

Meta-Learning for Multi-objective Reinforcement Learning

Xi Chen¹, Ali Ghadirzadeh^{1,2}, Mårten Björkman¹ and Patric Jensfelt¹

Abstract—Multi-objective reinforcement learning (MORL) is the generalization of standard reinforcement learning (RL) approaches to solve sequential decision making problems that consist of several, possibly conflicting, objectives. Generally, in such formulations, there is no single optimal policy which optimizes all the objectives simultaneously, and instead, a number of policies has to be found, each optimizing a preference of the objectives. In this paper, we introduce a novel MORL approach by training a meta-policy, a policy simultaneously trained with multiple tasks sampled from a task distribution, for a number of randomly sampled Markov decision processes (MDPs). In other words, the MORL is framed as a meta-learning problem, with the task distribution given by a distribution over the preferences. We demonstrate that such a formulation results in a better approximation of the Pareto optimal solutions, in terms of both the optimality and the computational efficiency. We evaluated our method on obtaining Pareto optimal policies using a number of continuous control problems with high degrees of freedom.

I. INTRODUCTION

Reinforcement learning (RL) is a framework to train an agent to acquire a behavior by reinforcing actions that maximize a notion of task-relevant future rewards. A reward function, i.e., the function that assigns a reward value to every action-decision made by the agent, is designed to guide the training to implement the behavior. For specific applications, there are learning frameworks that train with sparse rewards [1], [2]. However, in many real world problems, a behavior is best realized by finding a trade-off among a number of conflicting reward functions. This is known as multi-objective reinforcement learning (MORL), which is a fundamental problem in designing many autonomous systems. As an example, real systems consume energy and, in most cases, reducing energy consumption as a desirable behavior lowers the performance of the system. Other examples are safety and stability which, also in many cases, contradict the main objectives of the system.

In many RL problems, the trade-off between different objectives is found by constructing a synthetic reward function, e.g., a weighted linear sum of the objectives, using expert knowledge [3], [4]. This approach may not be suitable for problems that, (1) aim at realizing a multifaceted behavior consisting of several phases in sequence, therefore, preventing utility of a set of predetermined preferences over the objectives, (2) are required to adapt itself to operate in different modes, e.g., to operate in energy-saving versus high-performance mode, and (3) for which it is hard to find a

set of suitable preference over objectives manually, a problem known as preference elicitation (PE) [5].

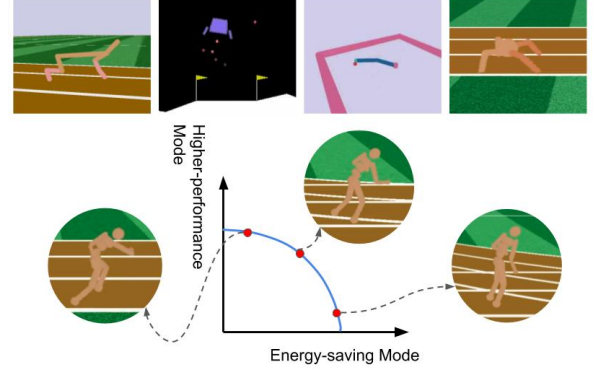


Fig. 1: The five continuous control tasks that are used in our experiments.

However, most of the current MORL solutions are limited to simple toy examples with discrete state-action spaces [5], [6], [7]. To the best of our knowledge, there is no MORL approach to train deep policies with different objectives for complex continuous control tasks.

In this work, we introduce a novel approach to solve deep MORL problems based on model-agnostic meta-learning [8]. We train a meta-policy with tasks sampled from a distribution over MDPs with different preferences over the objectives. We demonstrate that our proposed approach outperforms direct policy search methods, e.g., [9], at obtaining Pareto optimal policies for different continuous RL problems. Given a reward vector, a policy is Pareto optimal w.r.t. the expected return measure if it results in higher expected return for at least one component of the vector compared to all other policies.

The rest of this paper is organized as following: We outline related work in Section II. We introduce required background in Section III. Section V provides the details of our method. In Section VI, we explain our experimental results. Finally, in Section VII, we conclude the paper and introduce our future work.

