Monte-Carlo Tree Search for Policy Optimization

Xiaobai Ma[1] , **Katherine Driggs-Campbell**[2] , **Zongzhang Zhang**[3] and **Mykel J. Kochenderfer**[1]
[1]Aeronautics and Astronautics Department, Stanford University
[2]Electrical and Computer Engineering Department, University of Illinois Urbana-Champaign
[3]National Key Laboratory for Novel Software Technology, Nanjing University
maxiaoba@stanford.com, krdc@illinois.edu, zhangzongzhang@gmail.com, mykel@stanford.edu

Abstract

Gradient-based methods are often used for policy optimization in deep reinforcement learning, despite being vulnerable to local optima and saddle points. Although gradient-free methods (e.g., genetic algorithms or evolution strategies) help mitigate these issues, poor initialization and local optima are still concerns in highly nonconvex spaces. This paper presents a method for policy optimization based on Monte-Carlo tree search and gradient-free optimization. Our method, called Monte-Carlo tree search for policy optimization (MCTSPO), provides a better exploration-exploitation trade-off through the use of the upper confidence bound heuristic. We demonstrate improved performance on reinforcement learning tasks with deceptive or sparse reward functions compared to popular gradient-based and deep genetic algorithm baselines.

1 Introduction

Recent advances in deep reinforcement learning (DRL) have shown success on a variety of tasks, including games [Mnih *et al.*, 2015] and robotics [Wu *et al.*, 2017]. Artificial neural networks (ANNs) are often used to represent either a state-action value function [Hessel *et al.*, 2018] or a policy network [Schulman *et al.*, 2015a; Schulman *et al.*, 2017], and they are typically optimized through gradient-based methods [Kingma and Ba, 2015]. While gradient-based optimization can be extremely efficient on tasks with smooth or incremental rewards, they often require reward shaping [Grzes and Kudenko, 2009]. Such solvers struggle in environments with sparse or deceptive rewards, indicating that issues such as local optima and saddle points are limiting their success [Such *et al.*, 2017].

Recently, a category of gradient-free algorithms produced policies that outperformed those trained using the typical gradient-based methods on certain tasks [Chrabaszcz *et al.*, 2018]. The success of these methods implied that gradient-free algorithms could also work on deep and complex neural networks. Evolution Strategies (ES) were used to evolve a policy network by adding noise to the neural network parameters and performing parallel rollouts [Salimans *et al.*, 2017]. The evolution direction, i.e., the parameter update vector, is determined using the accumulated rewards of perturbed trajectories. Such strategies are tolerant to delayed rewards and long horizons because they directly use the trajectory return, in contrast that gradient-based methods often need immediate rewards for better credit assignment.

ES is similar to a gradient approach using finite differences over trajectory returns. As a truly gradient-free method, the deep genetic algorithm (Deep GA) applies a simple genetic algorithm to evolve neural networks [Such *et al.*, 2017]. Deep GA maintains a population of neural network parameters. The top performing individuals are iteratively mutated by adding random noise to create the next generation.

The success of Deep GA on difficult DRL tasks suggests that, in some cases, following the gradient direction may be misleading, especially in highly nonconvex spaces. Other population-based methods incorporate and adapt gradient-based approaches in the mutation steps [Khadka and Tumer, 2018].While this combined approach improves data efficiency, it also loses some randomness in the mutation and, more importantly, lacks the scalability to large population sizes and neural networks with deep structure.

This paper formulates the policy optimization problem as a deterministic Markov decision process (MDP) and uses Monte-Carlo tree search (MCTS) to find the optimal trajectories in the policy space. Using MCTS on population based methods has been applied to games with discrete state and action spaces [Baier and Cowling, 2018], but has not been effectively generalized to high-dimensional continuous state and action spaces.

Using a similar mutation process and network parameter representation as Such *et al.* [2017], the proposed method, referred to as Monte-Carlo tree search for policy optimization (MCTSPO), is shown to be a more efficient tree search algorithm than Deep GA. MCTSPO improves the exploration-exploitation tradeoff by adopting the UCT principle in choosing parents [Kocsis and Szepesvári, 2006], i.e., the individual with the highest upper confidence bound gets expanded first.

MCTSPO is compared to a state-of-the-art gradient based algorithm, trust region policy optimization (TRPO) [Schulman *et al.*, 2015a], and a genetic algorithm with safe mutation [Lehman *et al.*, 2018] on multiple continuous control tasks. We demonstrate how our method increasingly outperforms these baselines as the difficulty of the tasks increases.

In summary, we present the following contributions:

1. We format the policy optimization task as an MDP, which enables the application of existing MDP solvers to policy optimization;
2. We introduce new modifications to the MCTS algorithm to increase its efficiency and scalability when solving the policy optimization MDP; and
3. We show that MCTSPO is able to outperform existing methods on a set of difficult reinforcement learning benchmark tasks.

