

# Todo list

8 - 10 pages of introduction . . . . .	2
discuss policy . . . . .	2
4 - 5 pages . . . . .	3
(5 pages) . . . . .	8
also describes policy $\pi$ . . . . .	8
address this is the most common formulation and how different libraries implement the interface . . . . .	8
address OpenSpiel's design multiple agents . . . . .	9
address POMDP . . . . .	9
address OpenSpiel's design of random node . . . . .	9
barely seen in the literature, need more literature review . . . . .	9
(5 pages) . . . . .	9
elaborate formally . . . . .	9
pure functions are efficient . . . . .	9
(20 - 25 pages) . . . . .	9
add one or two more examples of using pure functions, also mention how tape in the MooZi is different from the tape in GII . . . . .	10
Data needed; I've seen multiple issues on GitHub complaining about the training speed. I once assigned the task of "actually running the project and gather run time data" to Jiuqi but no follow up yet. . . . .	10
include data from previous presentation to make the point . . . . .	10
boardgames are fine, Atari games are slow, but in either cases we can't control MCTS inference batching is slow due to IO overhead, include data here from previous presentation to make the point . . . . .	10
explain driver pseudo-code . . . . .	12
(20 pages) . . . . .	13
(3 pages) . . . . .	13
(1 page) . . . . .	13

# Contents

<b>1 Introduction</b>	<b>2</b>
1.1 Contribution . . . . .	3
<b>2 Literature Review</b>	<b>3</b>
2.1 Planning and Search . . . . .	3
2.1.1 Introduction . . . . .	3
2.2 Monte Carlo Methods . . . . .	4
2.3 Monte-Carlo Tree Search (MCTS) . . . . .	4
2.3.1 Selection . . . . .	5
2.3.2 Expansion . . . . .	5
2.3.3 Evaluation . . . . .	6
2.3.4 Backpropagation . . . . .	6
2.4 AlphaGo, AlphaGo Zero, and Alpha Zero . . . . .	6
2.5 MuZero . . . . .	7

<b>3</b>	<b>Problem Definition</b>	<b>8</b>
3.1	Agent-Environment Interface and Markov Decision Process . . . . .	8
3.2	Shortcomings of the Agent-Environment Interface for General Game Playing . . . . .	9
3.2.1	Multi-Agent Games . . . . .	9
3.2.2	Partial Observability . . . . .	9
3.2.3	Environment Stochasticity . . . . .	9
3.2.4	Episodic vs Continuous . . . . .	9
3.2.5	Self-Observability . . . . .	9
3.2.6	Environment Output Structure . . . . .	9
3.3	Our Approach . . . . .	9
3.3.1	Generalized Interaction Interface . . . . .	9
3.3.2	Advantages . . . . .	9
<b>4</b>	<b>Method</b>	<b>9</b>
4.1	Design Philosophy . . . . .	9
4.1.1	Use of Generalized Interaction Interface . . . . .	9
4.1.2	Use of Pure Functions . . . . .	9
4.1.3	Being User Friendly . . . . .	10
4.1.4	Training Efficiency . . . . .	10
4.2	Structure Overview . . . . .	10
4.2.1	Driver . . . . .	10
4.2.2	Environment Adaptors . . . . .	11
4.2.3	Rollout Workers . . . . .	11
4.2.4	Replay Buffer . . . . .	12
4.2.5	Parameter Optimizer . . . . .	12
4.2.6	Distributed Training . . . . .	12
4.2.7	Monte-Carlo Tree Search . . . . .	13
4.2.8	Logging and Visualization . . . . .	13
<b>5</b>	<b>Experiments</b>	<b>13</b>
<b>6</b>	<b>Conclusion</b>	<b>13</b>
6.1	Future Work . . . . .	13

# 1 Introduction

**Deep Learning (DL)** is a branch of **Artificial Intelligence (AI)** that emphasizes the use of neural networks to fit the inputs and outputs of a dataset. The training of a neural network is done by computing the gradients of the loss function with respect to the weights and biases of the network. A better trained neural network can better approximate the function that maps the inputs to the outputs of the dataset.

