MIMICRY CONSTRAINT POLICY OPTIMIZATION

Mingxuan Jing*, **Xiaojian Ma***
Department of Computer Science and Technology
Tsinghua University
{maxj14,jmx16}@mails.tsinghua.edu.cn

Fuchun Sun, Huaping Liu
Department of Computer Science and Technology
Tsinghua University
{fcsun,hpliu}@tsinghua.edu.cn

ABSTRACT

Exploration plays an important role in reinforcement learning. However, the exploration policy in most RL algorithms is usually randomly initialized, which can be extremely ineffective especially when the state-action space is tremendously large and environmental feedback is sparse. In contrast, imitation learning has shown promising results in balancing exploration and exploitation with expert demonstrations, albeit its resource-consuming drawback in collecting high-quality demonstrations. In this paper, we study the intersection of these two approaches and propose a novel *mimicry constraint* for guiding the agent to explore towards the area with high return by leveraging the expert demonstrations as an exploration **prior**. We show that our Mimicry Constraint Policy Optimization (MCPO) improves policy optimization by enforcing the occupancy measure matching to the expert, and can be integrated into many reinforcement learning methods. Considerable empirical results on challenging benchmarks with sparse reward demonstrate that our method attains consistent improvement over baselines, even when the demonstrations are few and imperfect.

1 INTRODUCTION

For robots and other autonomous agents, a crucial aspect of intelligence is the ability to learn from interacting with the surrounding environment and human (Mnih et al., 2015; Jing et al., 2019). Reinforcement learning (RL) (Sutton & Barto, 1998) and imitation learning (IL) (Schaal, 1997) are two major families of algorithms that were proposed to realize these learning paradigms. However, RL and IL both have some drawbacks: Most of the RL approaches may suffer from inefficient exploration when the feasible action and state spaces are large while the environmental feedback is sparse, and the resource consumption for satisfying the requirement of collecting enough high-quality demonstrations in IL can also be unaffordable. These drawbacks may prevent RL and IL from generalizing to complex real-world tasks.

To eliminate these drawbacks by combining the best of both RL and IL, reinforcement learning from demonstrations (RLfD) (Brys et al., 2015; Hester et al., 2018; Kang et al., 2018) has been paid close attention recently. It studies the intersection of RL and IL, and tries to improve the learning process of RL by leveraging the demonstrations data and learning methods from IL. In this paper, we mainly focus on RLfD and specifically concentrate on its most critical concern: improving the exploration efficiency of RL with expert demonstrations.

Most of the existing research on this problem either depends on the large number of demonstrations (Brys et al., 2015) or complex training strategies and models (Kang et al., 2018), these requirements, in our view, are **not** realistic to comply, especially in real-world complex domains, and may also induce low convergence efficiency. In this work, we aim to tackle the problem of RLfD with a lightweight and effective method, while only few demonstrations will be required. To

*These two authors contributed equally.

this end, inspired by the occupancy measure matching method in imitation learning (Ho & Ermon, 2016; Kim & Park, 2018), we introduce a novel *mimicry constraint* defined by the discrepancy between the state-action distribution between expert and agent into policy optimization, and propose Mimicry Constraint Policy Optimization (MCPO) algorithm. By leveraging the non-parametric distance metric MMD (Gretton et al., 2007) and the dual solving of the corresponding constrained policy optimization problem, MCPO is easy to implement and shown to be efficient. With only few and imperfect demonstrations, our method achieves consistent improvement over other counterparts on several challenging control benchmarks with sparse environmental feedback.

2 Related Work

RL from demonstrations. Recent research on reinforcement learning from demonstrations (RLfD) mainly focus on integrating the expert demonstrations into normal policy optimization methods, including providing prior for exploration (Brys et al., 2015; Chemali & Lazaric, 2015), facilitating the policy convergence (Cederborg et al., 2015) and improving final performance of learned policy (Hester et al., 2018; Sun et al., 2018; Kang et al., 2018). These approaches mostly share the same motivation as ours. However, they either require a significantly large amount of perfect demonstrations (Brys et al., 2015; Hester et al., 2018), which can be unrealistic for real-world tasks or depend on complex training strategies and models (Kang et al., 2018). In contrast, our MCPO is lightweight and can still achieve comparable performances on challenging tasks with sparse reward and few demonstrations; thus these drawbacks can be eliminated.