This paper is organized as follows: Section 2 gives an introduction to MDPs and Deep GA. Section 3 describes how to formulate policy optimization as an MDP and presents MCTSPO. Section 4 introduces the experimental setup and implementation details. The results and the corresponding discussion are presented in Section 5. Section 6 summarizes the contributions of the work and future work.

2 Background

2.1 Markov Decision Process

A Markov decision process (MDP) is a common model for sequential decision making problems [Kochenderfer, 2015]. An MDP is defined as a tuple $M = (S, A, P, R, T, \gamma, \rho_0)$, where S is the state space; A is the action space; $P : S \times A \times S \to [0,1]$ defines the transition probability; $R : S \times A \times S \to \mathbb{R}$ is the reward function; T is the horizon; $\gamma \in (0,1]$ is the discount factor; and $\rho_0 : S \to [0,1]$ is the initial state distribution. A trajectory τ is a tuple of sequential transitions in the MDP: $\tau = ((s_0, a_0, r_0, s_1), (s_1, a_1, r_1, s_2), \ldots, (s_{T-1}, a_{T-1}, r_{T-1}, s_T))$. The return of the trajectory is given by $\rho(\tau) = \sum_{t=0}^{T-1} \gamma^t r_t$.

A policy $\pi : S \times A \to [0,1]$ is a mapping from each state in S to a probability distribution over the actions. Let $\eta(\pi)$ denote its expected discounted reward:

$$\eta(\pi) = \mathbb{E}_{s_0, a_0, \ldots}\Big[\sum_{t=0}^{T-1} \gamma^t R(s_t, a_t, s_{t+1})\Big], \tag{1}$$

where $s_0 \sim \rho_0(s), a_t \sim \pi(a_t \mid s_t)$, and $s_{t+1} \sim P(s_{t+1} \mid s_t, a_t)$.

2.2 Deep Genetic Algorithm

Deep GA evolves a population of neural network policies, which are candidate optimal policies. At each iteration, each individual policy in the population is deployed in the environment to receive a fitness score, i.e., the average return of the rollout policy. Then, Deep GA performs truncation selection to choose the top k individuals as the parents of the next generation. The individual in the next generation is generated by uniformly choosing one parent and performing a mutation.

The mutations are in the form of additive Gaussian noise on the parent's parameter vector: $\theta' = \theta + \sigma\varepsilon$, where θ and θ' are parameter vectors of the parent and the child's neural network respectively, σ is the step size, and ε is sampled from a standard multivariate Gaussian distribution. Typically, a crossover step is also performed before mutation to combine the parameters of two parents to generate an offspring. Deep GA skips this step, since there is no trivial way to efficiently combine the learned information of two neural networks. Due to the high non-linearity of the neural networks, switching parameters arbitrarily between two neural networks could destruct the encoded information from both sides.

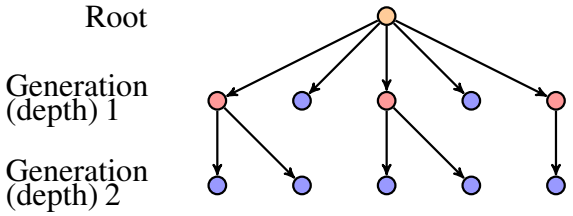

Figure 1: Tree structure in Deep GA. Each node represents an individual, and each arrow represents a mutation. Pink nodes are selected as parents for the next generation.

In a naive implementation, this method would not scale to large populations. Instead of explicitly storing parameter vectors of the entire population, Deep GA stores the series of random seeds that initialize and mutate the network. This idea is borrowed in MCTSPO and explained in Section 3.

3 Methods

3.1 Tree Structure Formulation

Since there is no crossover in Deep GA, each individual has a single parent from the previous generation, and may have several children in the next generation. If we add a dummy common parent to all individuals in the first generation, then we may represent the population evolution as a tree. In the tree, all individuals are represented by nodes and each edge represents a mutation. Each generation corresponds to a tree depth. Figure 1 shows the tree structure formed by a genetic algorithm with population size 5 and truncation size 3.

We observe that the genetic algorithm follows a simple tree search heuristic. At each depth, only a subset of nodes (the nodes selected as elites) is expanded and the rest of them are discarded. The number of nodes at each depth is fixed, determined by the population size. We introduce the idea of using Monte-Carlo tree search (MCTS) in the policy space, which balances exploration and exploitation using the upper confidence bound heuristic [Kocsis and Szepesvári, 2006].

3.2 Policy Optimization MDP

Formally, we consider the parameter optimization of a policy network as an MDP. There are two MDPs under consideration: one defined by the task environment (task MDP), and the other defined for policy optimization (policy optimization MDP). To differentiate, we use a tilde to denote elements related to the task MDP. Unless specified, all MDPs mentioned in the following text refer to the policy optimization MDP.

