Author Response: Stochastic Bandits with Context Distributions

First, we would like to thank all reviewers for their valuable feedback. We address all concerns raised below.

Reviewer #1

Thank you for pointing out the reference (Lamprier et al., 2018). We agree that the setting shares some similarity with ours, but also highlight some key differences. The setting of Lamprier et al. (2018) uses a fixed context distribution per action (profiles in their terminology) and only one context sample per round is observed. This allows/requires to build aggregated estimates of the mean context over time. In our setup, the algorithm is granted full access to the context distribution (either exactly or via samples), and the distribution can change from round to round (chosen by an adversary). Our setting specializes to the setting of Lamprier et al. (2018) if the context distribution is fixed and the feature distribution factors over the actions (and only a single sample is observed per round). However, our setting is much more general in that it allows an arbitrary sequence of distributions as well as correlation between the feature distributions of different actions. On a technical level, Lamprier et al. (2018, Proposition 5) controls the deviation in the feature estimates in each dimension separately (which in our opinion requires a union bound over $\mathcal{X}$ that leads to a deviation of $d \log(|\mathcal{X}|)$), whereas our analysis directly bounds the predictive errors and scales with $\log(|\mathcal{X}|)$ (e.g. eq (7)). Finally, the kernelized setting was not considered in the previous work, and we reveal an interesting connection to kernel mean embeddings. We will add discussion of this related work to the updated version of our paper.

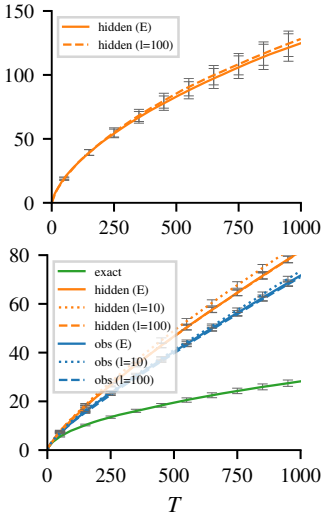
Regarding the example that illustrates the different notions of regret, a slightly more interesting case can be obtained by choosing the context as a biased coin flip. This leads to non-trivial regret for some algorithms. More complicated, lower-bound like constructions are possible, too; but this was not our intention at that point.

Reviewer #2

Regarding the stronger baseline, even if the true parameter is known to the algorithm, in general it is not possible to compete with a baseline that exploits the exact feature realization (as shown by our example). For this case, constant-per-round expected regret can only be avoided if the distribution uniquely specifies the best arm. With this constraint on the environment, our approach would already be competitive with the stronger baseline, because in this case the best arm is also identified by the best mean.

On the Movielens dataset, the unexpected ordering of the expected and sampled version is explained by the variance in the results (also note the errorbars bars). We re-ran both policies with 250 repetitions and obtained almost equal performance (no code changes).

In the crop experiment all approaches performed very similar due to an unfortunate choice in our setup. Namely, as context distribution we had chosen a Gaussian perturbation of a randomly chosen feature vector from the dataset (with variance informed from the dataset), and returned the chosen feature vector as context realization. With this setup, the context realization was always the mean of the observed distribution, which caused the different approaches to have very similar performance. We now changed the setup and center the context distribution around a fixed perturbation of the original vector, which is also a more realistic scenario. With this change, we observe a clear separation of the different configurations as shown on the right. The plots will be updated in the revised version of our paper, and we will add more discussion.



Top: Movielens with 250 repetitions. Bottom: Updated crop experiment.

Reviewer #3

We would like to emphasize that we identify a novel bandit setting that arguably covers many important applications. To someone familiar with the setting, our formulation leads to an analysis that might seem relatively straight forward, at least for the exact version of our algorithm. The sample-based version of our algorithm, however, poses additional technical challenges, as the sequential arm selection strategy can introduce bias into to estimated feature vectors. The bias carries on to the least squares regression and the regret, and needs to be controlled (Appendix A1 and A2, in particular eq (7), (11)-(12)). The improvement for the setting where the context realization is observed (Theorem 2) also requires a modification of the standard analysis (eq (17) and below). Finally, we provide a clean formulation of the kernelized setting, which we expect to be of interest for the Bayesian optimization community.

Lamprier, S., Gisselbrecht, T., and Gallinari, P. (2018). Profile-based bandit with unknown profiles. *The Journal of Machine Learning Research*, 19(1):2060–2099.