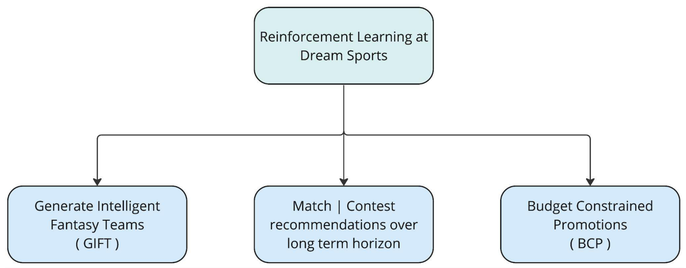
Reinforcement Learning Systems

Tanmaya Shekhawat, Dhruv Nigam, Hitesh Kapoor, Ilari Paasivirta

SCAN ME

Reinforcement Learning Systems in Dream Sports

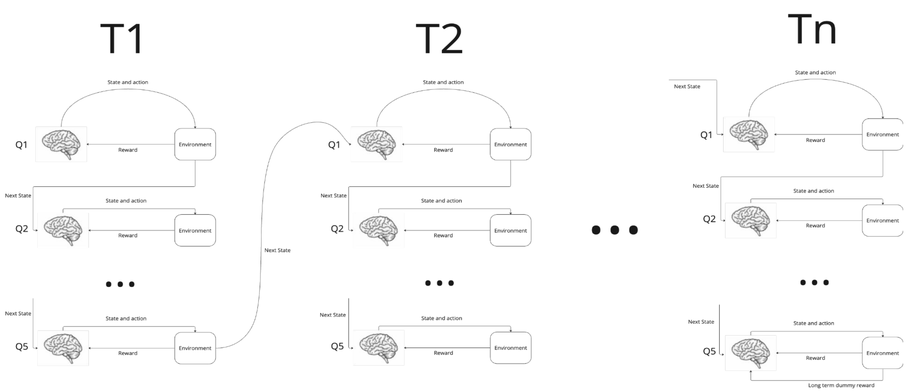
- Bandit-based systems are currently operational in production for intelligent traffic routing.
- Long-term horizon planning guides strategic actions that lead to enhancements in user engagement and business metrics.
- An imitation learning-based agent is undergoing experimentation to optimize promotions during its trial phase.



Cascading Q learning for Recommending contests

Utilization of Q learning algorithms to iteratively improve contest recommendations, creating a cascading effect where the recommendation of one contest influences the subsequent recommendations, aligning with a strategic long-term planning perspective.

Continuous optimization of contest suggestions through cascading Q learning, ensuring that if contest C1 is recommended, another Q learning process considers recommending contest C2, contributing to a coherent and personalized sequence of recommendations guided by a long-term horizon vision.



Budget Constrained Promotions

Discounts are employed as a mechanism to boost user engagement and foster retention. Considering the financial implications of discounts to the business, it is important to ensure that the benefits, such as improved retention, are comparable with the costs. Our goal is to establish policies that promote user journeys, effectively engaging customers over the long term under a given dynamic budget constraint.

This can be treated as Constrained Markov Decision Process

The CMDP formulation:

$$\max_{\pi} \sum_{t=0}^{\infty} \gamma^t E_{(s_t,a_t)\sim\pi}[R(s_t, a_t, s_{t+1})]$$

subject to:

$$E_{(s_t,a_t)\sim\pi}[C(s_t, a_t)] \leq \text{budget}$$

The Q-function :

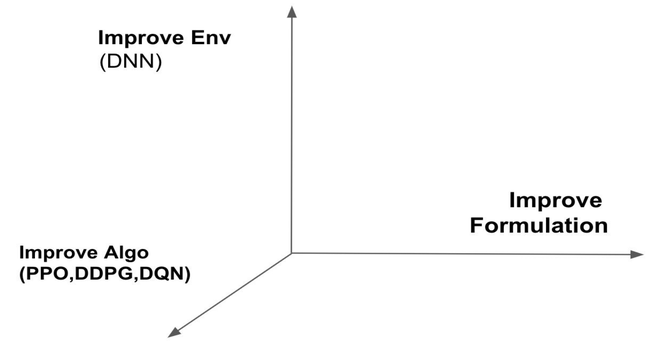
$$Q(s,a) = R(s,a) + \gamma \sum_{s'} P(s'|s,a) \max_{a'} Q(s',a') - \lambda \cdot C(s,a)$$

where:

- π is the policy to be optimized.
- $Q(s,a)$ is the Q-function for state s and action a. The state s will be comprised of user features and the overall Dream11 system features and the action a is the actual promotion given
- $R(s,a)$ is the immediate reward of taking action a in state s which signifies if the user availed the promotion and joined a contest,
- γ is the discount factor,
- $P(s'|s,a)$ is the transition probability from state s to s',
- $C(s,a)$ is the cost associated with taking action a in state s,
- λ is the Lagrange multiplier associated with the budget constraint.

This Q-function formulation reflects the trade-off between maximizing expected cumulative rewards and adhering to the budget constraint for promotions and discounts in Dream Sports.

Future Work



Constrained Optimization Post training Agent

Challenging to apply constraint at user level and keeping the budget at system level.

Alternate solution could to be to apply linear optimization once the RL agent is trained to adhere to the decided budget.

Advantage of this method is that budget can be dynamically updated [2].

References

1. Constrained Markov Decision Processes via Backward Value Functions Harsh Satija Proceedings of the 37th International Conference on Machine Learning, Online, PMLR 119, 2020.
2. Chen, Fanglin, et al. "BCRLSP: An Offline Reinforcement Learning Framework for Sequential Targeted Promotion." arXiv preprint arXiv:2207.07790 (2022).
3. Akifumi Wachi ,Yanan Sui Safe Reinforcement Learning in Constrained Markov Decision Processes Proceedings of the 37 th International Conference on Machine Learning, Online, PMLR 119, 2020