The state space of the MDP is the parameter space of the neural network. Each state is represented as the parameter vector of the network. The action of the MDP is the mutation applied to the network. Taking inspiration from Such *et al.* [2017] and Lehman *et al.* [2018], each action is represented as a vector consisting of a random seed and a magnitude, i.e., $a = [\text{seed}, v]^\top$.

The transition of the MDP is deterministic and is performed as follows. To create the first generation from the

root, we use the seed to initialize the neural network parameters with preferred weight initialization functions. For later generations, we first use the seed to randomly generate a direction vector $\hat{\delta}$ in the parameter space using the standard multivariate normal distribution. Then, the new state is given by $s_{t+1} = s_t + v\hat{\delta}$, where v is the mutation magnitude specified by the action. Thus, we may reproduce the parameter vector for any node in the tree by performing mutations specified by the action sequence connecting the root and the node. This representation avoids explicitly saving all nodes' parameters and makes the algorithm tractable.

The reward is defined as each mutation's performance improvement. For the tuple (s,a,s'), $R(s,a,s') = \tilde{\eta}(s') - \tilde{\eta}(s)$, where $\tilde{\eta}(s)$ is the expected return in the task environment produced by the policy specified by s. We assume $\tilde{\eta}(s_0) = 0$ for the root state. Assuming no discount, $\gamma = 1$, the return of a trajectory $\rho(\tau)$ equals the expected return of the last state, $\tilde{\eta}(s_T)$. Thus, we only need to rollout the policy represented by s_T in the task environment to get the trajectory return.

3.3 MCTS Methodology

We use MCTS to solve the MDP [Kocsis and Szepesvári, 2006]. The original algorithm is omitted here. This section focuses on introducing the important modifications to make MCTS work for policy optimization.

Since the action space is continuous, we use *progressive widening* to control the tree expansion [Chaslot *et al.*, 2008a]. We do not need to constrain the state widening since the state transition is deterministic. Similar to Couëtoux *et al.* [2011], we constrain the number of actions at each state node with $|N(s,a)| < kN(s)^{\alpha}$, where $N(s)$ is the number of times that s has been visited; $N(s,a)$ is the number of times that a has been chosen as the next action at state s; and $|N(s,a)|$ is the number of different actions tried at state s. Both positive parameters k and α are used to control the widening of the tree.

MCTS requires two meta policies: one for action selection and one for rollout. For action selection, we follow the upper confidence bound principle:

$$a' \leftarrow \arg\max_a Q(s,a) + c\sqrt{\frac{\log N(s)}{N(s,a)}}, \tag{2}$$

where c is an exploration parameter. The estimated return of choosing a at state s (in no discount and deterministic MDP) is $Q(s,a) = R(s,a,s') + V(s')$, where $V(s')$ is the estimated return at next state s'.

In the original MCTS algorithm, $V(s')$ is estimated by the *average* return collected by the subtree rooted at s'. Here, we use a variant of the MaxUCT estimation in the deterministic case, where $V(s')$ is the *maximum* return collected by the subtree [Keller and Helmert, 2013]. This estimation improves performance when the variance of returns among siblings may be very large. This large variance is common in policy space due to small differences and errors compounding during rollout, leading to diverging trajectories and returns.

For the rollout policy, a traditional approach is to apply random mutations to the current policy, and use the return of the final trajectory as an estimate of the value of the current policy, $V(s)$. This estimate is unbiased but has large variance. Thus, instead of random mutations, we employ a *no mutation* strategy. That is, in rollout, we just deploy the current policy in the environment and return the rewards. This approach requires less computation, and has lower variance.

Finally, we apply a technique called *safe mutation* to constrain the average performance divergence of a single mutation [Lehman *et al.*, 2018]. Due to the nonlinear nature of neural networks, adding random noise to neural network parameters can significantly and unpredictably influence its output. Constraining the average performance divergence ensures that each branch in the tree corresponds to a local region of the policy space, instead of distributing in the entire policy space randomly.

The performance divergence is defined as the expected mean square difference between the new and original policy outputs [Lehman *et al.*, 2018]. It is estimated by:

$$D(s,s_{old}) = \frac{1}{\tilde{T}}\sum_{t=1}^{\tilde{T}}\sum_{k=1}^{|\tilde{A}|}\Big[NN(\tilde{s_t};s)_k - NN(\tilde{s_t};s_{old})_k\Big]^2, \tag{3}$$

where s and s_{old} are the new and the original policy parameters. Recall $s = s_{old} + v\hat{\delta}$, where v and $\hat{\delta}$ are the mutation magnitude and direction. $NN(\tilde{s_t};s)_k$ is the k-th output of the policy network parameterized by s given input $\tilde{s_t}$. $\tilde{T}$ is the task environment's time horizon. $|\tilde{A}|$ is the dimension of the task environment's action space; and $\tilde{s_t}$, where $t \in \{1,2,\ldots,\tilde{T}\}$ is the sampled state at step t using the current policy.

