A competition between travelling salesmen

Cuong Nguyen & Daniel Dinjian

1 Project Overview

For our project, we will be teaching an agent how to play the popular board game ”Ticket To Ride”. This game consists of a problem that is very similar to the Travelling Salesmen Problem.
Travelling Salesmen Problem (TSP):
”Given a list of cities and the distances between each pair of cities, what is the shortest possible route that visits each city and returns to the origin city?”

Ticket To Ride Problem (TTRP): ”Given a list of cities and the distances between each pair of cities, what is the most rewarding possible train route that visits each city, while the limiting number of available train routes are shared among competing agents? ”

Comparing TSP to TTRP, we have made 3 important modifications:

1. ”shortest possible route” is replaced with ”most rewarding possible train route”
2. ”returns to the origin city” is removed
3. ”routes are shared among competing agents” is added

Before we dive into the problem, we must describe the game play of Ticket to Ride (TTR).

1.1 TTR Gampelay in Reinforcement Learning (RL) Lingo

1.1.1 Game Objective & Environment

At the beginning of the game, each agent (competing player) has to choose 2-3 secret travelling tasks. Each task consists of building railroad tracks between 2 cities within the United States. All traveling tasks are unique, so there are no duplicates. During game play, 2-5 agents will try to complete their individual secret tasks which may have overlapping routes. As a result, each agent must optimize their planning in order to complete its travelling tasks while antagonizing agents will do the same.

Agents’ conflicts of interest are:

- Limited cleared land for building railroad tracks between cities

- Limited resources (metal tracks)

Reward-based Objective

The objective of the game is to collect as many victory points as possible. Trivially, reward can be equivalent to victory points because maximizing victory points is the same as winning thus meeting the objective. Rewards are given to each agent in the following cases:

1. **Lump Sum Rewards**: An agent successfully connects its tasked sets of cities with railroad tracks. Each task has some predetermined positive reward value (+5 to +23) based on the difficulty/lengthiness of connecting the 2 cities.
2. **Sub Rewards**: An agent connects any 2 cities with railroad tracks, regardless if the 2 cities are tasked or not. The reward value is based on the number of tracks required to connect those 2 cities. See table 1.
3. **Longest Road Reward**: The agent with the longest connected railroad tracks receives an extra 10 victory points at the end of the game.
4. **Negative Reward (Penalty)**: For any incomplete tasks, the agent receives a negative value in reward (-5 to -23), rather than the aforementioned positive value.

Track Length	Sub Reward
1	+1
2	+2
3	+4
4	+7
5	+10
6	+15

Table 1: The sub reward granted when an agent connects any 2 cities with railroad tracks of varying track lengths

1.1.2 Action Space

The game proceeds in a round-robin manner where each agent takes a turn in a predetermined order. During each agent's turn, it has the option to do 1 of the following 3 actions:

- **Collecting tracks ("Resource/Track Cards")**: To prepare for the building costs of connecting tracks, an agent may use its turn to collect resource/track cards.

- **Expending collected tracks to build**: To build on a track between any 2 cities, an agent must use (and discard) the required track cards from their hand.
- **Taking on more tasks**: An agent may look at 3 additional tasks and is required keep 1-3 tasks. This is an action that requires the balancing of penalty and reward associated with each additional task.

1.1.3 State Space

To inform its decision making and planning, an agent can consider its state which will consist of some combination of the following features:

- **Available build routes**: An agent can only build on a route based on 2 criteria. Criteria one, no one has ever claimed the route before. Criteria two, the agent has enough track cards to match the specified expense of the route.
- **Available resources Obtainable**: Within the deck of track cards, there are 6 railroad track types that exists in limited numbers. Within a turn, an agent can either draw from the top of the face down deck, or pick up from the 5 face up replenish-able community pile, or do both***. The strategic trade-offs between drawing from the deck or community pile lays within the realm of secrecy and draw probabilities.
- **Available resources Owned**: At any time step, an agent must optimize its use of in-hand resources given the current state. Given the specified build expenses to certain routes and competition for claiming routes, it may be more advantageous to use an in-hand resource sooner rather than later, or vice-versa.
- **Antagonist agents' previous actions**: Observing the history of competing agents such as what cards they've collected or where they've been building can inform the protagonist agent's update of action-state values.

1.2 Project Aims

TTR poses the interesting problem that blends TSP within a competition where the necessary resources are limited. Our aim is to compare and contrast RL agents who differs in how they prioritize their sub-goals (options/strategy) that contribute to the overall selected travelling tasks.

- **Aim 1**: Show transfer learning's jumpstart properties for agents that were trained by a curriculum.[2][3][5]
- **Aim 2**: Explore different state-action trajectories for learning in multi-agent environments with varying amounts of cooperation.

- **Aim 3**: Show how different feature and state representations across adversarial agents affect their game play performance.

2 Background and Related Work

In a multi-agent environment, an agent can either behave in a cooperative or competitive manner. In a competitive setting such as TTR, it may seem intuitive for an agent to behave competitively and act independently. However, Austerweil et al.[1] has shown that individual agents can increase their reward objectives when their learning is done cooperatively. The results suggests that a cooperative learner will outperform an independent learner when given the same target task.

Similarly, Yang et al.[5] have trained multiple driving agents to cooperate in SUMO simulated traffic environment. Their goal is to teach cooperation to autonomous cars to reduce traffic and the number of accidental collisions. In a general sense, their goal is to explore multi-agent cooperation for a group of agents with different individual goals. This is highly relevant to the proposed TTRP because all TTR agents are trying to complete individual goals in an environment with limited resources. Although an agent should not be cooperative during game play, an agent may greatly benefit from learning a cooperative multi-agent curriculum.