II. RELATED WORK

MORL approaches mainly try to find Pareto optimal solutions which are defined as non-dominated solutions representing trade-offs among multiple conflicting objectives [5], [6], [7]. MORL methods can be divided into two main categories: single- and multi-policy approaches [5], [10], with the main difference being the number of policies that are learned at each run. Single-policy algorithms obtain solutions

¹RPL, KTH Royal Institute of Technology, Sweden

²Intelligent Robotics research group, Aalto University, Finland
[xi8][algh][celle][patric]@kth.se

with different trade-offs between the objectives by running the RL method several times with different parameters. In contrast, multi-policy algorithms update several policies simultaneously in each run to improve the current Pareto optimal solutions.

In case a scalar value can be computed from the multiple objectives using domain-specific expert knowledge, the problem can be converted to a single objective RL problem which can be solved based on standard RL solutions. As examples, Schulman et al., [3] and Duan et al., [4] applied policy search methods on a number of continuous control benchmarks by hand-designing a synthetic reward function from multiple objectives for different locomotion tasks. However, setting the parameters a priori requires fine-tuning and it is non-intuitive to find the right trade-off [5], [11]. Lizotte et al., [12] proposed a multi-objective Q-learning solution by linearly combining the objectives related to symptoms and side-effects of different treatment plans. They identified dominated actions, as well as, preferred actions at different areas of the objective space by using a tabular Q-function with an extra entry for the objectives. However, it is well-known that multi-objective Q-learning by the weighted-sum approach cannot approximate concave parts of the true Pareto front [13]. Van Moffaert et al., [14] introduced a similar multi-objective Q-learning method but based on the Chebyshev scalarization function as a replacement for the weighted-sum to approximate the Pareto front. Still the method updates the Q-function based on action-decisions made according to a preference over the objectives which is encoded by the weights of the Chebyshev function, resulting in sub-optimal Pareto fronts. In a later study [15], they improved the method by exploiting a hypervolume-based action-selection mechanism. The hypervolume is a measure used for performance assessment of the Pareto optimal solutions and not a single preference given by a set of weights. However, estimation of the hypervolume is a computationally expensive NP-hard problem [15]. Our approach trains stochastic policies with a weighted-sum scalarization method by sampling different random weights during the meta-learning phase to alleviate the aforementioned issue regarding the sub-optimality of the Pareto front, while still performing tractable computations.

The work done by Natarajan and Tadepalli [16] resembles our work, in that it also tries to obtain a number of policies, based on which a non-dominated Pareto front can be established for preferences over different objectives encoded by a weight vector. In their work, a bag of policies is constructed recursively to approximate the front, as well as to initialize new policies such that, the policy training can be resumed from the closest policy in the bag. We instead propose to initialize new policies, in a so-called adaptation phase, from the updated meta-policy which is trained with many different weight values. The meta-policy is proven to adapt to new conditions faster [8], and can be realized by both stochastic and deterministic policies.

Recently, policy search approaches have been studied in the context of MORL. Parisi et al., [9] proposed to estimate a number of Pareto optimal policies by performing gradient

ascent in the policy parameter space by solving different single-objective RL problems found by different convex combinations of the objectives. In contrast, our method does not initialize each policy randomly and it obtains a suitable initial policy parameter in the meta-learning phase, resulting in a more efficient way of estimating the Pareto optimal policies. Furthermore, as we demonstrate in the experiment section, our method outperforms [9] for complicated tasks requiring a hierarchy of skills to be obtained in sequences.

Pirotta et al., [17] introduced a manifold-based policy search MORL solution which assumes policies to be sampled from a manifold, i.e., a parametric distribution over the policy parameter space. They proposed to update the manifold according to an indicator function, such that the sampled policies construct an improved Pareto front. In a more recent work, Parisi et al., [18] extend the method to work with the hypervolume indicator function as a more well-suited measure when evaluating optimality of different Pareto fronts. However, the number of parameters of the manifold grows quadratically with the number of policy parameters. As an example, a policy with 30 parameters requires 7448 parameters to model the manifold [18] with Gaussian mixture models. As the result, these approaches may not be applicable to train deep policies with several thousands of parameters.

In this paper, we propose to frame policy search MORL for continuous control with large numbers of policy parameters, e.g., deep policies, as a meta-learning problem. In contrast to earlier policy search approaches, we train a meta-policy for different tasks sampled from a distribution of the preferences over the objectives, such that, by fine-tuning the meta-policy with a few gradient updates, the Pareto front can be constructed more efficiently.