Given $\hat{\delta}$, s, and the divergence constraint D_{max}, we solve the mutation magnitude efficiently using methods similar as in Schulman *et al.* [2015a]. The divergence D is approximated by $\hat{D}$ with:

$$\hat{D}(s,s_{old}) = \frac{1}{2}(s-s_{old})^{\top}U(s-s_{old}), \tag{4}$$

where $U_{ij} = \frac{\partial}{\partial s_i}\frac{\partial}{\partial s_j}D(s,s_{old})|_{s=s_{old}}$. Then $v = \sqrt{\frac{2D_{max}}{\hat{\delta}^{\top}U\hat{\delta}}}$, where $\hat{\delta}^{\top}U\hat{\delta}$ can be calculated efficiently as introduced in Schulman *et al.* [2015a]. Then, a line search is performed on v to ensure the satisfaction of the performance divergence constraint.

Calculating divergence needs rollouts in the task environment. Therefore, when expanding a node, we either must resample the trajectory or store the trajectory for each node when first reached. The first way is not sample efficient, since we deploy duplicate policies in the environment. The second suggestion, however, is inefficient in terms of memory.

To overcome this, we calculate feasible candidate actions when we first reach the node. When we want to expand a node, we randomly choose one from its candidate actions. Thus, we only need to store candidate actions for each node, which requires much less memory than storing the trajectory while avoiding duplicated sampling. For a node s, we add an additional buffer $CA(s)$ that stores the candidate actions. When a new state s is added to the tree, rollout is performed on s and a trajectory τ_s is collected from the task environment. We use τ_s to calculate n_{ca} candidate actions and store them in $CA(s)$, where n_{ca} is a predetermined constant. When we want to expand s with new actions, we simply pop an action from $CA(s)$. When $CA(s)$ is empty, we need to recollect a

Algorithm 1 MCTS for Policy Optimization (MCTSPO)

function MCTSPO(Task environment Γ, Initial state s_0)
  $Tree \leftarrow \emptyset$
  for $i \leftarrow 1$ **to** n_{itr}
    SIMULATE($Tree$, Γ, s_0)
  return $s^* \leftarrow \arg\max_{s \in Tree} \tilde{\rho}(\tau_s)$
function SIMULATE($Tree$, Γ, s)
  if $s \notin Tree$
    $Tree \leftarrow Tree \cup \{s\}$
    $(N(s), A(s), CA(s)) \leftarrow (0, \emptyset, \emptyset)$
    return ROLLOUT(Γ, s)
  $N(s) \leftarrow N(s) + 1$
  if $|N(s,a)| < kN(s)^{\alpha}$
    if $CA(s) = \emptyset$
      $\tau_s \leftarrow$ SAMPLE(Γ, s)
      $CA(s) \leftarrow$ GETCA(s, τ_s)
    $a \leftarrow$ POP($CA(s)$)
    $(N(s,a), Q(s,a)) \leftarrow (0, Q_0(s,a))$
    $A(s) \leftarrow A(s) \cup \{a\}$
  $a' \leftarrow \arg\max_{a \in A(s)} Q(s,a) + c\sqrt{\frac{\log N(s)}{N(s,a)}}$
  $N(s,a') \leftarrow N(s,a') + 1$
  $s' \leftarrow$ MUTATE(s, a')
  $q \leftarrow$ SIMULATE($Tree, \Gamma, s'$)
  if $q > Q(s,a')$
    $Q(s,a') \leftarrow q$
  return $Q(s,a')$

Algorithm 2 Rollout

function ROLLOUT(Γ, s)
  $\tau_s \leftarrow$ SAMPLE(Γ, s)
  $CA(s) \leftarrow$ GETCA(s, τ_s)
  return $\tilde{\rho}(\tau_s)$

Algorithm 3 Get candidate actions

function GETCA(s, τ_s)
  $CA \leftarrow \emptyset$
  for $i \leftarrow 1$ **to** n_{ca}
    $seed \leftarrow$ RANDSEED
    $v \leftarrow$ GETMAGNITUDE($s, \tau_s, seed$)
    $CA \leftarrow CA \cup \{[seed, v]^{\top}\}$
  return CA

trajectory from s and calculate a new set of $CA(s)$. Compared to sampling trajectories every time when expanding a state node, this increases the sample efficiency up to n_{ca} times.

In summary, based on the original MCTS, we apply progress widening on the actions. We use MaxUCT for value estimation. We deploy *no mutation* for rollout. We add performance divergence constraints on the mutation magnitude and use precalculated candidate actions to improve sample and memory efficiency. Algorithm 1 outlines the learning procedure for MCTSPO. Algorithms 2 and 3 show how to rollout and calculate candidate actions.

