Project Proposal

Graded

Group

Chris Lapop Salazar
Eric Peterson
Michelle Nhung Le
View or edit group

Total Points

9 / 10 pts

Question 1

Project Proposal

9 / 10 pts

– **0 pts** Correct

– **1 pt** Went over page limit

– **0.5 pts** Didn't use NeurIPS formatting

[R1]
Super interesting proposal! I would mention that I don't see why the categorical distribution in Phase 1 is necessary if there's nothing distinguishing the cards except their values: each card can essentially be identified with its actual value, returning the setup to an actual MAB. That being said, I would maybe skip this phase entirely; I think just one of the later parts would suffice to make an interesting project. Just be careful with time management; I would recommend starting on the later parts much earlier than you've indicated, as they will almost surely take _disproportionally_ more time. But the problem seems tractable. Make sure to leave yourself more time than you think you'll need. Good luck!
[R2]
Great job, Card Collecting is a great example of a game that can be modeled by an MDP. I really do love this idea. Your environment setup and potential algorithm choice is detailed and clear. I would be curious of a theoretical comparison to the coupon collector problem as a baseline. I particularly enjoy the extension to a particular deck or a particular merchant with trades - I think that the transformation from a MAB to a MDP setting is worth discussing in more detailed, the pros and cons of each. I would also cite some relevant literature and how you are improving upon or doing something different: a quick google search reveals https://arxiv.org/abs/1506.03379. How does your analysis differ from previous literature in this area? Why might we care about this problem in the real world, and what are the limitations of your model? This is an ambitious project - make sure to go step by step and it's okay if you don't make it through all your listed phases, just make sure each phase's analysis is solid and complete before moving to the next.

CS/STAT 184(0) Project Proposal

Christopher Lapop Salazar
Eric Peterson
Michelle Nhung Le
clapopsalazar@college.harvard.edu
peterson@g.harvard.edu
michellele@college.harvard.edu

Abstract

The current form of the project is to work through a model of hobby card-collecting. This would include purchasing box sets and pulling the cards within the box sets, each with its own market values and probability of being pulled. The model will later pivot toward the objective of specific collection of cards using these boxes.

1 Project Phases

1.1 Phase 1: Unknown box pull rates, minimize regret with different box choices

This phase will focus on a generalization of the Bernoulli Bandit. Instead of outputting 0 or 1 drawn from a Bernoulli distribution as we learned in class, a pull outputs a card, which is represented as a one-hot vector drawn from a categorical distribution. The distributions of the different arms are not known by the RL agent ahead of time, so exploration is required.

This is a multi-armed bandit problem with categorical outcomes. Environment settings are as follows:

State Space: stateless problem (MAB setting)

Action Space: $\mathcal{A} = \{1, ..., K\}$ where K is number of different box types. Each action represents selecting a box to open

Outcome Space: Each pull outputs a card $c \in \{1, ..., N\}$ where N is number of unique cards, represented as one-hot vector $\mathbf{e}_c \in \{0, 1\}^N$

Unknown Distributions: Each arm k has unknown categorical distribution θ_k. When pulling arm k, card c is drawn with probability $\theta_{k,c}$

$\sum_{c=1}^{N} \theta_{k,c} = 1$ for each arm k

Reward Function: Each card c has value v_c. Reward for pulling arm k and receiving card c is $r = v_c$. The goal is to minimize cumulative regret over T pulls:

$R(T) = \sum_{t=1}^{T}(v^* - v_{c_t})$

where: v^* is expected value of optimal arm: $v^* = \max_k \sum_c \theta_{k,c} v_c$; c_t is card received at time t

1.2 Phase 2: Minimize regret between difference boxes within constraint

The budget determines the time horizon, as there are a limited number of boxes that can be bought before running out of funds. In order to quantify the regret, we will try to maximize the value of the cards drawn during the time horizon where the "optimal pack" is simply the pack with the highest

expected market value. In later phases, we will consider optimization for more nuanced objective functions based on trying to build a desired hand rather than just maximizing the market value.

This is a constrained bandit problem with categorical outcomes and fixed budget. Environment settings are similar to above except Action a_t only valid if remaining budget $b_t \geq c_k$.

1.3 Phase 3: Maximize probability in completing a specific deck of cards, set budget

Introduce selling individual cards at market value, which allows the agent to extend the time horizon of the model by selling the cards for more card pulls. This marks a change from an MAB problem towards an MDP problem, as the current budget and cards drawn so far are used to make decisions. The state space encodes the cards that have been pulled so far and the remaining budget for buying cards. The action space consists of the arms to pull from Phases 1-2, and additionally the option to sell cards to a merchant to recoup money for more card packs.

Actions that result in acquiring needed cards could generate rewards, while actions that don't result in acquiring needed cards would generate either no reward or smaller rewards if they give high-value cards that can be sold to buy more card packs. The distributions of cards granted by the different packs are encoded in transition probabilities of the MDP. The state space consists of all possible combinations of cards, which is quite high-dimensional, so we will look into different policy gradient methods for optimizing the policy. An additional ability to purchase cards at market value price is being considered, which would allow guaranteed (but possibly overly costly) steps towards completing the specific deck of cards.

We can optimize card collection utility by applying Proximal Policy Optimization. Environment settings are:

State Space: $s_t = \mathbf{x}_t$, where $\mathbf{x}_t \in \mathbb{N}^N$ is current collection (count of each card) (N is total number of unique cards in possession)

Action Space:

- Buy box: $\mathcal{A}_{\text{buy}} = \{a_{\text{box}}(k) \mid k \in [K]\}$
- Sell card: $\mathcal{A}_{\text{sell}} = \{a_{\text{sell}}(i) \mid i \in [N]\}$
- Direct purchase: $\mathcal{A}_{\text{purchase}} = \{a_{\text{purchase}}(i) \mid i \in [N]\}$

Reward Function:

$$R(s_t, a_t, s_{t+1}) = \begin{cases} R_{\text{complete}} & \text{if collection completed} \\ R_{\text{progress}} \cdot \Delta\text{completion} + R_{\text{value}} \cdot \text{value}(c_t) & \text{if closer to target} \\ R_{\text{value}} \cdot \text{value}(c_t) & \text{otherwise} \end{cases}$$

where

- Δcompletion measures progress toward target collection
- value(c_t) is associated market value of acquired card

2 Timeline

11/16-11/18: Proposal Complete
11/19-11/27: Phase 1 & 2 Complete
11/28-12/05: Phase 3 Complete
12/06-12/09: Project Tuning or Phase 4/5

3 Why this Project is Compelling

This project is compelling due to its abundance of state-independent and state-dependent aspects, both in the form of pulling card packs purely subject to probability (MAB) and then finding an objective within the context to work towards, dependent on the current state (MDP). It would allow us to apply and personalize algorithms learned in class such as the Bernoulli Bandit and Proximal Policy Optimization methods to approach questions and objectives made within a novel environment on trading card collection.

4 Why this Works as a Project

4.1 (Stretch goal) Phase 4: Fluctuating Market Values of Individual Cards

Revisit minimization of box set regret, allowing the values of cards within a box to become higher or lower valued. Opportunity to explore restless bandit problems, as introduced in the project page.

4.2 (Stretch goal) Phase 5: Introduce Fluctuating Merchant Trades

Introduction of a merchant with fluctuating willingness to trade up or trade down from a card the user needs. Include a time step cost to ensure some degree of risk from a failed trade with the merchant (bus fare to visit shop)