8 - 10 pages of introduction

**Reinforcement Learning (RL)** is a branch of AI that emphasizes on solving problems through trials and errors with delayed rewards. RL had most success in the domain of **Game Playing**: making agents that could play boardgames, Atari games, or other types of games. An extension to Game Playing is **General Game Playing (GGP)**, with the goal of designing agents that could play any type of game without having much prior knowledge of the games.

discuss policy

**Deep Reinforcement Learning (DRL)** is a rising branch that combines DL and RL techniques to solve problems. In a DRL system, RL usually defines the backbone structure of the algorithm especially the Agent-Environment interface. On the other hand, DL is responsible for approximating specific functions by using the generated data.

**Planning** refers to any computational process that analyzes a sequence of generated actions and their consequences in the environment. In the RL notation, planning specifically means the use of a model to improve a policy.

A **Distributed System** is a computer system that uses multiple processes with various purposes to complete tasks.

## 1.1 Contribution

In this thesis we will describe a framework for solving the problem of GGP. We first define a interaction interface that's similar to the Agent-Environment Interface established in the current literature. We will also detail **MooZi**, a system that implements the GGP framework and the **MuZero** algorithm for playing both boardgames and Atari games.

# 2 Literature Review

4 - 5 pages

## 2.1 Planning and Search

### 2.1.1 Introduction

Most, if not all, AI problems can be reduced to a search problem [10, p.39]. Such search problems could be solved by determining the best plan, path, model, function, and so on, based some metrics of interest. Therefore, search has been playing a vital role in AI research since its dawn.

The term **planning** and **search** are widely used across different domains, especially in AI, and sometimes interchangeable. Here we adopt the definition by Sutton and Barto [9].

**Planning** refers to any process by which the agent updates the action selection policy  $\pi(a | s)$  or the value function  $v_\pi(s)$ . We will focus on the case of improving the policy in our discussion. We could view the planning process as an operator  $\mathcal{I}$  that takes the policy as input and outputs an improved policy  $\mathcal{I}\pi$ .

Planning methods could be categorized into types based on the focus of the target state  $s$  to improve. If the method improves the policy for arbitrary states, we call **background planning**. That is, for any timestep  $t$ :

$$\pi(a | s) \leftarrow \mathcal{I}\pi(a | s), \quad s \in \mathcal{S}$$

Typical background planning methods include dynamic programming and Dyna-Q. In the case of dynamic programming, a full sweep of the state space is performed and all states are updated. In the case of Dyna, an random subset of the state space is selected and all states in the subset are updated.

The other type of planning focuses on improving the policy of the current state  $S_t$  instead of any state. We call this **decision-time planning**. That is, for any

timestep  $t$ :

$$\pi(a \mid s) \leftarrow \mathcal{I}\pi(a \mid s), \quad s = S_t$$

We could also use blend both types of planning together. Technically, we could say that algorithms like AlphaGo use both types of planning when they self-play for training. The decision-time part is straight-forward: a tree search is performed at the root node and updates the policy of the current state. At the same time, the neural network is trained on past experience and all states are updated. The updates of this background planning is applied when the planner pulls the latest weights of the neural networks.

Early AI research has been focused on the use of search as a planning method. In 1968, Hart, Nilsson, and Raphael designed the **A\*** algorithm for finding shortest path from a start vertex to a target vertex [5]. Although A\* works quite well for many problems, especially in early game AI, it falls short in cases where the assumptions of A\* do not hold. For example, A\* does not yield optimal solution under stochastic environments and it could be computationally infeasible on problems with high branching factors. More sophisticated search algorithms were developed to cater to the growing complexity of use cases.

In 1990, Korf framed the problem of **Real-Time Heuristic Search**, proposed the Real-Time-A\* algorithm, and pioneered the study of search algorithms with bounded computation [7].

Monte-Carlo techniques were adopted to handle environment stochasticity. Tree-based search algorithms such as **MiniMax** and **Alpha-Beta Pruning** were also designed to better play two-player games.

## 2.2 Monte Carlo Methods

In 1873, Joseph Jagger observed the bias in roulette wheels at Monte Carlo Casino. He studied the bias by recording results of roulette wheels and capitalized over 2 million francs over several days by betting on the most favorably biased wheel [8]. Therefore, **Monte Carlo (MC)** methods gained their name as a class of algorithms based on building expectation on random processes.