III. PRELIMINARIES

A multi-objective sequential decision making process can be represented by a Markov decision process (MDP) which is defined by the tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, r, p(s_0), \gamma)$, where \mathcal{S} is a set of states, \mathcal{A} is a set of actions, $\mathcal{P} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow [0, 1]$ is the action-conditioned state transition probability, $r : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}^q$ represents the reward functions corresponding to q different objectives, $p(s_0) : \mathcal{S} \rightarrow [0, 1]$ is the distribution of initial states, and $\gamma \in [0, 1]$ is a discount factor. We consider episodic tasks in which an episode starts by sampling an initial state from $p(s_0)$, and then for every time-step $t < H$, sampling actions a_t from $\pi_\theta(a_t|s_t)$, a parametric stochastic policy. The successor state in each time step is given according to $p(s_{t+1}|s_t, a_t)$ and a reward vector $r_t = r(s_t, a_t)$ is provided at every time-step by the environment. For every state-action pair in the episode, the return is a vector defined as the discounted sum of future rewards, $R_t = \sum_{t'=t} \gamma^{t'-t} r(s_{t'}, a_{t'})$.

IV. MULTI-OBJECTIVE POLICY SEARCH

Reinforcement learning is based on performing exploratory actions and reinforcing the actions that result in return outcomes exceeding the expectation of the agent.

The expectation of the agent is modeled by a state value function, represented by a neural network in our work, and is continually updated to model the expected return for a given state s while following the policy π_θ ,

$$V^\pi(s) = \mathbb{E}[\sum_{t=0}^T \gamma^t r_t | s_0 = s].$$

The agent performs exploratory actions and compare the actual return outcomes for every state-action pair with the expected value to form the advantage function,

$$A(s_t, a_t) = R_t - V^\pi(s_t).$$

Here, the advantage is a vector which can be converted to a scalar value by a parametric scalarization function $f_\omega(A)$, e.g., the weighted sum, $f_\omega(A) = \sum_{i=1}^q \omega_i A_i$. The policy parameters are updated such that state-action pairs with $f_\omega(A) > 0$ are reinforced, i.e., become more probable in the future. This can be achieved by minimizing a loss function $\mathcal{L}(\theta, \omega)$, in our case the clipped version [3] of TRPO loss [19],

$$\mathcal{L}_{trpo}(\theta, \omega) = \mathbb{E}_\tau[\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)} f_\omega(A_t) - \beta D_{KL}(\pi_{\theta_{old}} || \pi_\theta)],$$

where, D_{KL} represents Kullback Leibler (KL)-divergence and τ represents the state-action trajectories and β is a scalar parameter found empirically. The parameters are updated such that the new policy assigns higher probability to state-action pairs resulting in positive scalarized advantages while the assigned policy distribution is not drastically changed from the old policy $\pi_{\theta_{old}}$.

V. MULTI-OBJECTIVE META-RL (META-MORL)

Meta-RL algorithms train a meta-policy π_θ by multiple tasks sampled from a given task distribution. As the result, the agent can learn a novel task, sampled from the same task distribution, more efficiently with limited number of training data [8]. In this section, we describe how this paradigm can be applied to the MORL domain to efficiently find a number of Pareto optimal policies. The scalarization function $f_\omega(\cdot)$, introduced in the previous section, converts a MORL problem to different single-objective RL problems by different values of ω . In fact, a new task is defined by setting the weight parameter ω to arbitrary (non-negative) values. Therefore, a distribution over tasks can be found by assigning a distribution over the weight parameter $p(\omega)$. In this case, meta-learning can be applied to find a meta-policy which is trained with all the tasks sampled from the task distribution. The meta-policy, once it is trained, does not contribute to the construction of the Pareto front by itself, but it is used as an optimal initial value of the policy parameters to efficiently train multiple policies with different objectives by a few learning iterations. These policies then contribute to construct the Pareto optimal solutions.

We follow the model-agnostic meta-learning approach introduced in [8]. In this formulation, the learning consists of three different phases, (1) an adaptation phase, in which a number of policies are updated for a few iterations from

the meta-policy, (2) a meta-learning phase, in which the meta-policy is updated by aggregating data generated by the policies trained in the previous phase, and (3) a fine-tuning phase, in which, the Pareto optimal policies are trained by being initialized with the meta-policy parameters, once the meta-policy is converged. In the following sections, different phases of the training are introduced in more details.

