MACHINE TEACHING OF ACTIVE SEQUENTIAL LEARNERS

Tomi Peltola, Mustafa Mert Çelikok, Pedram Daee, Samuel Kaski
Helsinki Institute for Information Technology HIIT, Department of Computer Science, Aalto University, Finland
`firstname.lastname@aalto.fi`

TL;DR:
- **HOW TO STEER AN ACTIVE MACHINE LEARNER THAT QUERIES LABELS SEQUENTIALLY?**
- **WE FORMULATE THE TEACHING PROBLEM AS A MARKOV DECISION PROCESS, WITH LABEL CHOICE AS ACTION.**
- **A TEACHER TEACHING WITH INCONSISTENT LABELS CAN BEAT CONSISTENT LABELS.**
- **WE FURTHER ENDOW THE LEARNER WITH A MODEL OF THE TEACHER.**
- **APPLIED TOWARDS MODELLING STRATEGIC USER BEHAVIOUR IN INTERACTIVE INTELLIGENT SYSTEMS.**
- **ACCEPTED TO NEURIPS 2019, PREPRINT:** `https://arxiv.org/abs/1809.02869`

INTRODUCTION

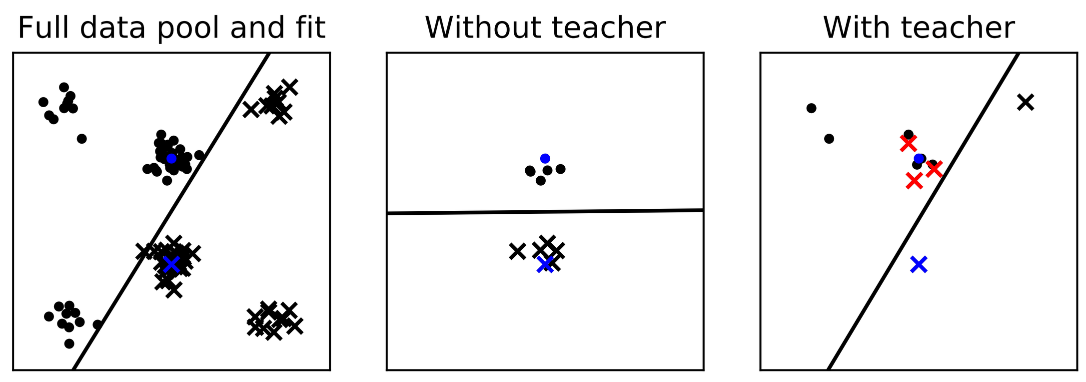
Machine teaching addresses the problem of finding the best training data that can guide a learning algorithm to a target model with minimal effort.

- Traditionally, the teacher provides data by sampling labels from the true data distribution (consistent teacher).
- Providing true labels can be sub-optimal in finite–horizon tasks for sequential learners that actively choose their queries (see right "Without teacher" panel).

Contributions

- We formulate this sequential teaching problem, as a Markov decision process, and allow the teacher to provide data inconsistent with the true distribution (see right "With teacher" panel).
- We address the complementary problem of the teaching-aware learner by endowing the learner with a model of the teacher. The final inference problem reduces to inverse reinforcement learning.
- We evaluate the formulation with multi-armed bandit learners in simulated experiments and a user study.

The approach gives tools to taking into account strategic (planning) behaviour of the users in interactive intelligent systems, such as recommendation engines.



Example of teaching effect on pool-based logistic regression active learner. Starting from blue data,

- the learner without teacher, fails to sample useful points from the pool to learn a good decision boundary.
- a planning teacher can help the learner by switching some labels (red points).

MODELLING

Task: Goal of the learner and the teacher is to learn and teach the best possible model of the true data distribution respectively. We consider Bayesian Bernoulli Bandits as a specific case of this task.
Interaction: At each iteration t, the bandit learner queries an arm i_t and the teacher provides a stochastic reward y_t.
Learner model:

- Logistic regression based contextual multi-armed bandit with Thompson sampling for exploration–exploitation trade-off.
- Teacher's stochastic reward responses to queried arms i are modelled as $\pi_i = \sigma(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{w})$, where $\boldsymbol{w}$ is a model parameter.
- Thompson sampling is used to select the next arm: (1) sample $\tilde{\boldsymbol{w}}$ from $p(\boldsymbol{w} \mid \mathcal{D}_t)$, and (2) choose $i_{t+1} = \arg\max_k \pi_k^{(\tilde{\boldsymbol{w}})}$.

