We warmly thank the reviewers for their careful reading and their remarks.

General comments We would like to point out that while Taylor's expansion or Nyström have already been explored for i.i.d. data, the existing results are limited in the adversarial context. No algorithm reached optimal regret with a per-round computational complexity $o(n^2)$. In particular, the existing analysis of Nyström (for Pros-N-Kons [13] or FOGD [16]) must restart the algorithm to update the approximation. This leads to deterioration of the regret bounds and poor practical performance. Removing these restarts was not trivial but crucial to achieving optimal rates.

We agree that the algorithms and approximation techniques used in the paper are not new (Kernel Ridge, Nyström, Taylor's approximation). Yet, surprisingly they outperform on the 1) computational complexity, 2) theoretical guarantees (regret) and 3) practical performance (see experiments) more complex algorithms (using restarts, additional projections,...) that are widely used in practice and constitute the actual state of the art. We think it is is interesting in itself to show new optimal theoretical guarantees for adversarial variants of these well-known methods.

Finally, we will take all the time necessary to carefully improve the writing and add appropriate references if necessary.

To reviewer 4

- *When $n^2 = O(e^d)$ none of the algorithms will be runnable.* We believe this is runnable in many real-life scenarios with a small number of explanatory variables but very long time series or large datasets. In online algorithms, a (close to) constant cost per iteration is of great importance. For example, when $d = 10$, we have $n^2 \gg e^d$ as soon as $n \gg 150$ which is likely to happen. In this article, we propose compelling experiments on large-scale datasets from many contexts that adapt to our context. In addition, as shown in Figure 1 or Figure 2 with $M = 2$, there is an approximation trade-off: less complexity is always possible at the cost of a sub-optimal regret.
- *Similar results for Pros-N-Kons and interest of AWV.* We guess (though it was not proved) that Pros-N-Kons may achieve similar results to Cor. 8. But this is not as clear for Thm. 9 which allows optimal rates to be obtained for many Kernels when the features are revealed sequentially. The analysis of changing approximations (without restarting the algorithm) was not trivial. It was quite specific to squared loss and KAWV. As shown in Fig. 1 (left), due to restarts, the performance of Pros-N-Kons is capped. Its current analysis does not allow it to get optimal regret.
- *Why is the m term not there in the projection error term bound of Pros-N-Kons?* The analysis of Pros-N-Kons is different and can only deal with fixed approximations and the algorithm needs to be restarted m times. Hence a regret bound deteriorated by a factor m. Our algorithm is not restarted but the approximation term is directly controlled but get multiplied by m. It yields a tighter upper-bound on the regret. Besides, note that μ and m are linked since $m \approx d_{\text{eff}}(\mu) \leq (n/\mu)^\gamma$. The approximation term of Pros-N-Kons therefore also depends on m.

To reviewer 5 First, we will have our submission proof-read for English style and grammar issues. Thanks for the typos, we will check the paper carefully.

- *Narrow audience (restricted to squared loss, improved speed but same regret bounds)* We believe that our work may be of interest to a wide audience because: 1) squared loss is widely used 2) existing algorithms that achieve the same (optimal) regret suffer from a total time complexity of order n^3 that is prohibitive for many applications.
- *Squared loss vs Lipschitz loss and comparison to [16].* Much faster rates are possible for the squared loss. In particular, the regret bound provided by [16] for Lipschitz losses (see Th. 1) is much worse than ours because:
  - it depends on the ℓ_1-dual norm which can be arbitrarily large for ill-conditioned data while ours only depends on the ℓ_2-norm and the effective dimension;
  - it is of order $\sqrt{n}$ while ours can be logarithmic.
- *Fixed Nyström dictionary and noise sensitivity.* The dictionary for Nyström is not fixed but growing randomly as new samples are observed. We consider an adversarial setting, the algorithm should therefore be robust to noise.
- *Why are our approaches faster than FOGD?* FOGD is actually the fastest algorithm. However, it has worse theoretical regret bound and poor performance in our experiments.
- *Classification experiments with squared loss.* The squared loss was used for learning but the 0/1 loss was used to evaluate the algorithms. We re-used the experimental setups of previous work [13,16] to ease the comparison and reproducibility. It should be noted that Pros-n-Kons makes a *curvature* assumption on the loss that prevents the use of losses such as hinge or logistic (possible but at the price of a exponentially small hyperparameter). So it is also natural to use Pros-n-Kons with squared loss. For the sake of coherence we used FOGD with squared loss too.
- *The experiments use PKRR rather than PKAWV.* Our experiments on real datasets are closer to an i.i.d. setting rather than an adversarial one. In this case, Kernel Ridge Regression (KRR) seems to be more suited than PKAWV (its adversarial counterpart). We will report the results of PKAWV which are very similar to PKRR. Furthermore, we can run the experiments longer. But since the results are reported in a loglog scale and since the rates are already visible, this will not add much information.