Given the complexities of the TTR tasks, the learning curve for the TTR task may be too steep for novice agents to overcome. As a result, convergence upon a beneficial strategic policy may require overly-extensive computations. In response, we will introduce a curriculum of sub tasks to aid novice agents in their learning of the game objective. Our approach will be similar to the approaches shown by Narvekar et al.[3].

3 Problem Formulation and Technical Approach

We wish to train an agent that can optimally answer the TTR question of most valuable path. However in any instance of playing the game, your ideal actions could conflict with opponents' actions. There are many successful strategies to the game that may or may not overlap. Maybe one strategy is to make the longest train that combines all of your tasks, but maybe another strategy is just to build tracks that interfere with your opponents' tasks. Maybe a third strategy is to ignore your travelling tasks and just collect sub rewards. There are multiple options to how you'll address game-play and it's not immediately obvious which is best.

Furthermore, maybe you want to build along a certain path and so you need certain resource cards. To what extent should you consider this need in your actions? Should you draw cards in hopes of covering the expenses of that action?

Should you take your next-most optimal action because your opponents' next move may change the environment? Is it worth spending all of your highly valuable rainbow cards to make that action possible now? Should you consider the probability of drawing a desired resource card, given the resource cards that you've previously seen in play?

There's a trade-off between storage complexity and speed complexity of your state space. To explore this we'll use different state-representations for various agents. Maybe the game's objective could be learned with tabulated state representations, or maybe not.

We will also consider the applications of curriculum learning to ease the entry of our agents into this relatively complex environment. There is a worry that starting naively, agents would be incapable of making objective-based decisions for the episode to terminate frequently enough. Therefore in addition to comparing various state-representations we'll also use various training curricula to give agents a vague understanding of how to maximize reward and score points.

Our curricula will primarily consist of learning three sub-goals.

1. Individually drawing cards and building train tracks to gain the sub rewards
    - Agent will be alone taking turn after turn, trying to maximize average points per turn.
    - Agents will learn to make educated decisions about when and where to build trains.
    - Agents will learn when and where to draw cards from.
    - Agents will learn how to connect routes across multiple cities to score longest road.
2. Accounting for additional opponents
    - Agents will compete in this sub game without their secret traveling tasks
    - Agents will learn to consider their opponent's potential moves when drawing resource cards or building tracks.
    - Agents may learn or be encouraged to sabotage each other.
    - Losing agents will have their end game rewards stripped away and penalized by point differentials
3. Completing travelling tasks to gain the Lump Sum Reward
    - Agent will be placed in situation where game has almost ended and agent has a variable number of turns left to complete objective travelling tasks.
    - Agent will receive reward or penalty based on the completion of their assinged tasks.

- Environment may end prematurely to ensure urgency in agent's strategic approach.

After completing the curriculum, the agent is considered "educated" and we expect educated agents to make wise decisions in the real environment among drawing cards, building tracks, completing travelling tasks, and mutually competing with other agents for victory.

We now have three questions to ask about our agent, as specified in our aims. Are they educated? What is their state representation? And have they coordinated their learning with other agents or learned individually? This third question refers to how multiple agents are learning different policies, but as programmers we have access to all of the diverse policies. We can combine or ensemble policies to leverage the experiences of adversarial/cooperative agents as they train, and more [4].

Our agents will have state representations built from some combination of the options in **1.1.3** and for each state representation we will train multiple adversarial models, some first through the curriculum and others kept naive. We will experiment with sharing experiences and policies between agents of the same state-space to create even higher performance agents. All agents will use an n-step Sarsa policy with n briefly tuned as a hyper parameter. We intend to tune this n parameter based on cumulative rewards across multiple episodes.

4 Evaluation and Expected Outcomes

We will compare learning rate of educated and naive agents as well as according to their state representation. We will also consider how mutual cooperation affects learning rate and how policy sharing/ensembling affects performance. We will measure trained model success according to average score, win-rate, and performance against human players. To explore the importance of curriculum ordering, we will compare asymptotic performances of curricula with differently ordered sub-goals. Badly ordered curricula are likely to confuse agents and hinder their asymptotic performance, while properly ordered curricula should help.

In the hopes of building a successful curriculum, we expect to see educated agents outperform uneducated agents. As suggested by previous work, we will demonstrate how cooperative learners will have better game play performance than independent learners while trained within the same number of episodes. We expect to show that certain state representations will be more effective regardless of the curriculum formats. Although, it may be interesting to see if some state representations are more compatible with cooperation than others.

References

[1] Joseph L Austerweil, Stephen Brawner, Amy Greenwald, Elizabeth Hilliard, Mark Ho, Michael L Littman, James MacGlashan, and Carl Trimbach. How

other-regarding preferences can promote cooperation in non-zero-sum grid games. In *Proceedings of the AAAI Symposium on Challenges and Opportunities in Multiagent Learning for the Real World*, 2016.

[2] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48. ACM, 2009.

[3] Sanmit Narvekar, Jivko Sinapov, Matteo Leonetti, and Peter Stone. Source task creation for curriculum learning. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*, pages 566–574. International Foundation for Autonomous Agents and Multiagent Systems, 2016.

[4] Ming Tan. Multi-agent reinforcement learning: Independent vs. cooperative agents. In *In Proceedings of the Tenth International Conference on Machine Learning*, pages 330–337. Morgan Kaufmann, 1993.

[5] Jiachen Yang, Alireza Nakhaei, David Isele, Hongyuan Zha, and Kikuo Fujimura. Cm3: Cooperative multi-goal multi-stage multi-agent reinforcement learning. *arXiv preprint arXiv:1809.05188*, 2018.