4 Experiments

We compare the performance of MCTSPO to two state-of-the-art baselines: TRPO as a representative of gradient-based methods [Schulman *et al.*, 2015a] and Deep GA using *safe mutation* [Lehman *et al.*, 2018]. The test environments are divided into two classes: classic control and Roboschool.

4.1 Classic Control

Three classic continuous control tasks, Acrobot [Geramifard *et al.*, 2015], Mountain Car [Moore, 1991], and Bipedal Walker [Brockman *et al.*, 2016], are tested with the same time horizon $\tilde{T} = 100$ and reward function. The agent receives a positive reward of 1.0 if the goal is reached, otherwise a small control penalty proportional to the action magnitude is applied. With this reward function, there is a trivial local optimum where zero control effort is applied at each step with zero reward. Some of the task environments' parameters are changed to increase the difficulty of the task:

- Acrobot: We increase the vertical position that needs to be reached for the end point of the robot arm, y_{goal}. The maximum reachable height is 2, meaning the two arms need to be strictly vertical. Here we set $y_{\text{goal}} = 1.999$.
- Mountain Car: We decrease the power of the car, p_{car}. A smaller p_{car} means a smaller acceleration limit in one time step, which makes reaching the flag harder. Here p_{car} is set to 0.0015.
- Bipedal Walker: We increase the horizontal length that the robot needs to move forward for success, x_{goal}. x_{goal} is set to 30 in our experiments.

4.2 Roboschool Robotics

To test the performance on high dimensional tasks, we adapt three robotics environments, Ant [Schulman *et al.*, 2015b], HalfCheetah [Wawrzynski, 2007], and Hopper [Murthy and Raibert, 1984], from OpenAI Roboschool [Schulman *et al.*, 2017]. The environment goal is to control the robot to walk forward for 2 m. The agent receives a positive reward of 1000 when the goal is achieved, otherwise it is penalized for control effort and invalid actions like feet collisions. The time horizon $\tilde{T}$ is set to 500.

4.3 Algorithm Implementation

For TRPO, we use the Gaussian multilayer perceptron architecture from RLLab [Duan *et al.*, 2016] with hidden layer sizes of 128, 64, and 32 with tanh activations. It is trained for 5000 iterations using step sizes 0.1 and 1.0. The batch size is set to 1000 for classic control tasks and 5000 for Roboschool.

For Deep GA, we use the deterministic multilayer perceptron architecture with the same network structure as used in TRPO. The population sizes are 100, 500, and 1000 with 500, 100, and 50 training iterations, respectively. The truncation size for parent selection is 20. At each iteration, the top three individuals persist to the next generation with no mutation, following a technique called *elitism* [Such *et al.*, 2017]. Since both the environment and the policy are deterministic, we use a batch size which is equal to the horizon, i.e., only one trajectory is sampled per rollout. The divergence constraint for the mutation step is set to 1.0 through preliminary tests.

For MCTSPO, we use the same architecture and batch size as in Deep GA. We use an exploration constant of $\sqrt{2}$ for classic control tasks and 10 for Roboschool.[1] The progressive widening parameters are set to $\alpha = k = 0.3$, 0.5, and 0.8, respectively.[2] The number of candidate actions is set to $n_{\text{ca}} = 4$ to balance the computation complexity and the sample efficiency. We train for 50,000 iterations with the same divergence constraint as used in Deep GA.

All three algorithms have the same number of environment calls in total for the training.[3]

5 Results

Each classic control task is tested for 20 trials and each Roboschool task is tested for 10 trials, using different random seeds. Performance is evaluated using the average return of the best trajectory found. Table 1 summarizes the results of different algorithms. To compare the performance and data efficiency of each algorithm, for each environment setting, we plot its training curve in fig. 2, showing the hyperparameter setting with the highest final return. The horizontal axis is the number of environment calls and the vertical axis is the average return of the best policy found so far. The shaded region of each curve is the error bound given by $\Delta/\sqrt{n_{\text{trial}}}$, where Δ is the standard deviation of the best returns.

In the classic control environments, MCTSPO outperforms Deep GA and TRPO. First, we note that TRPO almost gets stuck in local optima in all the experiment trials, which indicates the limitation of gradient based approaches on sparse and deceptive rewards. In Acrobot, MCTSPO is slightly better than Deep GA. In Mountain Car and Bipedal Walker, the advantage of MCTSPO is more evident. In Mountain Car, all MCTSPO trials converge to the optimal solution using approximately 4×10^6 environment calls fewer than Deep GA. In Bipedal Walker, MCTSPO achieves a much higher average return than Deep GA. In Roboschool, MCTSPO as has the best performance. It is able to reach the goal in most of the experiment trials except for the Half Cheetah, which is the hardest task attempted here.