MC methods are used in many domains but in this thesis we will primarily focus on its usage in search. In a game where terminal states are usually unreachable by the limited search depth, evaluations has to be performed on the leaf nodes that represent intermidate game states. One way of obtaining an evaluation on a state is by applying a heuristic function. Heuristic functions used this way are usually hand-crafted by human based on expert knowledge, and hence are prone to human error. The other way of evaluating the state is to perform a rollout from that state to a terminal state by selecting actions randomly. This evaluation process is called **random rollout** or **Monte-Carlo rollout**.

## 2.3 Monte-Carlo Tree Search (MCTS)

Kocsis and Szepesvári developed the **Upper Confidence Bounds applied to Trees (UCT)** method as an extension of the Upper Confidence Bound (UCB) employed in bandits [6]. Rémi Coulom developed the general idea of **Monte-Carlo Tree Search** that combines both Monte-Carlo rollouts and tree search [3] for Go program CrazyStone. Shortly afterwards, Gelly et al. implemented MoGo that uses the UCT selection formula [4]. MCTS is then generalized by Chaslot et al. as a

framework for game AI [2]. This framework requires less domain knowledge than other classic approaches to game AI while also being competent in strength. The core idea of this framework is to gradually build the search tree by iteratively applying four steps: **selection**, **expansion**, **evaluation**, and **backpropagation**. The search tree built this way will have an emphasize on more promising moves and game states. This means these more promising states are visited more often, have more children, and have deeper estimates that are aggregated to a more accurate value.

Here we detail the four steps in the MCTS framework by Chaslot et al. (see Figure 1).

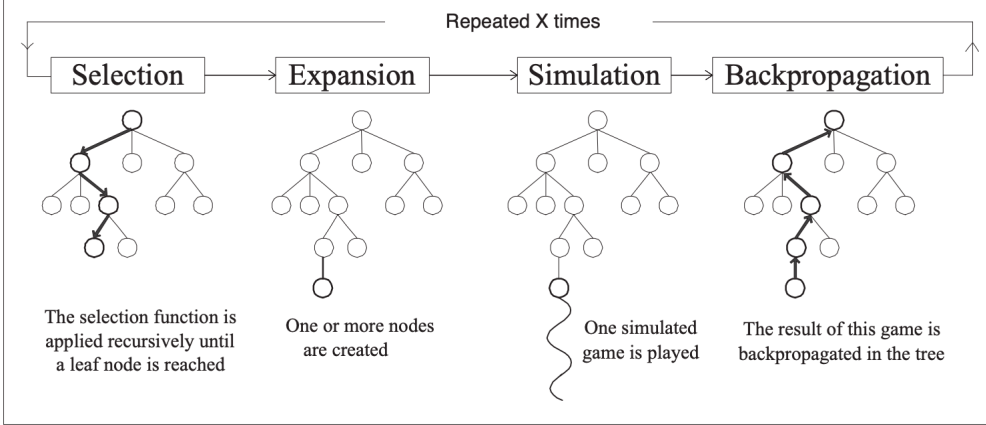


Figure 1: The Monte-Carlo Tree Search Framework

### 2.3.1 Selection

This selection process starts at the root node and repeats until a leaf node is reached. At each level of the tree, a child node is selected based on a selection formula. A selection formula usually has two parts, one part is based on the prior score  $x$  of the state, and the other part is the exploration bonus  $b$ . For a parent node  $p$  and children nodes  $C$ , the selection  $I$  is based on

$$I_p = \operatorname{argmax}_{c \in C} (x_c + b_c) \quad (1)$$

The prior score could be based on the value of the child, the accumulated reward of the child, or the prior selection probability based on the policy  $\pi(c | p)$ . The exploration bonus is usually based on the visit count of the child and the parent. The more visits a child gets, the less the exploration bonus will be.

For example, Gelly et al. used this following formula for selection in their MoGo implementation:

$$I_p = \operatorname{argmax}_{c \in C} \left( \frac{v_c}{n_c} + \sqrt{\frac{2 * \log(n_b)}{n_c}} \right)$$

where  $v_c$  is the value of the node,  $n_b$  and  $n_c$  are the visit counts of the parent and child, respectively.