Constrained optimization in RL. Although most of the policy optimization problems in RL do not have explicit constraints, we notice some recent related work on constrained policy optimization and its efficient solving methods (Achiam et al., 2017; Tessler et al., 2019). However, these approaches are all built upon CMDP (Altman, 1999), in which only the cost-like constraints are studied. As our proposed mimicry constraint does not belong to this family, we develop a new optimization method for solving policy optimization with the mimicry constraint efficiently.

3 Methodology

In this section, we will first model the reinforcement learning from demonstrations (RLfD) as a constrained optimization problem with a novel *mimicry constraint* defined by the provided demonstrations. To solve this challenging optimization problem, we present a practical solution based on local approximation. Finally, an efficient implementation will be provided.

3.1 Reinforcement Learning from Demonstrations via Mimicry Constraint

On solving RLfD, the main idea behind our MCPO is to leverage few demonstrations from an expert policy π_E as an exploration prior and guide the agent to explore towards the area with higher return and converge much faster. To achieve this prior, we introduce a *mimicry constraint* constructed with the provided demonstrations to realize a new constrained optimization in addition to the original problem setting of local policy search (Kakade, 2002; Schulman et al., 2015), a popular category of RL approaches.

More specifically, suppose π_θ and π_E are a θ-parameterized policy and an expert policy respectively, we could denote their corresponding occupancy measure as ρ_{π_θ} and ρ_E, and they will also be the (unnormalized) density of exploration trajectory under π_θ and expert demonstrations.To encourage the agent with π_θ to explore nearer to the area specified by the demonstrations, we thus define the *mimicry constraint* as

$$\mathbb{D}\left(\rho_{\pi_\theta}(s,a) \| \rho_E(s,a)\right) \leq d, \tag{1}$$

where $\mathbb{D}(\cdot\|\cdot)$ can be any discrepancy measure, and d is the tolerance that controls the range of exploration w.r.t. the demonstrations, which are samples to ρ_E. We will discuss the choice of d lately.

At first glance, the constraint requires the agent policy to not being so far away from the expert policy by limiting the discrepancy between their corresponding occupancy measure ρ_{π_θ} and ρ_E. Therefore,

the learned policy will have to stay close to the expert to satisfy the constraint during the policy optimization procedure, which also makes the meanwhile on-policy exploration close to the area with potentially high return and speeds up the convergence. On the other hand, since the constraint is likely to be unsatisfied due to either the initialization (i.e., random policy) or approximation, the solving of policy optimization with this constraint will usually contain a recovery phase to satisfy it, namely, treating it as an objective instead. This can be seen as a policy imitation with demonstrations mechanism (Silver et al., 2016; Cruz Jr et al., 2017) that has been shown to be effective in improving the exploration in subsequent policy optimization.

Overall optimization problem. Notice that, the proposed *mimicry constraint* depends on current policy π_θ; thus we can integrate the constraint into the optimization problem of RL, i.e., local policy search. This concludes as the following optimization problem in k-th optimization step

$$\begin{aligned} \theta_{k+1} = \arg\max_{\theta} \ & \mathbb{E}_{\pi_{\theta_k}}\left[A_{\pi_{\theta_k}}(s,a)\right] \\ \text{s.t. } & \mathbb{D}\left(\rho_{\pi_\theta}(s,a) \| \rho_E(s,a)\right) \leq d_k \\ & \mathbb{D}_{\mathrm{KL}}\left(\pi_{\theta_k} \| \pi_{\theta_{k+1}}\right) \leq \delta. \end{aligned} \tag{2}$$

Furthermore, since only the samples from ρ_{π_θ} (exploration rollout) and ρ_E (demonstrations) are available during policy optimization, we adopt the non-parametric distance metric MMD as the discrepancy measure in the mimicry constraint to avoid the density estimation bias

$$\mathrm{MMD}[\mathcal{H}, \rho_{\pi_{\theta_k}}, \rho_E] \leq d_k, \tag{3}$$

and it can be approximated via its empirical estimation introduced in (Gretton et al., 2007). The optimization problem equation 2 is the core of our proposed Mimicry Constraint Policy Optimization (MCPO). Later, a practical solution for it will be provided.