A. Adaptation phase

In the adaptation phase a number of tasks ω_i , sampled from $p(\omega)$, are defined. For each task, a policy is initialized with the parameters of the meta-policy. The policies are updated for one iteration using state-action trajectories generated by running the meta-policy, and return values found according to the assigned task (specified by ω_i). In short, each policy π_{θ_i} is trained as,

$$\theta_i = \operatorname{argmax}_{\theta'} \mathcal{L}(\theta', \omega_i) |_{\theta_{init}=\theta}, \quad (1)$$

where, θ is the meta-policy parameter.

B. Meta-learning phase

The meta-policy is updated in the meta-learning phase by aggregating the information of the policies π_{θ_i} trained in the previous step,

$$\theta = \operatorname{argmax}_{\theta'} \sum_i \mathcal{L}(\theta', \omega_i) |_{\theta_{init}=\theta}. \quad (2)$$

C. Fine-tuning phase

Finally, once the meta-policy is trained, a set of Pareto optimal policies is found by fine-tuning the meta-policy for a number of iterations with different tasks randomly sampled from the task distribution $p(\omega)$.

Algorithm 1 Meta-MORL policy training

Input $p(\omega)$: the distributions over the tasks

- 1: Randomly initialize the meta-policy π_θ
 - 2: **while** training π_θ **do**
 - 3: Sample trajectories \mathcal{D} following π_θ
 - 4: **for** each task i **do**
 - 5: Sample $\omega_i \sim p(\omega)$
 - 6: Update π_{θ_i} with \mathcal{D} according to Eq. 1
 - 7: Sample trajectories \mathcal{D}_i following π_{θ_i}
 - 8: Update meta-policy with \mathcal{D}_i, ω_i according to Eq. 2
 - 9: Construct the Pareto front by the fine-tuning phase
-

The details of our method are provided in Algorithm 1. The meta-policy is initialized randomly and it is updated in the training loop. In each iteration, trajectories of states, actions and rewards are sampled by running the meta-policy. For every task, a weight vector is sampled and a new policy is trained, in the adaptation phase. The trained policies are used to sample new trajectories based on which the meta-policy is updated in the meta-learning phase. Finally, a set of Pareto front policies is obtained by initializing the policies with the meta-policy and training them for a number of iterations.

Our method resembles the multi-policy approaches, in that it optimizes a meta-policy to estimate the Pareto front implicitly, but it also resembles the single-policy approaches, since the final Pareto front policies are trained individually for each objective during the fine-tuning phase.

VI. EXPERIMENT

In the experiments, we aim to answer the following questions: (1) does meta-MORL successfully train a set of policies with a large number of parameters that estimate the Pareto front for continuous control problems? (2) compared to direct training using a set of fixed weights over the objectives, does meta-MORL perform better considering the training time and the optimality of the resulting Pareto front? To answer these questions, we prepared five continuous control tasks with several conflicting objectives in different simulated environments (Fig. 1). We compared the results of our approach against the Radial Algorithm (RA) [9] regarding the training time and the quality of the estimated Pareto front. For all of the tasks, we exploited the Proximal Policy Optimization (PPO) [3] algorithm as the policy learning method.

A. Environments

The simulated environments are provided by Roboschool [20] and OpenAI gym [21]. The original environments return a single reward, which is the summation of rewards with respect to several objectives. Instead, we decompose the objectives to construct a reward vector. For each task, there is a minimum requirement that a policy needs to achieve in order to be considered as a valid policy (e.g., neither causing crashes nor joint stuck during the execution time). A detailed description of each task is presented in the followings.

1) *Reacher*: We start with a simple control task using a simulated arm with 2 degrees-of-freedom (DoFs). The goal is to move the tip of the arm as close as possible to a random target position. The environment returns a 3-dimensional reward vector, which is related to the distance from the arm tip to the target position, an energy consumption cost and a fixed joint stuck term penalizing the stuck motors. Among the 3 rewards, the distance to the target and the energy consumption cost are conflicting. A policy is considered valid if it receives an average discontented return of the joint stuck penalty less than -0.05 .

