View Meta-Reviews

Paper ID

542

Paper Title

Decision Making in Hybrid Environments: A Model Aggregation Approach

Track Name

COLT2025

META-REVIEWER #1

META-REVIEW QUESTIONS

2. Meta-review

The paper considers a “hybrid” variant of the general decision making (subsuming bandits and RL) setting of Foster et al. (2021, 2022) in which the environment is has a combination of stochastic and adversarial components (e.g., a fixed MDP with adversarial, time-varying rewards across episodes). The main contribution is to give a generalization of the Decision Estimation Coefficients (DEC) that characterizes the complexity of learning in these environments, which requires new, non-trivial algorithmic modifications of the AIR framework of Xu and Zeevi.

Reviewers were uniformly positive about the paper and felt that it advances the DEC theory in a non-trivial way, and thus merits acceptance.

View Reviews

Paper ID
542

Paper Title
Decision Making in Hybrid Environments: A Model Aggregation Approach

Track Name
COLT2025

Reviewer #1

Questions

1. Rating (scale of 1-7)
5: Good paper, accept

3. Review
The paper advances the line of work on decision estimation coefficients along several axes. First the paper develops a generalization that characterizes hybrid stochastic-adversarial environments, such as MDPs where the dynamics are fixed and stochastic, but the reward function changes from episode to episode. This is done with a more general approach where the learner considers a partition over the hypothesis/model class, where the choice of partition results in a tradeoff between estimation complexity (regret of the prediction component) and decision-making complexity (actual value of the DEC). This approach is flexible enough to capture both model-based and model-free settings and as a demonstration, the paper recovers the \sqrt{T} regret for the well-studied linear Q*/V* setting.

The paper primarily operates in the AIR framework of Xu-Zeevi. One starts with a model class M and a policy class \Pi and defines a partition \Phi over Mx\Pi. The environment is restricted to choose a \phi \in \Phi and then can choose (M_t,\pi_t) \in \phi adversarially in each round. AIR is defined in the natural way, rather than maintain a posterior over the optimal policy \pi^\star or the true model M^\star (which no longer exist), we maintain a posterior over the true partition cell \phi. The algorithm is to find the distribution that optimizes the \phi-restricted AIR objective (given the current estimate of the partition cell and deploy the induced policy. This algorithm has regret bounded by the phi-restricted AIR plus the estimation complexity of estimating the partition cell \log |\Phi|. The key tradeoff is that estimation complexity can decrease, demonstrated via log |\Phi|, while decision-complexity can increase as demonstrated via the \phi-AIR.

The paper then studies an important special case of a "independent comparator" where the choice of model and choice of comparator are decoupled. Bounds that scale with |M||\Pi| can be highly suboptimal here. To address this the paper presents a decomposition into two games, one where the learner choose policy for every model (partition cell) but only gets observations from the correct one and another where the environment chooses a policy for every model, reveals this to the learner, who then chooses the decision making policy. The paper shows that the latter game can be solved independently of the policy space in general. At the same time, in some settings, like adversarial MDPs with full information on the reward, the former game is also easy to solve.

Lastly the paper applies these general results to two main settings. The first is MDPs with fixed, stochastic transitions, and adversarial rewards, both with full and bandit feedback. A perhaps surprising finding is that these results improve on the best known results for this setting, specifically for the low rank MDP variant of this problem. The second is model free RL in stochastic and hybrid MDPs.

Overall, I think the paper is pretty interesting and should be accepted to the conference. The results advance our understanding of DEC and related information-theoretic approaches to decision making, as well as improve state of the art in traditional RL formulations like linear/low-rank MDPs. Further, the results is clearly presented with sufficient exposition and discussion regarding connections to prior work.

---- Post Rebuttal ----
Thanks for the rebuttal. I remain positive on the paper.

Reviewer #2

Questions

1. Rating (scale of 1-7)

4: Marginally above acceptance threshold

3. Review

This paper considers the complexity of general decision making in environments which lie between adversarial and stochastic setups, for example, adversarial MDPs where transitions are stochastic but rewards are adversarial.

Strength

1. Based on the AIR framework, this paper generalizes it via defining a family of disjoint subsets of (M,\Pi). This definition swiftly captures many setups, including the existing

stochastic and adversarial setups as well as the stochastic-transition adversarial-reward setups induced by adversarial MDPs.
2. Various specific examples, including model-based adversarial MDPs, model-free stochastic MDPs, and model-free full-information adversarial MDPs. Specifically, when applied to adversarial low-rank MDPs as an example, the general algorithm crafted for this setup directly induces an $O(\sqrt T)$ UB which is better than existing $T^{2/3}$ ones. When applied to adversarial general-function approximation MDPs, this paper gives a first step in the literature.

Weakness

1. While the definition of Φ is good, it is questionable whether everything in this paper is solely due to the new definition of Φ. Specifically, what would be the technical contributions of these results?
2. The application to adversarial MDPs does not result in computationally efficient algorithms and cannot motivate the improvement of existing RL algorithms. Thus it's hard to say whether these results are helpful enough.

Reviewer #3

Questions

1. Rating (scale of 1-7)
5: Good paper, accept

3. Review
The paper proposes a general extension to the Decision Estimation Coefficient (DEC) framework for decision-making in hybrid environments, where transitions remain fixed while rewards change adversarially. Their framework involves learning over partitions in the joint model and policy space and allows for trade-offs between estimation complexity and decision complexity as a result. The authors apply this framework to model-based and model-free learning in reinforcement learning (RL) settings and provide improved bounds for the hybrid setting.

The paper is well-organized and presents a natural and flexible extension of the DEC framework through the introduction of the Phi-restricted environment.

Can the authors elaborate on the challenges in extending their framework to match known bounds in general stochastic settings? Are the limitations fundamental to the Phi-restricted environment?