3.2 Approximated Solving for MCPO

Compared to the original local policy search, solving our new optimization problem with mimicry constraint equation 2 can be much more challenging due to: 1. **Feasibility**, it may be difficult to find a feasible solution with the additional constraint. 2. **Scalability**, policies that are characterized by a model with high-dimensional parameter space, i.e., neural networks, the computation cost of the new constraint will become unaffordable. To this end, we propose an approximated solving for MCPO by linearizing around π_{θ_k} at the k-th optimization step. Denoting the gradient of the objective as g, the current MMD at θ_k as d_{θ_k} and its gradient as b, the Hessian of the KL-divergence as H[1], the approximation to equation 2 is

$$\begin{aligned} \theta_{k+1} = \arg\max_{\theta} \ & g^T(\theta - \theta_k) \\ \text{s.t. } & b^T(\theta - \theta_k) + d_{\theta_k} \leq d_k \\ & \frac{1}{2}(\theta - \theta_k)^T H(\theta - \theta_k) \leq \delta. \end{aligned} \tag{4}$$

The approximated optimization problem above is convex as the Fisher information matrix H is always positive semi-definite (Schulman et al., 2015). Therefore, compared to its original form equation 2, a feasible solution can be found more easily using duality. In particular, given λ and ν as the Lagrange multipliers for KL-divergence and MMD constraints, a corresponding dual to equation 4 can be written as

$$\max_{\substack{\lambda \geq 0 \\ \nu \geq 0}} -\frac{1}{2\lambda}(g^T u + 2\nu b^T u + \nu^2 b^T r) - \nu c - \lambda\delta, \tag{5}$$

where $u = H^{-1}g$, $r = H^{-1}b$, $c = d_k - d_{\theta_k}$. Since the number of variables in this dual problem is much less than the dimension of θ, the computation cost will also be much less than solving equation 2. The closed-form expression of optimal solution λ^*, ν^* can be derived by firstly obtaining

[1]The KL constraint should be approximated via second-order expansion since its first order gradient is zero at $\pi_{\theta_k} = \pi_\theta$. More details on the solving of duality can be found in the supplementary materials.

and substituting ν^*, then discussing the sub-case and finally gets λ^*. Suppose we have the optimal solution λ^*, ν^* of this dual problem, the solution to the primal one will be

$$\theta^*_{k+1} = \theta_k - \frac{1}{\lambda^*}(u + r\nu^*). \tag{6}$$

When both the constraints are feasible, we update the policy parameter θ by solving the dual for λ^* and ν^* equation 6. However, due to the initialization and approximation error, the proposed update rule may sometimes not satisfies the constraints in equation 2, especially at the beginning of optimization. In the next section, we will provide more details on ensuring the feasibility.

3.3 Implementation

In this section, we will present some implementation details of the proposed method, including the techniques about ensuring the feasibility of solving MCPO, the kernel selection in MMD and finally the choice of the tolerance factor d in the mimicry constraint.

Feasibility. The main sources that account for the feasibility issues in MCPO are twofold. One lies in the beginning phase. As the parameter θ is usually randomly initialized, it may induce infeasibility when the optimization starts. Therefore, a recovery method is necessary. In our implementation, we propose to transform the constraint into an objective to decrease it

$$\theta^* = \arg\min_{\theta} \mathrm{MMD}[\mathcal{H}, \rho_{\pi_\theta}, \rho_E]. \tag{7}$$

As we mentioned in Section 3.1, this recovery approach can be regarded as a policy imitation with demonstrations, and we also notice that this is exactly the objective of a recently proposed imitation learning algorithm (Kim & Park, 2018).

Another source of infeasibility comes from equation 6. The update rule may not satisfy the constraints due to the approximation error. To this end, a backtracking linesearch along $\Delta\theta = -\lambda^{*-1}(u + r\nu^*)$ is used to ensure the constraint satisfaction. To further reduce the computation cost, we also adopt the conjugate gradient method like (Schulman et al., 2015) to approximately compute the inverse of H and its products.