### 2.3.2 Expansion

The selected leaf node is then expanded by adding one or more children to the leaf node, each represents a succeeding game state of the selected leaf's state.

### 2.3.3 Evaluation

Each of the expanded nodes is now evaluated by using one of these methods:

- A game is played until the game ends using a random policy.
- A game is played until the game ends using a rollout policy.
- A evaluation function is applied to the node. This evaluation function could be either a hand-crafted heuristic or learned by a neural network

### 2.3.4 Backpropagation

After the expanded nodes are evaluated, the traversed nodes to reach the expanded nodes are updated. The statistics updated usually include visit count, estimated value and accumulated reward of the nodes.

These four steps are repeated until the budget runs out. After this search is performed, the agent acts by selecting the action associated with the most promising child of the root node. This could be the most visited child, the child with the greatest value, or the child with the highest lower bound.

## 2.4 AlphaGo, AlphaGo Zero, and Alpha Zero

**AlphaGo** is the first Go program that beats a human Go champion by 5 games to 0. The AlphaGo is trained with a machine learning pipeline with multiple stages. For the first stage of training, a supervised learning policy (or SL policy) is trained to predict expert moves. This SL policy  $p$  is parametrized by weights  $\sigma$ , denoted  $p_\sigma$ . The input of the policy network is a representation of the board state, denoted  $s$ . Given a state  $s$  as the input, this network outputs a probability distribution over all legal moves  $a$  through the last softmax layer. During the training of the network, randomly sampled expert moves are used as training targets. The weights  $\sigma$  are then updated through gradient ascend to maximize the probability of matching the human expert move:

$$\Delta\sigma \propto \frac{\partial \log p_\sigma(a | s)}{\partial \sigma}$$

For the second stage of training, the supervised policy  $p_\sigma$  is duplicated further trained with reinforcement learning. This reinforcement learning trained policy (or RL policy) is parametrized by weights  $\rho$  which is initialized to the same values as  $\sigma$ . In other words, the RL policy is initialized so that  $p_\rho = p_\sigma$ . The training data is then generated through self-play using  $p_\rho$  as the rollout policy. For each game, the game outcome  $z_t = \pm r(s_T)$ , where  $s_T$  is the terminal state, +1 for winning, -1 for losing from the perspective of the current player. Weights  $\rho$  is then updated using gradient ascend to maximize the expected outcome using the update formula:

$$\Delta\rho \propto \frac{\partial \log p_\rho(a_t | s_t)}{\partial \rho} z_t$$

For the last stage, a value function is trained to evaluate board positions. This value function is modeled with a neural network with weights  $\theta$ , denoted  $v_\theta$ . Given a state  $s$ ,  $v_\theta(s)$  predicts the outcome of the game as if both players act according to the policy  $p_\rho$ . The neural network is trained with stochastic gradient descent to

minimize the mean squared error (MSE) between the predicted value  $v_\theta(s)$  and the outcome  $z$ .

$$\Delta\theta \propto \frac{\partial v_\theta(s)}{\partial\theta} (z - v_\theta(s))$$

AlphaGo combines the policy network  $p_\rho$  and the value network  $v_\theta$  with MCTS for acting. AlphaGo uses a MCTS variant similar to that described in 2.3. In the search tree, each edge  $(s, a)$  stores an action value  $Q(s, a)$  a visit count  $N(s, a)$ , and a prior probability  $P(s, a)$ . At each time step, the search starts at the root node and simulates until the budget runs out. In the select phase of each simulation phase, an action is selected for each traversed node using the same base formula (1). In AlphaGo, the base score of the selection formula is the estimated value of the next states after taking the actions, namely  $Q(s, a)$ . The exploration bonus of edge  $(s, a)$  is based on the prior probability and decays as visit count grows.

$$u(s, a) \propto \frac{P(s, a)}{1 + N(s, a)} \quad (2)$$

The action taken at time  $t$  is the action that maximizes the sum of the base score and the exploration bonus

$$a_t = \underset{a}{\operatorname{argmax}} (Q(s_t, a) + u(s_t, a)) \quad (3)$$

**AlphaGo Zero** is a successor of AlphaGo with the main difference of not learning from human.

**Alpha Zero** reduces game specific knowledge of AlphaGo Zero so that the same algorithm could be also applied to Shogi and chess.

