Monte Carlo Rollout Policy for Recommendation Systems with Dynamic User Behavior

Rahul Meshram (IITM)
Kesav Kaza (IITB)

COMSNETS 2021, Bangalore

5th January 2021

Motivation

- User visits to a recommendation system (RS)
- RS recommends item sequentially to the user
- The user provides feedback based on his interest.
- Recommendation systems examples—Netflix, Amazon Prime, Spotify, Youtube, etc.
- Recommendation system generates personalized playlists.
- The playlist is generated using information from watch history or information from social networking sites.

Recommendation systems models

- Collaborative filtering approach:
- Matrix completion method for RS:
- In these models: it is implicitly assumed that the user interest is static and the current recommendation does not influence future behavior of user interest.
- Our model: We assume that user behavior is dynamic, i.e., the user interest is influenced by current recommendation as well as preceding recommendations.
- We consider Markov model for user interest, a state describes the intensity level of preferences.
- Objective: Model and analyze the dynamic playlist generation systems using binary feedback from user

Our model assumptions

- There are N independent items
- There are different items, the user state for different items will be different
- State of the user interests for items is not observable
- Belief about the state is maitained by RS
- State evolution for each item is different and this evolution depends on whether item played or not
- Only one item is played to the user at a time
- Each item can be modeled as partially observable Markov decision process (POMDP)
- RS has many items and these are weakly coupled since only one item is allowed to play

Connection to Restless multi-armed bandit

- There are N independent arms.
- Each arm can be in one of many states.
- At each time step, one arm is played.
- Playing of arm yields a unit reward that depends on the state of that arm.
- State of each arm evolves at every time step.
- This evolution depends on whether an arm is played or not.
- Goal: Determine sequence of arms to play that maximizes a long run reward function.

Solution Approach

- RMAB is PSPACE complete problem. Hence the optimal solution is difficult.
- Popular heuristic policies are
  - Myopic policy: Play item with highest immediate expected payoff
  - Whittle index policy [**?**]: Play item with highest index, the index is mapping from state to a real number
    - Difficulty: Need to show indexability and obtain closed form expression for index in terms of state and rewards
    - In general, no closed form solution for index with POMDP model
    - Literature: [**?**], specialized model is studied with two state POMDP, special structure is assumed, and index formula is derived
  - Our approach: Monte Carlo rollout policy
    - This is simulation based approach for complex problem
    - No structural assumption is made
    - Recent literature [**?**], [**?**], [**?**]

Monte Carlo rollout policy

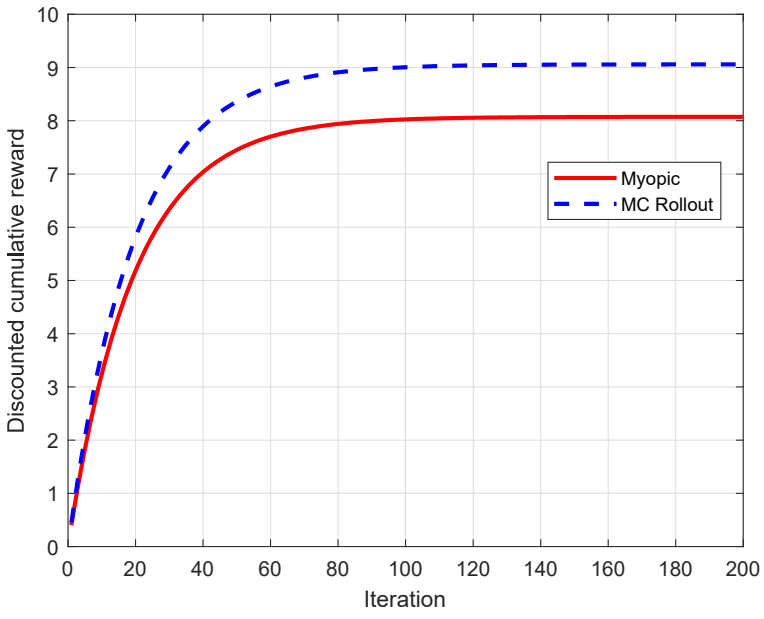
- L trajectories are simulated for a fixed horizon length H using a fixed policy ϕ
- The information obtained from a single trajectory is $\{\pi_{j,t,l}, a_{j,t,l}, R^{\phi}_{j,t,l}\}_{j=1,t=1}^{N,H}$
- Then, the value estimate for state π and action a over L trajectories under policy ϕ is

$$\widetilde{Q}^{\phi}_{H,L}(\pi, a) = \frac{1}{L}\sum_{l=1}^{L} Q^{\phi}_{H,l}(\pi, a, W) = \frac{1}{L}\sum_{l=1}^{L}\left[\sum_{h=1}^{H}\beta^{h-1} r(\pi_{h,l}, a_{h,l}, \phi)\right]$$

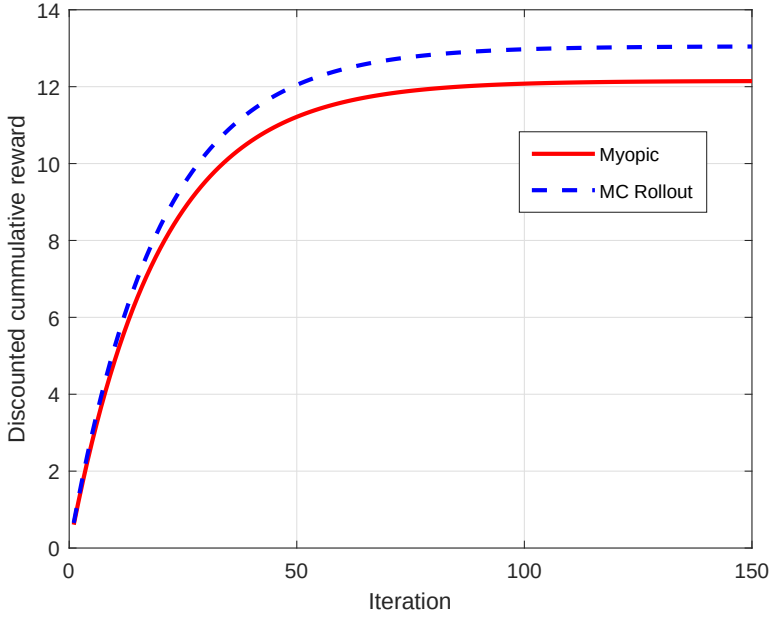
- A one step policy improvement is performed, and the optimal action is selected according follow rule.

$$j^*(\pi) = \arg\max_{1\le j\le N}\left[r(\pi, a=j) + \beta\widetilde{Q}^{\phi}_{H,L}(\pi, a=j)\right].$$

Numerical Examples



a) $N = 5$



b) $N = 15$

Remark

- Monte Carlo rollout policy performs better than myopic
- No structural assumption on transition dynamics is assumed

Concluding remarks and future work

- Concluding remarks:
  - We presented a new Monte Carlo rollout algorithm for RS with Markov model.
  - We demonstrated the performance of the algorithm on a small scale example.
  - We observed that Monte Carlo rollout policy performs better for arbitrary transition dynamics
- Future work:
  - How to design RS assuming there is cognitive limitation of human and information overload problem
  - Modeling of user dynamic behavior using other models (Non Markovian) which can characterize complex behavior of user.

Bibliography I