2) *LunarLanderContinuous*: The goal of this task is to control the two engines of a rocket to land safely on the ground. The reward vector is a 4-dimensional vector which includes (1) a shaping reward given when the rocket is upright and moving to the center of the landing area, (2) an energy consumption cost for the main engine, (3) an energy consumption cost for the side engine, and (4) a landing reward which can be positive or negative depending on the success or failure of the landing task. There is a weak conflict between the reward of shaping and the energy consumption cost for the engines. A policy is considered to be valid, i.e., not crashing, in case it receives an average discontented return of the landing reward greater than 20.

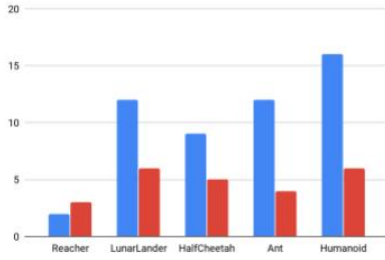
3) *HalfCheetah*, *Ant* and *Humanoid*: To study the suitability of the meta-MORL method to solve more complex RL problems, we prepared a set of high-dimensional locomotion tasks using agents with different DoFs. The goal of the tasks is to control different agents to move forward. The agents are a half cheetah (6 DoFs), an ant (8 DoFs) and a humanoid robot (17 DoFs). All the agents use 5-dimensional reward vector consisting of (1) being alive, i.e., staying in an upright position, (2) gaining forward speed, (3) energy consumption penalty, (4) joint limit penalty, and (5) collision penalty. For every time step, the agent receives a fixed alive reward if it stays upright. The forward speed reward is proportional to the forward speed (negative when moves backward). The energy consumption penalty is proportional to the amount of forces applied to all joints. The joint limit penalty punishes the agent by a fixed negative value when a joint is positioned outside its limit. The collision cost punishes the agent for self collision. Among these rewards, the speed and the energy consumption rewards are conflicting, and cannot be optimized at the same time. A policy is valid if it keeps the agent upright. This corresponds to an average discontented return of the alive reward greater than 80 for the ant and the half cheetah, and 150 for the humanoid.

B. Results

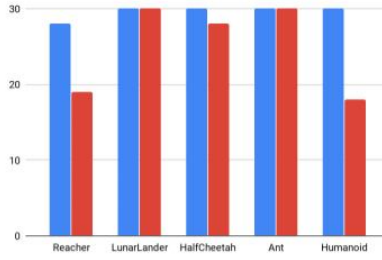
Here, we present meta-MORL results to obtain the Pareto optimal policies and compare it with RA [9] as the baseline. As described in Sec. II, RA is a single-policy approach that estimates the Pareto front using multiple runs each of which solves a single-objective RL problem found by the scalarization function with different weights. The starting policy of RA is initialized randomly. Meta-MORL method work similar to RA but with the difference that, instead of randomly initializing the policy, the policy is initialized with the meta-policy. The meta-policy is trained as explained in the previous section with 30 random scalarization functions.

1) *Estimating the Pareto front*: We use the hypervolume indicator ([15]) to evaluate the quality of the Pareto fronts estimated by the two methods. The hypervolume indicator calculates the volume encapsulated between a reference point and the Pareto front points. A larger hypervolume indicates a more optimal Pareto front. The reference point is defined as the minimum value of each objective found by all policies. The hypervolume of the Pareto front constructed by the two methods are listed in Table I. The results show that, for simple tasks, like LunarLander and Reacher, the performance of RA is equal or slightly better than the meta-MORL. However for complex tasks, the meta-MORL finds solutions resulting in larger hypervolumes, i.e., more optimal Pareto fronts.

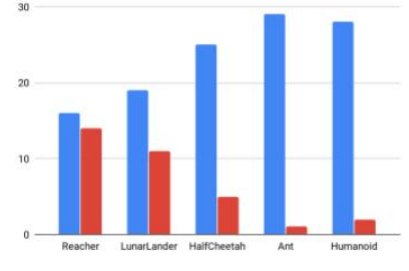
To analyze the optimality of the Pareto front for each task, we only demonstrated three of the objectives in Fig. 3. We evaluated the dominance of each policy compared to all other policies and marked the non-dominated policies with black circles. The number of non-dominated policies found by the two methods is given in Fig. 2a. Considering both the hypervolume and the number of non-dominated



(a) The number of non-dominated policies found by two methods



(b) The number of valid policies found by two methods



(c) The number of policies that achieves higher scalarized return given a weight vector.