## 2.5 MuZero

**MuZero** is different from the Alpha Zero family where the model used for planning is learned instead of given. This difference further reduces game specific knowledge required for the algorithm. More specifically, MuZero no longer requires the access to a perfect model of the target game. Instead, a neural network that learns from the game experience is used in the search to replace the perfect model.

Central to MuZero’s design are three functions trained for acting and planing. The **representation function**  $h$  encodes a history of observations  $o_1, o_2, \dots, o_t$  into a hidden state  $s_t$ . The **dynamics function**  $g$ , given a hidden state  $s^k$  and action  $a^k$ , produces an immediate reward  $r^k$  and the next hidden state  $s^{k+1}$ . The **prediction function**  $f$ , given a hidden state  $s^k$ , produces a probability distribution  $p^k$  of actions and a value associated to that hidden state  $v^k$ .

MuZero plans with a search method based on the MCTS framework (discussed in 2.3). Due to the lack of access to a perfect model, MuZero’s MCTS differs from a standard one in numerous ways. The nodes are no longer perfect representations of the board states. Instead, each node is associated with a hidden state  $s$  as a learned representation of the board state. The transition is no longer made by the perfect model but the dynamics function. There are technically no terminal states in the tree as the learned model does not predict game termination. For timestep  $t$ , the search starts by creating the root node of the search tree with a history of observations using the representation function

$$s^0 \leftarrow h_\theta(o_0, o_1, o_2, \dots, o_t)$$

Then an action is selected base on a variant of formula 1

$$a^k = \arg \max_a \left\{ Q(s, a) + P(s, a) \frac{\sqrt{\sum_b N(s, b)}}{1 + N(s, a)} \left[ c_1 + \log \left( \frac{\sum_b N(s, b) + c_2 + 1}{c_2} \right) \right] \right\} \quad (4)$$

Where the base score of an edge is the  $Q$  value, and the exploration bonus is computed based on the prior probability of the edge and the visit counts of both sides of the edge.

### 3 Problem Definition

(5 pages)

#### 3.1 Agent-Environment Interface and Markov Decision Process

Traditionally, RL problems and solutions are frame with the **Agent-Environment Interface** (Figure 2).

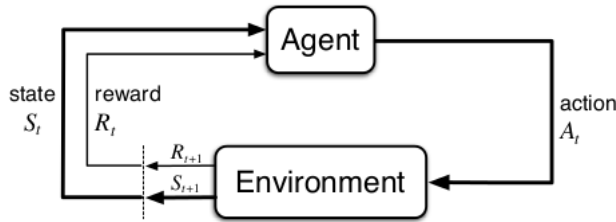


Figure 2: Agent-Environment Interface

The decision maker in this interface is called the **Agent**, and the agent interacts with the **Environment** continuously. A problem implements the interface could be represented as a **Markov Decision Process (MDP)**, which is tuple of four elements where:

- $\mathcal{S}$ , a set of states that forms the *state space*
- $\mathcal{A}$ , a set of actions that forms the *action space*
- $P(s, a, s') = Pr[S_{t+1} = s' | S_t = s, A_t = a]$ , the transition probability function
- $R(s, a, s')$  the reward function

At each time step  $t$ , the agent starts at state  $S_t \in \mathcal{S}$ , picks an action  $A_t \in \mathcal{A}$  transitions to state  $S_{t+1} \in \mathcal{S}$  based on the probability function  $P(S_{t+1} | S_t, A_t)$ , and receives a reward  $R(S_t, A_t, S_{t+1})$ .

