Adaptive Experimental Design with Temporal Interference: A Maximum Likelihood Approach

Response to reviewers

We are grateful to the referees for their thoughtful comments regarding our paper. Regarding typos and other detailed suggestions, we plan to incorporate those prior to submission of our camera ready version. Below we have also provided feedback in response to major comments and requests from the referees.

Practical considerations. Several referees commented on practical considerations (Reviewers 1, 3, and 4): in particular, the performance of maximum likelihood estimation (MLE) with large state spaces and on finite horizons, as well as its computational complexity. More broadly, we note that in followup work, we have developed an alternative approach to experimental design for temporal interference using sample average estimation (SAE). In this work, we study sampling strategies based on the *regenerative method*; together, SAE and regenerative policies ensure consistency while being practically implementable. The idea is that we commit to a single state (the *regeneration state*), and only allow ourselves to switch chains in this state. This method can be used together with sample average estimation (SAE) of rewards, rather than (MLE). Further, it requires no advance knowledge of the state space, other than the regeneration state. This is a much more practical, scalable solution; although it is not as sample efficient as the policy and MLE in our present paper, we can use similar techniques to obtain the optimal regenerative policy with SAE. Though we could not include this work due to space constraints, we will add some discussion of this extension to the paper to address concerns regarding practical implementation.

Reviewer 1. Regarding how to interpret our results, note that consistency follows under very general conditions: *any* sampling procedure that samples both experiments infinitely often in each state will ultimately yield consistent estimates via the MLE. By using the MLE instead of just averaging rewards obtained in runs of each chain, we avoid temporal interference completely. The main contribution of our paper is to provide a strong characterization of the *sample efficient* experimental design for the MLE.

Regarding practical considerations and state space complexity, note that our theory shows that for any finite state space our policy eventually outperforms any other TAR policy. Nevertheless, you are right that for a given sample size, our policy may be computationally complex; see our comments above on practical considerations.

Regarding multiple treatments, we conjecture that the optimal policy in the multiple treatments setting eventually looks essentially like ours, once the best and second best treatments have been identified.

Reviewer 2. Thank you for your review. Regarding making the paper more accessible to the Neurips community: we will plan to provide some more intuition for the main results in the final version.

Reviewer 3. Regarding the convex optimization problem in Theorem 13 and Section 6, we expect that in many applications the switching time is slow enough relative to computation time, so that standard convex optimization techniques can be employed. That said, we agree that computational complexity is an important practical issue; see our discussion above.

The set $\mathcal{K}$ is a closed, compact, convex polytope in $2|S|$-dimensional space (where S is the state space).

Regarding what policies are time-average regular (TAR), we emphasize that TAR is a weak regularity requirement: it says that the fraction of time steps in which chain ℓ is sampled in state x converges to a well-defined random variable (possibly deterministic). Virtually any reasonable policy will satisfy this requirement. Policies which, e.g., switch chains on exponentially increasing timescales will not be TAR, but such policies are not likely to be used in practice.

Reviewer 4. Regarding more general optimality results, we agree with your conjecture; this remains an important open direction. Note that for any fixed TAR policy, the MLE is asymptotically efficient given the samples collected by that policy. Further, our results show how to compute the optimal TAR policy when estimating via the MLE.

Regarding obtaining higher reward by combining the two chains, our work is entirely focused only on *estimation* of the difference in steady state rewards of the two chains, rather than *optimization* of the cumulative reward obtained. An interesting question concerns whether regret optimal policies can be designed to maximize the cumulative reward using only the two chains; this is an interesting reinforcement learning problem for future study.

Other comments: Synthetic evaluation. We agree with Reviewers 1 and 3 that synthetic evaluation would be valuable to add to our paper. Synthetic evaluation would primarily be valuable to study finite horizon performance, as our paper provides a full characterization of optimal asymptotic sample efficiency. We also emphasize that our paper is primarily a theoretical study of optimal experimental design in this setting; practical considerations can lead to different preferred designs (cf. our discussion above). Intuitively we believe the regenerative method leads to designs with improved finite horizon performance; we plan to carry out a synthetic study to compare and contrast finite horizon performance of various designs as part of our future work.