Kernel selection. When $k(\cdot, x)$ is a characteristic kernel, $\mathrm{MMD}[\mathcal{H}, p, q] = 0$ if and only if $p = q$ (Smola et al., 2007). This property can help eliminate the inconsistency between minimizing MMD and morphing two distributions. Therefore, we use a characteristic radial basis function kernel

$$k(x', x) = \exp(-\frac{\|x - x'\|_2^2}{2\sigma_X \sigma_{X'}}). \tag{8}$$

The bandwidth parameter σ_X and $\sigma_{X'}$ are chosen as the standard deviation of given sample set X and X' instead of a fixed value for better adaptability on various state-action pair distributions under different experiments.

Constraint choice. For tolerance d in mimicry constraint, a straightforward choice is handcrafting a fixed d. As discussed on previous sections, a smaller tolerance would restrict the distribution of state-action pair under learned policy to be closer to expert, which leads to narrower exploration area; thus a larger tolerance should be used for the imperfect demonstrations. On the other hand, although a fixed tolerance could guide the exploration at the beginning, it would hurt the subsequent policy learning when the demonstrations are imperfect. We thus optionally adopt an annealing mechanism to dynamically adjust d along the training procedure. Specifically, we use an update rule $d_{k+1} \leftarrow d_k + d_k \cdot \epsilon$ with a factor ϵ to gradually increase d for avoiding this issue.

4 Discussion

In this section, we will discuss some prior counterparts on RLfD, including policy pre-training (Silver et al., 2016) and POfD (Kang et al., 2018), and show their connections to us. We will also provide some intuitions of our method.

Algorithm 1 MCPO

Input: Expert demonstrations $\mathcal{D}_E = \{\zeta_i^E\}$, policy π_{θ_0}, initial constraints tolerance d_0, δ, annealing factor ϵ, maximal iterations N.
for k = 0 to N **do**
  Sample rollout $\mathcal{D}_\pi$ with π_{θ_k}.
  Estimate $\hat{g}$, $\hat{b}$, $\hat{H}$ with samples from $\mathcal{D}_E$ and $\mathcal{D}_\pi$.
  if the optimization problem equation 4 is feasible **then**
    Solve the dual problem equation 5 to obtain λ^*, ν^*.
    Compute update step proposal $\Delta\theta$ as equation 6.
    Update the policy by backtraching linesearch along $\Delta\theta$ to ensure the satisfaction of constraints.
  else
    Update the policy via the recovery objective equation 7.
  end if
  Annealing the tolerance d_k: $d_{k+1} \leftarrow d_k + d_k \cdot \epsilon$.
end for

A straightforward solution to RLfD will be **Pre-training** the policy with demonstrations via imitation learning, i.e., behaviour cloning (Schaal, 1997; Atkeson & Schaal, 1997), then proceeding with normal reinforcement learning (Silver et al., 2016). The first step is similar to MCPO when the constraints are unsatisfied at the beginning. However, this approach cannot guarantee the exploration quality in the later RL step, and thus the subsequent training can still suffer from poor sample efficiency in the case with large exploration space and sparse feedback.

POfD (Kang et al., 2018) is the work that closest to us. It separately trains a GAN-based distance metric model to measure the discrepancy between ρ_π and ρ_E and integrates it into the original task reward to provide an extra penalty cost. Compared to POfD, our MCPO has two significant advantages: 1. **Lightweight**, as we use non-parametric MMD as the distance metric, there is no need to additionally train a complex model, which is easier to implement and can eliminate the error introduced by the training bias. 2. **Effective**, by leveraging the demonstrations as a *mimicry constraint* instead of a penalty, the discrepancy between the occupancy measure of expert and agent can be optimized more efficiently, which can better guide the exploration. Finally, our method can be extended to other value-based methods like DQN (Mnih et al., 2013) via normalized advantage functions (Gu et al., 2016), a general approach that turns a Q function into a stochastic policy. Therefore, MCPO can also be integrated into any other RL methods as a universal component as POfD does.