The agent interacts with enviornment and generates a sequence of actions, states, and rewards:  $S_0, A_0, R_1, S_1, A_1, R_2, \dots$ . We call this sequence a **trajectory**. In the finite case, this interaction terminates until a terminal state is reached at time  $t = T$ , and this sequence is called an **episode**. In the infinite case, the interaction continues indefinitely. The goal of the agent in the problem is either to maximize the accumulative reward in the finite setting, or to maximize the average reward in the infinite case.

also describes policy  
 $\pi$

address this is  
the most common  
formulation and



## 3.2 Shortcomings of the Agent-Environment Interface for General Game Playing

### 3.2.1 Multi-Agent Games

address OpenSpiel's design multiple agents

### 3.2.2 Partial Observability

address POMDP

### 3.2.3 Environment Stochasticity

address OpenSpiel's design of random node

### 3.2.4 Episodic vs Continuous

### 3.2.5 Self-Observability

barely seen in the literature, need more literature review

### 3.2.6 Environment Output Structure

## 3.3 Our Approach

(5 pages)

### 3.3.1 Generalized Interaction Interface

We propose the **Generalized Interaction Interface (GII)**. We define the **tape**  $E$  as the data storage of the interface, and a **law**  $L$  as a pure function that operates on the tape. An instance of such interface could consists of exactly one tape and multiple laws, and we define such an instance a **universe**. A universe **ticks** by applying the laws on the tape.

elaborate formally

We implement a simplified version of this interface in **MooZi**.

### 3.3.2 Advantages

pure functions are efficient

## 4 Method

(20 - 25 pages)

### 4.1 Design Philosophy

#### 4.1.1 Use of Generalized Interaction Interface

One of the goals of the project is to demonstrate the use of Generalized Interaction Interface (GII). All modules in the project will be implemented to align with the interface. Third-party libraries that include game environments are wrapped with special wrappers that converts the outputs into the GII format.

#### 4.1.2 Use of Pure Functions

One of the most notable difference of MooZi implementation is the use of pure functions. In GII, **laws** are pure functions that read from and write to the **tape**. Agents implemented in Agent-Environment Interface usually do not separate the storage of data and the handling of data. In MooZi, we separate the storage of data and the handling of data whenever possible, especially for the parts with heavy

computations. For example, we use **JAX** and **Haiku** to implement neural network related modules. These libraries separate the **specification** and the **state** of a neural network. The **specification** of a neural network is a pure function that is internally represented by a fixed computation graph. The **parameters** of a neural network includes all variables that could be used with the specification to perform a forward pass.

4.1.3 Being User Friendly

4.1.4 Training Efficiency

One common problem with current open-sourced MuZero projects is their training efficiency. Even for simple environments, these projects could take hours to train.

There are a few major bottlenecks of training efficiency in this type of project. The first one is system parallelization.

The second one is the environment transition speed.

The third one is neural network inferences used in training.