We observe that Deep GA also suffers from local optima. This sub-optimal behavior happens when the individuals chosen as parents are all in the "valley" around the local optimum. Then, the descents are likely all distributed in the valley as well. Escaping the valley requires going against the direction of increasing return. Even if some individuals in the next generation evolve in the right direction, they are not likely to be chosen as the parents. Thus, the entire population will be stuck in the local optimum. This situation is often caused by a poor initialization.

Novelty search is helpful in this case [Such *et al.*, 2017], where rewards are replaced by the novelty of the policy behavior, which encourages exploration. However, this method requires a behavior characteristic, which is domain-specific. In addition, Deep GA with novelty search essentially sweeps over the entire policy space, thus preventing exploitation or focusing search efforts on more promising regions.

This problem is mitigated by MCTSPO. States or nodes with the same ancestor are in nearby regions of the policy space through constrained mutation. When a local optimum is found, the node around the local optimum is expanded more than other nodes. However, as more descendants are added to this node, the value estimation, $Q(s,a)$, associated with this node quickly reaches its limit, where $N(s,a)$ keeps increasing. As a result, the upper confidence bound estimation of s decreases, which prevents more descendants being added to s in the later iterations. The algorithm then expands nodes in other regions. Since each mutation path is stored in the search tree, there is a non-negligible chance to find a long enough path that escapes the local optimum.

As we decrease the difficulty of tasks, e.g. increase the power of the car in Mountain Car or decrease the target distance in Roboschool tests, the advantage of MCTSPO becomes less obvious. Actually when the environment goal could be easily achieved without much exploration, the sample efficiency of TRPO and Deep GA is higher than MCTSPO, although they converge to similar final performance. In this case, the exploration effort done by MCTSPO could be redundant. This could be mitigated by reducing the exploration constant as well as the progressive widening parameter in MCTSPO.

6 Conclusion

This paper discussed the limitations of popular gradient-based DRL solvers on environments with sparse or deceptive rewards. We propose an approach called MCTSPO to balance exploration and exploitation of policy optimization using Monte-Carlo tree search in the policy space. We demonstrate how we can escape local optima and efficiently train policies in continuous control tasks. The algorithm shows promising results in different classic control and robotics tasks, outperforming TRPO and Deep GA.

In our implementation of MCTS, sampling is not executed in parallel, which limits the efficiency of MCTSPO as well as the training clock time. Parallel MCTS techniques are likely to further improve the time efficiency [Chaslot *et al.*, 2008b]. The combination of gradient-based methods with MCTSPO might further improve sample efficiency. For example, in the rollout, instead of *no mutation*, we could run several iterations of gradient-based optimization to increase the accuracy of the value estimates.

[1] In the case of an MDP with finite state and action spaces and returns in the range of $[0,1]$, an exploration constant of $\sqrt{2}$ gives a theoretical guarantee of finding the (global) optimal policy [Kocsis and Szepesvári, 2006]. In classic control, we consider environments with a continuous state and action space and the returns are approximately in this range. Nonetheless, we observed reasonable results with this parameter. For Roboschool robotics, the return ranges are much larger, so we scaled the exploration constant accordingly.

[2] Note that α and k are not necessarily equal.

[3] Although the total number of environment calls is the same for each algorithm, the training clock time for Deep GA and MCTSPO is approximately twice that of TRPO. This difference is mainly caused by the single-threaded sampling in our Deep GA and MCTSPO implementation. Parallel MCTS is an important area of future work.

Policy	Hyperparameter	Acrobot	MountainCar	BipedalWalker	Ant	HalfCheetah	Hopper
TRPO	step size 0.1	**0.0**	**0.0**	0.0	**−14.3**	**−1.90**	−1.92
	step size 1.0	0.0	−0.024	**0.022**	−19.6	−2.01	**−1.48**
Deep GA	population 100	0.364	0.8	0.103	**35.0**	−3.69	−2.76
	population 500	0.320	0.894	**0.233**	−14.1	−3.12	−2.54
	population 1000	**0.593**	**0.942**	0.153	−14.5	**−2.98**	**−2.46**
MCTSPO	$k = \alpha = 0.3$	0.684	0.934	**0.562**	**741**	**161**	**544**
	$k = \alpha = 0.5$	**0.772**	**0.938**	0.3	−16.5	−3.37	276
	$k = \alpha = 0.8$	0.638	0.937	0.042	−15.8	−2.85	91.1

Table 1: Average best return of TRPO, Deep GA, and MCTSPO using different hyperparameters on six task environments. The bold entries indicate the best performance found for each algorithm in each environment.

