

Reviewer 1. C: *"I do agree that full information feedback is hard to expect in real scenarios,... However, the current model ...is just as unreasonable to expect. Is there an application where this is a more realistic assumption?"*
R: The main motivation for our model is a setting that is in between the full information and bandit feedback. In such settings, opponents' actions, or some similar aggregate information, can often be observed in practice. In the considered routing application (Section 4.2), the number of agents that traverse routes can be monitored and made available to the agents, while the delay functions are in general unknown to the agents. Also, when recommending items (Section 4.3), users' preferences are unknown, but one can typically observe the profile of the user who rates the recommendation. The proposed feedback model is also present in other practical applications. For example, in repeated auctions, bidders do not know their reward functions since these functions depend on complex auction-clearing mechanisms, but can often observe their payoffs as well as some aggregate information about the opponents' bids [26]. In competitions between firms in a shared market, each firm sets a price and subsequently observes its experienced revenues as well as the prices set by the other firms [27]. Finally, the feedback required by our model is natural in *cooperative* multi-agent systems where the agents communicate their actions to each other.

C: *"What is the regularity parameter γ_T in the routing example?"* **R:** Computing the *kernel-dependent* quantity γ_T is in general a hard problem, and bounds only exist for most widely-used kernels [23]. To the best of our knowledge, no bounds exist for the used polynomial kernels. We note that we observed comparable experimental performance when we switched to the Squared Exponential (SE) kernel (for which a bound on γ_T exists).

Reviewer 2. C: *"...when both players play no-regret algorithms faster convergence rates can be obtained. Is something similar possible for this model of feedback?"* **R:** Regardless of other players' strategies, the individual regret of GP-MW always contains an additive $\mathcal{O}(\sqrt{T})$ term due to learning the unknown reward function. Therefore, differently from [24], an overall convergence rate of $\mathcal{O}(\sqrt{T})$ seems inevitable.

C: *"It is interesting to understand how much feedback one really needs to get fast bounds and the trade-off between feedback and convergence rate is a research worth pursuing."* **R:** We agree that the trade-off between feedback and convergence rate is a research worth pursuing and we see our work as a contribution in this direction.

Reviewer 3. C: *"The authors do not spend enough effort to justify why the kernel assumption makes sense. Do most forms of games have this property?"* **R:** Assuming the reward function $r^i(\cdot)$ has a bounded RKHS norm with respect to some kernel k, is a standard assumption used in kernelized stochastic bandit literature (e.g., [23], [9], etc.) to effectively enforce smoothness on the reward function. This leads to the fact that similar game outcomes would produce similar rewards, and allows a player to use the observed history of play to learn about $r^i(\cdot)$ and generalize for unseen game outcomes. Note that our results are not restricted to any specific kernel function, and depending on the application at hand, various kernels can be used to model different types of reward functions. We will elaborate more on the made assumption both formally and intuitively in our paper.

We answer to the remaining questions of Reviewer 3 using the same numbering provided in the review:
(1): The intention of Table 1 was to summarize the regret bounds of algorithms that require *different feedback*. We will clarify this aspect more in the paper. Exp3 algorithm works in the standard multi-armed bandit setting and it does not exploit potentially present correlations between outcomes and rewards. Moreover, we are not aware of sharper regret bounds for Exp3 when rewards are *noisy* and satisfy the kernel assumption. Furthermore, it is not obvious how to modify Exp3 to make use of the additional feedback considered in GP-MW. On the other hand, under the proposed feedback model, GP-MW is able to exploit such correlations and achieve a kernel-dependent regret bound. Additionally, in the case in which rewards are not correlated (corresponding to a diagonal kernel), the constant γ_T in Theorem 1 grows as $\mathcal{O}(K_i)$ and hence, our approach does not provide any improvement over the regret bound of Exp3.
(3): The obtained regret guarantees also hold under *adaptive* opponents' strategies. We mentioned this in Footnote 1 and proof of Theorem 1. We will further clarify this in the paper.
(4): Yes, the improvement in Line 170 arises from the kernel assumption which allows emulating the full information feedback guarantees, similarly as in Theorem 1.
(5): This is true when it comes to our first experiment, while this is **not** the case in our real-world experiments (Sections 4.2, 4.3). For example, in 4.2, the reward function is obtained via the Bureau of Public Roads congestion model [14]. GP-MW is then run with a polynomial kernel whose parameters are learned from observed data. We experimented with different kernels and found out that similar results can also be obtained with the Squared Exponential (SE) kernel (which has universal function approximation properties and is typically a default choice when no additional domain knowledge is available).

Additional References

[26] J. Weed, V. Perchet, and P. Rigollet. Online learning in repeated auctions. *COLT*, 2016.
[27] U. Nadav and G. Piliouras. No Regret Learning in Oligopolies: Cournot vs. Bertrand. *Algorithmic Game Theory*, 2010.