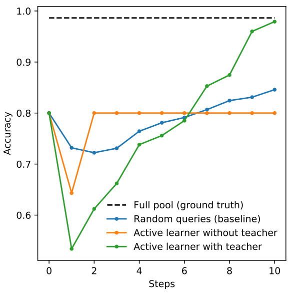
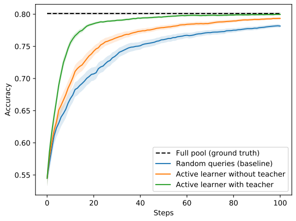
We would like to thank the reviewers for the constructive reviews.

R1, R3: *provide theoretical arguments.* **R6**: *"explanation why [...] true distribution can be sub-optimal is insufficient."* – We agree that developing theory is important. Yet, the focus of the paper is more conceptual and applied: we outline the new machine teaching setting, propose solutions to it, and give empirical validation, including a user study. We hope our paper will also entice others to contribute to this line of work, also more theoretically. In ideal settings (exact computation, etc.), the solution to the teaching MDP will (on average) be at least as good as the true distribution, up to the planning horizon, for the objective determining the reward of the MDP (since by definition, the solution maximizes that). The illustrative examples in the paper show cases where the teacher can do better than the true data labels, and our empirical experiments show improved performance.



R1: *empirical study in teaching an active learner* – Good point, our main focus was on the bandit setting, but we will add the empirical results for teaching of the uncertainty-sampling-based logistic regression active learner for the illustrative example (upper figure on the right; full horizon planning teacher) as 4th panel to Fig 1, and for the Wine dataset (lower figure on the right; 1-step planning teacher) to the supplement. Both use independent test sets to measure performance, averaging over 100 repetitions of runs, and the teachers optimize for prediction accuracy (based on having knowledge of the full pool, including labels).



R1: *"body of theoretical work on teaching a strategic (teacher-aware) learner [...] missing from the related work"* – Thank you for the references, we will include them to the related works. Their theory has important assumptions such as concept-consistent labels, which we relax. Thus, they are not directly applicable, albeit highly relevant.

R3: *"[...] the teacher's knowledge comprises the reward probabilities of individual arms, with model parameters θ, which allows the teacher to strategically alter responses y_i, by means of an MDP. For this to work, one "bandit" must inform another [...] Exactly how does this work for Bernoulli bandits? [...]" And question about inferring θ^*/θ.* – Indeed, we consider Bernoulli bandits with arm dependencies, where a response to an arm gives also information about other arms' rewards. Knowing the arm dependencies allows the teacher to better predict which arms the learner is more likely to query in the future, and thus direct the learner. The ground truth (reward generating) θ^* and θ are equal: given enough queries, the learner can infer θ without a teacher. But a teacher can make the learning faster (and, further, a teacher-aware learner can improve over baseline learner).

R3: *source code* – We will provide a link upon acceptance.

R5: *"1) From the abstract, I was not able to tell what the paper is about and what its main ideas and contributions are. [...] 2) l. 98: Why is a deterministic learning algorithm considered? [...] "3) Is the teaching MDP formulation of Sec. 3.2 a novel contribution [...] 4) l. 139: Is the teacher's reward the same as the reward previously defined for the learner? [...] 5) l. 160: The choice of the reward function as the scalar product between next sample location and true parameters is unclear to me. [...] 6) Why is the model in (3) chosen? [...]"* – Thanks, we will clarify and discuss these in the revision: 1) We will rephrase the abstract to: first, clearly state the new machine teaching setting we propose; second, our solutions to teaching problem and learning from a teacher; third, its advantages and increase in empirical performance. 2) We focused on deterministic learning algorithms, but there is no inherent reason why stochastic learning algorithms could not be considered, apart from making the problem more complex. 3) Yes; (PO)MDPs have been used to model teaching, as we mention in the Related works, but not in a setting with an active sequential learner. 4&5) The teacher's reward is the same as the learner's expected reward (both include the scalar product for the chosen arm x_t at current iteration), except we didn't include the logistic function $\sigma(\cdot)$ in the teacher's reward to simplify the formulas for the teaching model. We will clarify this in the revision. 6) The softmax policy is a common model in probabilistic and MaxEnt reinforcement learning. It has also been shown to model human choice behaviour well.

R6: *clarity* – We will use the added 9th page for clearly commented pseudo-code algorithms.

R6: *"The bandit algorithm seems not correct. [...] assumption that the x_t received in all iterations are independent"* – It is correct; our Thompson sampling algorithm is the standard version used for Bayesian models (Alg. 4 in our ref. 34: `https://arxiv.org/abs/1707.02038`), here sampling from the posterior of the generalized linear model (GLM). We only assume the independence of the "noise" in the Bernoulli distributions generating the rewards at each time step, thus the likelihood terms factor; on the other hand, the posterior distribution is conditioned on the observed arm feature vectors x_t and the distribution of x_t is not modelled (as is usual for GLMs). We will add pseudo-code algorithms in the revision, which will also make this clearer. Thanks for the GLM reference, we will add it in the revision.

General: – Thank you for the minor comments which we will naturally fix.