On few and imperfect demonstrations. In our experiments[2], we find MCPO can facilitate policy optimization even when there are only few and imperfect demonstrations. The intuitions behind these observations come from the design of *mimicry constraint*. Firstly, since MMD does not rely on a large number of samples for training or evaluation (Smola et al., 2007), it can generalize well to the few samples setting. Secondly, imperfect demonstrations can still indicate an area with a relatively higher return as an exploration prior. With tolerance annealing strategy, MCPO can benefit from this prior while final performances will not be affected by the drawbacks in the demonstrations.

5 Conclusion

We have presented MCPO, a policy optimization algorithm that aims to improve the exploration process in reinforcement learning with a novel *mimicry constraint*. By introducing the discrepancy between the occupancy measure between the expert and agent as a constraint and providing an efficient solution for the constrained policy optimization problem, our algorithms can facilitate the policy learning process by improving the quality of exploration. Experiments on challenging control benchmarks demonstrate the effectiveness of our proposed method. As the main part of this paper is solving an optimization problem with mimicry constraint, an exciting direction of future work could

[2]We defer the empirical results to the supplementary due to the space limit.

be theoretical analysis on the dynamics and efficiency during the optimization procedure, which may bring more interpretability to our method.

REFERENCES

Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *International Conference on Machine Learning (ICML)*, 2017.

Eitan Altman. *Constrained Markov decision processes*, volume 7. CRC Press, 1999.

Christopher G Atkeson and Stefan Schaal. Robot learning from demonstration. In *International Conference on Machine Learning (ICML)*, 1997.

Tim Brys, Anna Harutyunyan, Halit Bener Suay, Sonia Chernova, Matthew E Taylor, and Ann Nowé. Reinforcement learning from demonstration through shaping. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2015.

Thomas Cederborg, Ishaan Grover, Charles L Isbell, and Andrea Lockerd Thomaz. Policy shaping with human teachers. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2015.

Jessica Chemali and Alessandro Lazaric. Direct policy iteration with demonstrations. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2015.

Gabriel V Cruz Jr, Yunshu Du, and Matthew E Taylor. Pre-training neural networks with human demonstrations for deep reinforcement learning. *arXiv preprint arXiv:1709.04083*, 2017.

Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex J Smola. A kernel method for the two-sample-problem. In *Advances in neural information processing systems (NIPS)*, 2007.

Shixiang Gu, Timothy Lillicrap, Ilya Sutskever, and Sergey Levine. Continuous deep q-learning with model-based acceleration. In *International Conference on Machine Learning (ICML)*, 2016.

Todd Hester, Matej Vecerik, Olivier Pietquin, Marc Lanctot, Tom Schaul, Bilal Piot, Dan Horgan, John Quan, Andrew Sendonaris, Ian Osband, et al. Deep q-learning from demonstrations. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2018.

Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. In *Advances in Neural Information Processing Systems (NIPS)*, 2016.

Mingxuan Jing, Xiaojian Ma, Wenbing Huang, Fuchun Sun, and Huaping Liu. Task transfer by preference-based cost learning. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2019.

Sham M Kakade. A natural policy gradient. In *Advances in neural information processing systems (NIPS)*, 2002.

Bingyi Kang, Zequn Jie, and Jiashi Feng. Policy optimization with demonstrations. In *International Conference on Machine Learning (ICML)*, 2018.

Kee-Eung Kim and Hyun Soo Park. Imitation learning via kernel mean embedding. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2018.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

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, 2015.

Stefan Schaal. Learning from demonstration. In *Advances in neural information processing systems (NIPS)*, 1997.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International Conference on Machine Learning (ICML)*, 2015.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484, 2016.

Alex Smola, Arthur Gretton, Le Song, and Bernhard Schölkopf. A hilbert space embedding for distributions. In *International Conference on Algorithmic Learning Theory (ALT)*, 2007.

Wen Sun, J. Andrew Bagnell, and Byron Boots. Truncated horizon policy search: Combining reinforcement learning and imitation learning. In *International Conference on Learning Representations (ICLR)*, 2018.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT Press, 1998.

Chen Tessler, Daniel J. Mankowitz, and Shie Mannor. Reward constrained policy optimization. In *International Conference on Learning Representations (ICLR)*, 2019.