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.

Reviewer 1. C: *"Can you provide some intuitions on how to deal with a setting with completely-partial (in the sense that the learner does not even observe the opponent's type) feedback?"* **R:** In such a setting (also denoted as 'bandit feedback') the learner could play according to the Exp3 algorithm [2], as discussed in Lines 137-140. Compared to StackelUCB, the reward estimates obtained by Exp3 do not exploit of the rewards bilevel structure, yielding a higher variance and an unavoidable $\mathcal{O}(\sqrt{|\mathcal{X}|})$ term in the resulting regret bound, as discussed in Line 169. **C:** *"- Can you provide more intuition as to why your regret bounds (see e.g. Theorem 1) do not depend on the size of the opponent's type space?"* **R:** Compared to state-of-the-art results (e.g., [3]), our regret bounds do not depend on the number of types, but only on the *dimension* of the corresponding type space Θ, via the maximum information gain γ_T. For instance, in case of the squared exponential kernel we have $\gamma_T = \mathcal{O}((\log T)^{d+1})$, where d is the dimension of $\mathcal{X} \times \Theta$. This is because, compared to [3], our algorithm can exploit the present correlations among different types (i.e., the fact that similar types lead to similar opponent responses) through the RKHS model.

Reviewer 2. C: *"...the opponent's type may depend on not only the learner's previous actions but also the learner's overall strategy (the learning algorithm per se), right?"* **R:** Indeed, the sequence of types can be chosen by an adaptive adversary who knows the learner's past actions and the learner's algorithm (but not the realization of its internal randomization). We will make sure to better emphasize that our model accommodates this fact. **C:** *"The assumption that the learner observes the opponent's type is a very strong one. It is unclear why this makes sense in real world problems."* **R:** We agree that observing the opponent's type is a stronger assumption than the standard bandit feedback, however in some applications (such as the ones studied in our experimental section) one may receive information about the opponent a-posteriori, that can be utilized to improve the playing strategy. In the considered traffic example, for instance, the network operator can reconstruct the past demands in the network. In security domains, one may acquire information about the attacker after an attack has taken place (e.g., as in [3]). This information can be encoded as opponent's *type*, and our work shows that it can significantly improve the learner's performance (when available) compared to only using the bandit feedback.

Reviewer 3. C: *"...while the learner agent is allowed to randomise, the opponent is not. Why? "* **R:** We considered deterministic responses since the opponent plays second, i.e., only after observing the learner's play, and hence there are no advantages in considering randomized strategies for the opponent. **C:** *"The practicality of these assumptions should be discussed in paper"* **R:** The main assumptions of our model are observing the opponent's types and assuming its response function has a small RKHS norm. Observing opponent's types is of practical interest, e.g., in security domains (see response to Reviewer 2), and a key contribution of our work is to show that such observations can significantly improve the learner performance. Assuming a small RKHS norm is a typical non-parametric assumption used in black-box Bayesian optimization to efficiently learn and optimize an unknown function by lifting it to a higher dimensional feature space. It has found several practical relevance during the past years (see, e.g., [30, 34]). The optimal kernel choice is problem-specific, although squared exponential kernels have universal approximation properties. **C:** *"In the wildlife task, why is the proposed algorithm no longer compared to other bandit algorithms as was done in the traffic task? Here, none of the baselines are learning algorithms."***R:** In the wildlife task the learner faces a single type of opponent and hence this leads to different algorithmic benchmarks than the traffic routing task. The Reviewer is correct in that Figure 2 compares our method only against offline strategies, however we have also considered the GP-UCB [34] bandit algorithm as a natural learning benchmark, as discussed in Section 4.2. A direct comparison with GP-UCB, under different learning rates, is included in Appendix F due to space limitations.
Finally, we thank the Reviewer for pointing out the relevant literature on "opponent modeling" and "type-based reasoning" which we will include in the paper. We have identified our setup as a general 'sequential game' since the key component is learner and opponent playing sequential moves, the second observing the action of the first but not necessarily playing a best-response function (as in Stackelberg games). We will clarify the distinctions and connections with the mentioned literature result in the paper.

Reviewer 4. C: *"Some related work seems to have been overlooked: Playing Repeated Security Games with No Prior Knowledge (Xu et al, AAMA 2016)."* **R:** We thank the Reviewer for bringing up this related work and we will add a reference and discussion in our paper. We would like to point out that such work focuses on playing repeated *security* games, i.e., where the learner's reward structure and corresponding feedback information follow the specific combinatorial model of allocating security resources to protect a given set of targets. The Follow-the-Perturbed Leader (FPL) online learning algorithm proposed by Xu et al. exploits this specific combinatorial structure. When applied to our general sequential games framework, such FPL-based algorithm essentially corresponds to the bandit Exp3 [2] (see Lines 137-140) which we have compared both theoretically (see discussion after Theorem 1) and experimentally (see Section 4.1) with the proposed method. **C:** *'In the experiments, different kernels were used. How does the choice of kernels affect the performance of the approach?"* **R:** In general, we observed that certain kernels are more suitable than others depending on the application. In the traffic experiment we observed similar performance with polynomial kernels of different degrees and with squared exponential kernels (which have the property of being universal approximators), while in the wildlife example we experienced similar results with Matérn kernels with different hyperparameters.