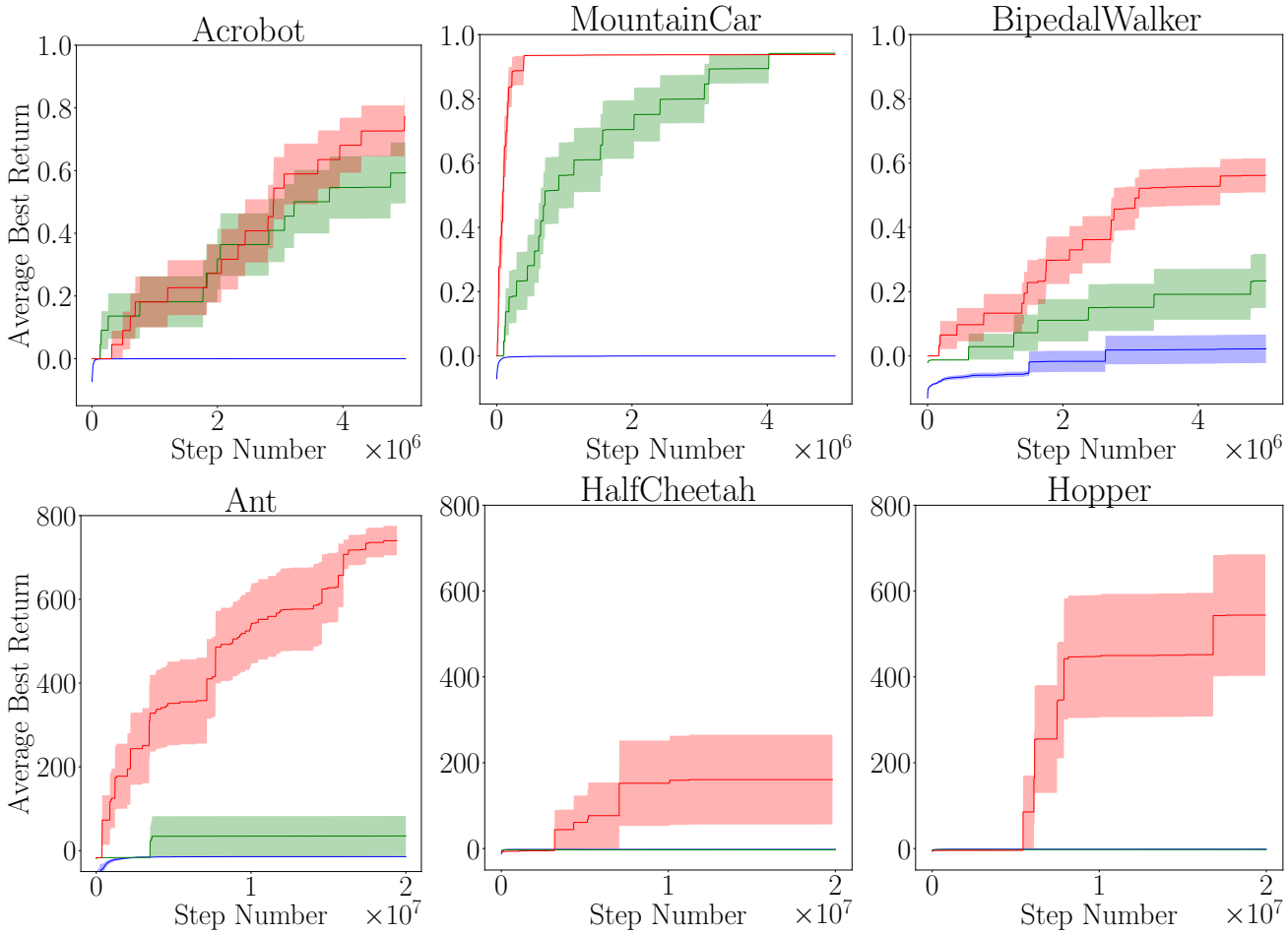

Figure 2: Performance curve for task environments using TRPO (blue), Deep GA (green), and MCTSPO (red) with best hyperparameters found. The hyperparemters for each algorithm are shown in Table 1 in bold.

Acknowledgments

We thank anonymous reviewers for their helpful feedback and suggestions. This work is sponsored through the Stanford Center for AI Safety. Zongzhang Zhang is in part supported by the National Natural Science Foundation of China under Grant No. 61876119, and the Natural Science Foundation of Jiangsu under Grant No. BK20181432, and the China Scholarship Council.

References

[Baier and Cowling, 2018] Hendrik Baier and Peter I. Cowling. Evolutionary MCTS for multi-action adversarial games. In *IEEE Conference on Computational Intelligence and Games (CIG)*, pages 1–8. IEEE, 2018.

[Brockman *et al.*, 2016] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. OpenAI gym. *arXiv:1606.01540*, 2016.

[Chaslot *et al.*, 2008a] Guillaume M.J.B Chaslot, Mark H.M. Winands, H. Jaap Van Den Herick, Jos W.H.M Uiterwijk, and Bruno Bouzy. Progressive strategies for Monte-Carlo tree search. *New Mathematics and Natural Computation*, 4(03):343–357, 2008.

[Chaslot *et al.*, 2008b] Guillaume M.J.B. Chaslot, Mark H.M. Winands, and H. Jaap Van Den Herik. Parallel Monte-Carlo tree search. In *International Conference on Computers and Games*, pages 60–71, 2008.

