Abstract

We consider the dueling bandits problem, a sequential decision task where the goal is to learn to pick `good' arms out of an available pool by actively querying for and observing relative preferences between selected pairs of arms. The noisy observed preferences are assumed to be generated by a fixed but unknown stochastic preference model. Motivated by applications in information retrieval, e-commerce, crowdsourcing, etc., a number of bandit algorithms have been proposed in recent years for this task. These have mostly addressed restricted settings wherein the underlying preference model satisfies various structural assumptions. such as being based on a random utility function or a feature-space embedding, or satisfying transitivity or sparsity properties, or at least possessing a Condorcet winner { a single `best ‘arm that is preferred to all others. In seeking to move beyond such restricted settings, there has been a recent shift towards alternative notions of `good' arms (including Borda, Copeland and von Neumann winners).

In this work, we extend the dueling bandits problem by adopting, as the desired target set of good (or 'winning') arms, a number of tournament solutions that have been proposed in social choice and voting theory literature as embodying principled and natural criteria for identifying good arms based on preference relations. We then propose a family of upper confidence bound (UCB) based dueling bandit algorithms that learn to play winning arms from several popular tournament solutions, the top cycle, uncovered set, Banks set and Copeland set.

We derive these algorithms by first proposing a generic UCB-based framework algorithm that can be instantiated for different tournament solutions by means of appropriately designed `selection procedures. We show sufficiency conditions for the resulting dueling bandit algorithms to satisfy distribution-dependent, horizon-free bounds on natural regret measures defined w.r.t. the target tournament solutions. In contrast to previous work, these bounds do not require restrictive structural assumptions on the preference model and hold for a range of different tournament solutions.

We develop selection procedures that satisfy the sufficiency conditions for a number of popular tournament solutions, yielding dueling bandit algorithms UCB-TC, UCB-UC, UCB-BA and UCB-CO for the top cycle, uncovered set, Banks set and the Copeland set respectively.

The $O\_{K^2 \ln T\, g^2}\_$ bounds we derive are optimal in their dependence on the time horizon T. We show that for all of these tournament solutions, the distribution-dependent `margin' g is lower bounded by the separation or the relative advantage of top cycle arms over non-top cycle arms. While O(K ln T) bounds are known for Condorcet models, our $O(K^2 \ln T)$ bounds extend to more general models as well as other tournament solutions. We empirically validate these claims and evaluate the proposed algorithms, comparing them to dueling bandit algorithms RUCB, SAVAGE and BTMB over synthetic and real-world preference models. We show that the UCB-TS algorithms perform competitively over models that possess a Condorcet winner, but out-perform the other algorithms over more general models that do not possess a Condorcet winner.