Teacher model:

- Teacher models (right) interpret the teacher's actions (likelihood for $\boldsymbol{w}$).
- Naive: Teacher provides the reward response by sampling from the true distribution.
- Planning: Teacher anticipates next arms based on a Markov decision process, with transition dynamics defined by a nested, simpler model of the bandit, and chooses her action to steer towards good arms in the future.
- Mixture: Learns whether the teacher uses naive or planning strategy.

Computation: Laplace approximation implemented in the probabilistic programming language Pyro.

Teacher models

Naive:

$$p_{\mathcal{B}}(a_t \mid i_t, \boldsymbol{\pi}) = \text{Bernoulli}(a_t \mid \pi_{i_t})$$

Planning:

$$p_{\mathcal{M}}(a_t \mid i_t, \mathcal{M}_t, \boldsymbol{w}) = \frac{\exp\left(\beta Q^*_{\mathcal{M}_t}(s'_0, a'_0; \boldsymbol{w})\right)}{\sum_{a'} \exp\left(\beta Q^*_{\mathcal{M}_t}(s'_0, a'; \boldsymbol{w})\right)}$$

Mixture:

$$p_{\mathcal{M}/\mathcal{B}}(a_t \mid i_t, \mathcal{M}_t, \boldsymbol{w}, \boldsymbol{\pi}, \alpha) = \alpha p_{\mathcal{M}}(a_t \mid i_t, \mathcal{M}_t, \boldsymbol{w}) + (1-\alpha) p_{\mathcal{B}}(a_t \mid \pi_{i_t})$$

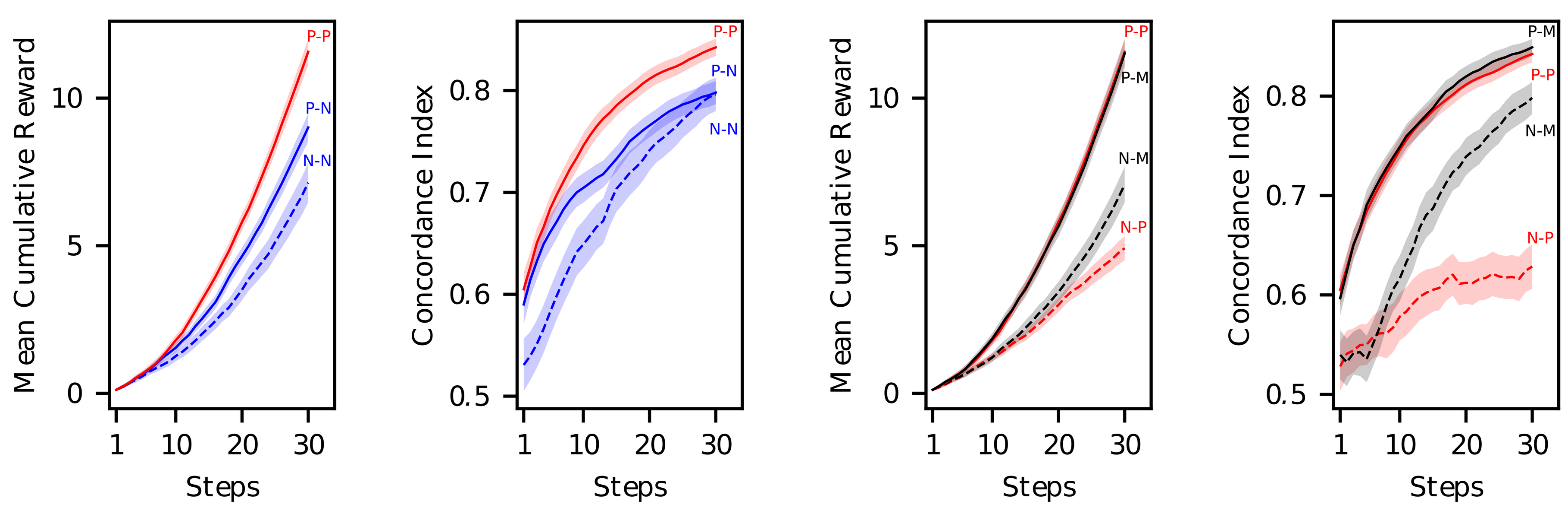
EXPERIMENTS

Setup:

- Word search study: the teacher selects a target word and the learner tries to guess the word by asking sequential questions.
- Learner: "Is this word relevant to the target?", Teacher: Yes/No
- **Below**: Simulated teachers and learners.
- **Right**: Human teachers ($n = 10$) with learners having mixture and naive teacher models.