4.2 Structure Overview

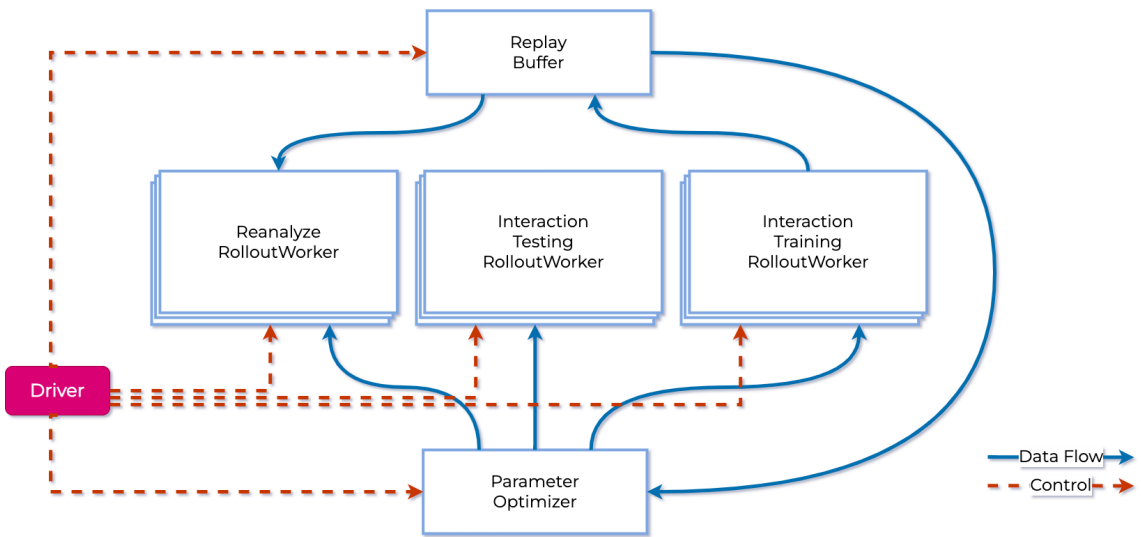


Figure 3: MooZi Architecture

4.2.1 Driver

In a distributed system with centralized control, a single process is responsible for operating all other processes. This central process is called the **driver**. Other processes are either **tasks** or **actors**. **Tasks** are stateless functions that takes inputs and return outputs. **Actors** are statefull objects that group several methods that take inputs and return outputs. In RL literature, **actor** is also a commonly used term to describe the process that stores a policy and interacts with an environment. Even though MooZi does not adopt the concept of a RL actor, we will use the term **ray task** and **ray actor** to avoid confusion. In contrast to distributed systems with distributed control, ray tasks and ray actors are reactive and do not have busy loops. The driver decides when a ray task or ray actor is activated and what data

add one or two more examples of using pure functions, also mention how tape in the MooZi is different from the tape in GII

Data needed; I've seen multiple issues on GitHub complaining about the training speed. I once assigned the task of "actually running the project and gather run time data" to Jiuqi but no follow up yet.

include data from previous presentation to make the point

boardgames are fine, Atari games are slow, but in either cases we can't control

MCTS inference batching is slow due to IO overhead, include data here from previous presentation to make the point

should be used as inputs and where the outputs should go. In other words, the driver process orchestrates the data and control flow of the entire system, and ray tasks and ray actors merely response to instructions.

#### 4.2.2 Environment Adaptors

Environment adaptors unify environments defined in different libraries into a unified interface. In the software engineering nomenclature, environment adaptors follow the adaptor design pattern [1].

More specifically, in our project we implement environment adaptors for three types of environments that are commonly used in RL research: (1) OpenAI Gym (2) OpenSpiel (3) MinAtar. The adaptors convert the inputs and outputs of these environments into forms that GII accepts.

The adaptors have the same signature as follows:

- Inputs

**is\_first:** A boolean signals the episode start.

**is\_last:** A boolean signals the episode end.

**action:** An integer indicates the last action taken by the agent. The valid range of the action is  $[0, \text{number\_of\_available\_actions})$ .

- Outputs

**obs:** An N-dimensional array that represents the observation of the current timestep.

**is\_first:** A boolean signals the episode start.

**is\_last:** A boolean signals the episode end.

**to\_play:** An integer indicates the next player to take a move.

**reward:** A float indicates the reward of taking the given action.

**legal\_actions\_mask:** A bit mask of legal action indicies.

#### 4.2.3 Rollout Workers

A **rollout worker** is a ray actor that:

- stores a collection of universes, including tapes and laws in the universes
- stores a copy of the neural network specification
- stores a copy of the neural network parameters
- optionally stores batching layers that enable efficient computation

A rollout worker does not inherently serve a specific purpose in the system and its behavior is mostly determined by the list of laws created with the universes.

There are three main patterns of rollout workers used in MooZi: **interaction training rollout worker**, **interaction testing rollout worker**, and **reanalyze rollout worker**.

#### 4.2.4 Replay Buffer

The replay buffer:

- stores trajectories generated by the rollout workers
- processes the trajectories into training targets
- stores processed training targets
- computes and updates priorities of training targets
- responsible for sampling and fetching batches of training targets

#### 4.2.5 Parameter Optimizer

The parameter optimizer:

- stores a copy of the neural network specification
- stores the latest copy of neural network parameters
- stores the loss function
- stores the training state
- computes forward and backward passes and updates the parameters