[Chrabaszcz *et al.*, 2018] Patryk Chrabaszcz, Ilya Loshchilov, and Frank Hutter. Back to basics: Benchmarking canonical evolution strategies for playing Atari. *arXiv:1802.08842*, 2018.

[Couëtoux *et al.*, 2011] Adrien Couëtoux, Jean-Baptiste Hoock, Nataliya Sokolovska, Olivier Teytaud, and Nicolas Bonnard. Continuous upper confidence trees. In *Learning and Intelligent Optimization (LION)*, pages 433–445, 2011.

[Duan *et al.*, 2016] Yan Duan, Xi Chen, Rein Houthooft, John Schulman, and Pieter Abbeel. Benchmarking deep reinforcement learning for continuous control. In *International Conference on Machine Learning (ICML)*, pages 1329–1338, 2016.

[Geramifard *et al.*, 2015] Alborz Geramifard, Christoph Dann, Robert H. Klein, William Dabney, and Jonathan P. How. RLPy: A value-function-based reinforcement learning framework for education and research. *Journal of Machine Learning Research*, 16:1573–1578, 2015.

[Grzes and Kudenko, 2009] Marek Grzes and Daniel Kudenko. Theoretical and empirical analysis of reward shaping in reinforcement learning. In *International Conference on Machine Learning and Applications*, pages 337–344. IEEE, 2009.

[Hessel *et al.*, 2018] Matteo Hessel, Joseph Modayil, Hado Van Hasselt, Tom Schaul, Georg Ostrovski, Will Dabney, Dan Horgan, Bilal Piot, Mohammad Azar, and David Silver. Rainbow: Combining improvements in deep reinforcement learning. In *AAAI Conference on Artificial Intelligence (AAAI)*, pages 3215–3222, 2018.

[Keller and Helmert, 2013] Thomas Keller and Malte Helmert. Trial-based heuristic tree search for finite horizon MDPs. In *International Conference on Automated Planning and Scheduling (ICAPS)*, pages 135–143, 2013.

[Khadka and Tumer, 2018] Shauharda Khadka and Kagan Tumer. Evolution-guided policy gradient in reinforcement learning. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1196–1208, 2018.

[Kingma and Ba, 2015] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

[Kochenderfer, 2015] Mykel J. Kochenderfer. *Decision Making Under Uncertainty: Theory and Application*. MIT Press, 2015.

[Kocsis and Szepesvári, 2006] Levente Kocsis and Csaba Szepesvári. Bandit based Monte-Carlo planning. In *European Conference on Machine Learning (ECML)*, pages 282–293, 2006.

[Lehman *et al.*, 2018] Joel Lehman, Jay Chen, Jeff Clune, and Kenneth O. Stanley. Safe mutations for deep and recurrent neural networks through output gradients. In *Genetic and Evolutionary Computation Conference*, pages 117–124. ACM, 2018.

[Mnih *et al.*, 2015] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

[Moore, 1991] Andrew Moore. *Efficient Memory-based Learning for Robot Control*. PhD thesis, Carnegie Mellon University, 1991.

[Murthy and Raibert, 1984] Seshashayee S. Murthy and Marc H. Raibert. 3D balance in legged locomotion: Modeling and simulation for the one-legged case. *ACM SIGGRAPH Computer Graphics*, 18(1):27–27, 1984.

[Salimans *et al.*, 2017] Tim Salimans, Jonathan Ho, Xi Chen, Szymon Sidor, and Ilya Sutskever. Evolution strategies as a scalable alternative to reinforcement learning. *arXiv:1703.03864*, 2017.

[Schulman *et al.*, 2015a] John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International Conference on Machine Learning (ICML)*, pages 1889–1897, 2015.

[Schulman *et al.*, 2015b] John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. *arXiv:1506.02438*, 2015.

[Schulman *et al.*, 2017] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv:1707.06347*, 2017.

[Such *et al.*, 2017] Felipe Petroski Such, Vashisht Madhavan, Edoardo Conti, Joel Lehman, Kenneth O. Stanley, and Jeff Clune. Deep neuroevolution: Genetic algorithms are a competitive alternative for training deep neural networks for reinforcement learning. *arXiv:1712.06567*, 2017.

[Wawrzynski, 2007] Pawel Wawrzynski. Learning to control a 6-degree-of-freedom walking robot. In *EUROCON 2007-The International Conference on " Computer as a Tool"*, pages 698–705. IEEE, 2007.

[Wu *et al.*, 2017] Yuhuai Wu, Elman Mansimov, Roger B. Grosse, Shun Liao, and Jimmy Ba. Scalable trust-region method for deep reinforcement learning using Kronecker-factored approximation. In *Advances in Neural Information Processing Systems (NIPS)*, pages 5279–5288, 2017.