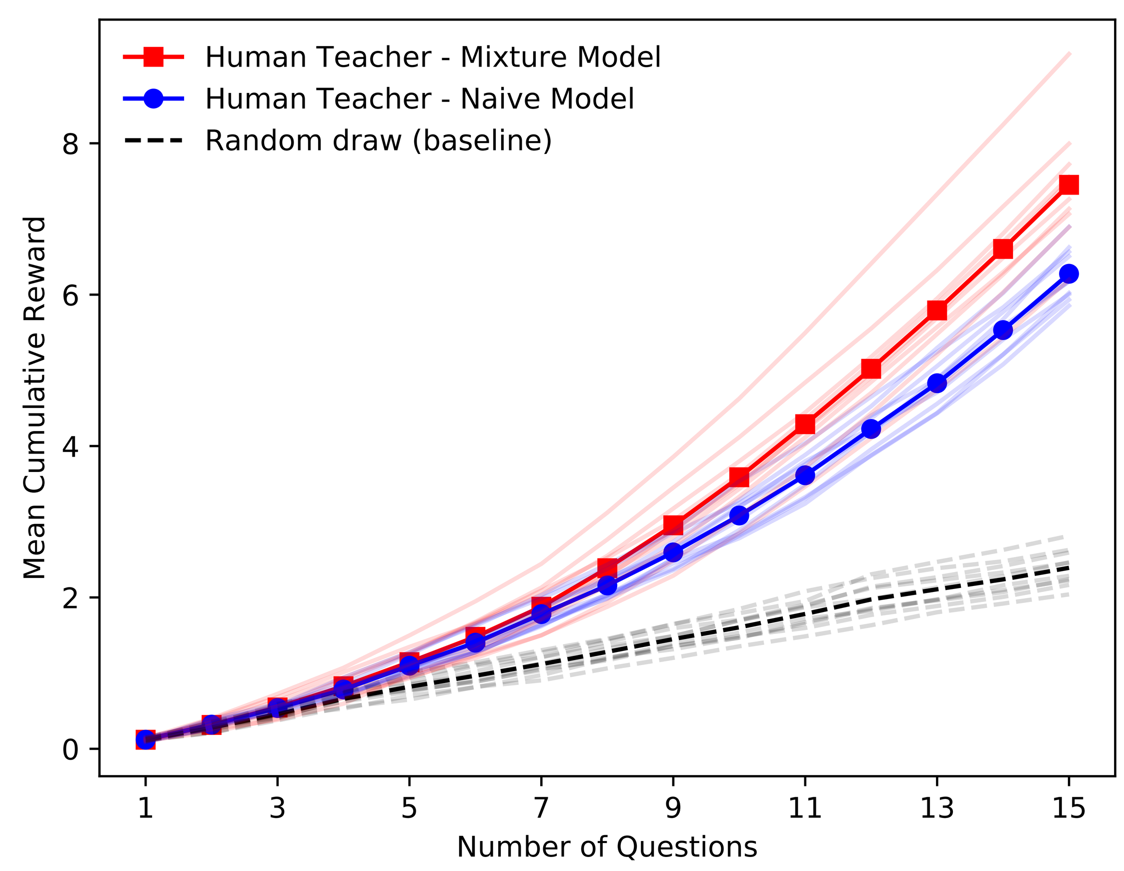
Simulation Results:

- The planning teacher can steer a teacher-unaware learner to achieve a marked increase in performance compared to a naive teacher (P-N vs N-N; left-side panels)
- The performance increases markedly when the learner models the planning teacher (P-P; left-side panels)



User Study Results:

- Participants achieved noticeably higher rewards while interacting with a learner having the mixture teacher model (red), compared to the naive teacher model (blue).



CONCLUSION

- We have introduced a new sequential machine teaching problem, where the learner actively chooses queries (e.g., in active learners and multi-armed bandits) and the teacher provides responses.
- The teaching problem is formulated as a Markov decision process, the solution of which provides the optimal teaching policy.
- Teacher-aware learning from the teacher's responses is formulated as probabilistic inverse reinforcement learning which illustrated a performance improvement.
- The proposed teaching framework holds promise for a feasible and natural computational approach in modelling active user behaviour in interactive intelligent systems.

Acknowledgments: This work was financially supported by the Academy of Finland (grants 313195, 305780, 292334, 294238, 319264, and 295503). Mustafa Mert Çelikok was partially funded by the KAUTE Foundation. We acknowledge the computational resources provided by the Aalto Science-IT Project. We thank Antti Oulasvirta for comments that improved the manuscript.