#### 4.2.6 Distributed Training

```
for epoch in range(num_epochs):
    for w in workers_env + workers_test + workers_reanalyze:
        w.set_params_and_state(param_opt.get_params_and_state())

    while traj_futures:
        traj, traj_futures = ray.wait(traj_futures)
        traj = traj[0]
        replay_buffer.add_trajs(traj)

    if epoch >= epoch_train_start:
        train_batch = replay_buffer.get_train_targets_batch(
            big_batch_size
        )
        param_opt.update(train_batch, batch_size)

    env_trajs = [w.run(num_ticks_per_epoch) for w in workers_env]
    reanalyze_trajs = [w.run() for w in workers_reanalyze]
    traj_futures = env_trajs + reanalyze_trajs

    if epoch % test_interval == 0:
        test_result = workers_test[0].run(120)

    for w in workers_reanalyze:
        reanalyze_input = replay_buffer.get_train_targets_batch(
            num_trajs_per_reanalyze_universe
            * num_universes_per_reanalyze_worker
        )
        w.set_inputs(reanalyze_input)
```

explain driver  
pseudo-code

#### 4.2.7 Monte-Carlo Tree Search

#### 4.2.8 Logging and Visualization

### 5 Experiments

(20 pages)

### 6 Conclusion

(3 pages)

#### 6.1 Future Work

(1 page)

### References

- [1] *Adapter Pattern*. In: *Wikipedia*. May 31, 2022. URL: [https://en.wikipedia.org/w/index.php?title=Adapter\\_pattern&oldid=1090799086](https://en.wikipedia.org/w/index.php?title=Adapter_pattern&oldid=1090799086) (visited on 06/09/2022).
- [2] Guillaume Chaslot et al. “Monte-Carlo Tree Search: A New Framework for Game AI”. In: (2008), p. 2.
- [3] Rémi Coulom. “Efficient Selectivity and Backup Operators in Monte-Carlo Tree Search”. In: *Computers and Games*. Ed. by H. Jaap van den Herik, Paolo Ciancarini, and H. H. L. M. Donkers. Red. by David Hutchison et al. Vol. 4630. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 72–83. ISBN: 978-3-540-75537-1 978-3-540-75538-8. DOI: 10.1007/978-3-540-75538-8\_7.
- [4] Sylvain Gelly et al. “Modification of UCT with Patterns in Monte-Carlo Go”. In: (2006). URL: <http://citeseerx.ist.psu.edu/viewdoc/citations?doi=10.1.1.96.7727> (visited on 06/03/2022).
- [5] Peter E. Hart, Nils J. Nilsson, and Bertram Raphael. “A Formal Basis for the Heuristic Determination of Minimum Cost Paths”. In: *IEEE Transactions on Systems Science and Cybernetics* 4.2 (July 1968), pp. 100–107. ISSN: 2168-2887. DOI: 10.1109/TSSC.1968.300136.
- [6] Levente Kocsis and Csaba Szepesvári. “Bandit Based Monte-Carlo Planning”. In: *Machine Learning: ECML 2006*. Ed. by Johannes Fürnkranz, Tobias Scheffer, and Myra Spiliopoulou. Red. by David Hutchison et al. Vol. 4212. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 282–293. ISBN: 978-3-540-45375-8 978-3-540-46056-5. URL: [http://link.springer.com/10.1007/11871842\\_29](http://link.springer.com/10.1007/11871842_29) (visited on 04/05/2021).
- [7] Richard E. Korf. “Real-Time Heuristic Search”. In: *Artificial Intelligence* 42.2 (Mar. 1, 1990), pp. 189–211. ISSN: 0004-3702. DOI: 10.1016/0004-3702(90)90054-4.
- [8] *Monte Carlo Casino*. In: *Wikipedia*. May 28, 2022. URL: [https://en.wikipedia.org/w/index.php?title=Monte\\_Carlo\\_Casino&oldid=1090263293](https://en.wikipedia.org/w/index.php?title=Monte_Carlo_Casino&oldid=1090263293) (visited on 06/03/2022).

- [9] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. Second edition. Adaptive Computation and Machine Learning Series. Cambridge, Massachusetts: The MIT Press, 2018. 526 pp. ISBN: 978-0-262-03924-6.
- [10] Georgios N. Yannakakis and Julian Togelius. *Artificial Intelligence and Games*. Cham: Springer International Publishing, 2018. ISBN: 978-3-319-63518-7 978-3-319-63519-4. DOI: 10.1007/978-3-319-63519-4.