Fig. 2: Different policies found by the two methods. The blue bars denote the results of meta-MORL and the red bars denote the result of RA.

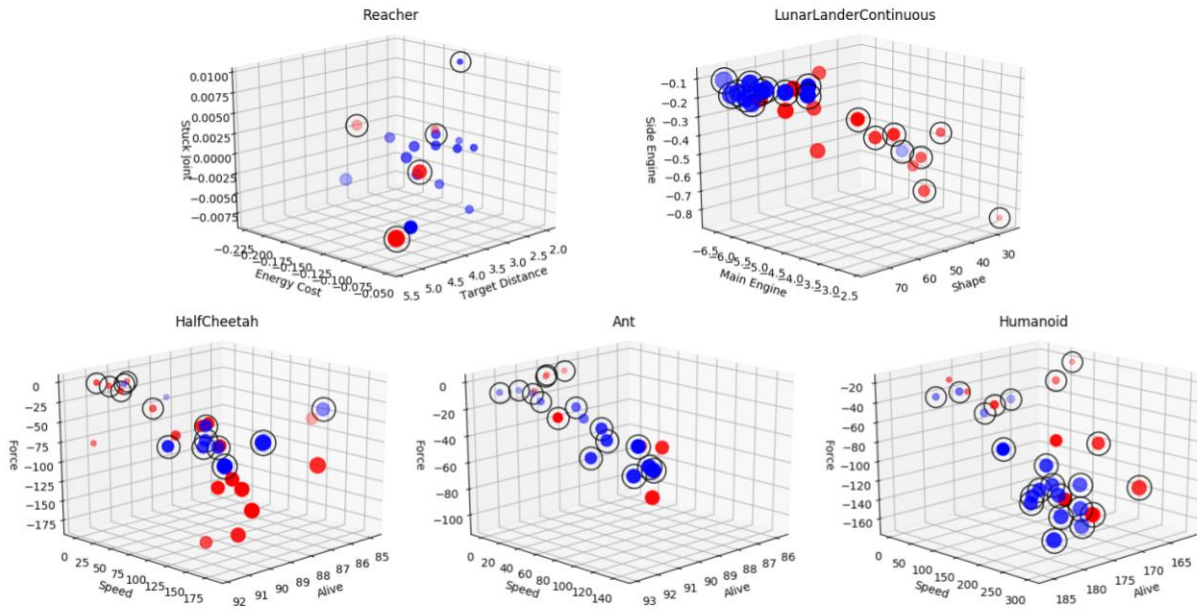


Fig. 3: The Pareto optimal policies found by the two methods. For tasks with more than three objectives, we only illustrate the three most informative ones. The blue points denote the policies found by the meta-MORL, and the red points denote the policies found by the baseline. The points with black circle denote the non-dominated policies. The size of points is proportional to the distance to the origin to better visualize in 3D.

policies measures confirms that meta-MORL outperforms the baseline in complex continuous control tasks.

2) *Data-efficiency*: Here, we demonstrate the superiority of the proposed method to construct the Pareto front using less data. Fig. 4 illustrates the hypervolume measure calculated by the meta-MORL method during the fine-tuning phase. The red dashed lines denote the hypervolume found by the baseline method when all policies are converged. To compare the data-efficiency of the methods, we observed how much data is used in each case.

Baseline: Each policy is trained with 1K iterations for Reacher and LunarLanderContinuous, 2K iterations for HalfCheetah and Ant, and 4K iterations for Humanoid. In

total 30 different policies are trained for each task. Therefore, as an example, for the Humanoid task 120K iterations are collected to obtain the hypervolume indicated by the red dashed line.

Meta-MORL: The meta-policy is trained by the following number of meta-updates, i.e., the number of iterations for the adaptation and the meta-learning phases: 400 for Reacher and LunarLanderContinuous, 200 for HalfCheetah, 500 for Ant and 690 for Humanoid. In our implementation, a meta-update requires five times more training data compared to a regular update. Therefore, as an example, with $(690 \times 5 + 70) \times 30 \sim 100K$ of iterations, meta-MORL reaches similar performance as achieved by 120K iterations of the

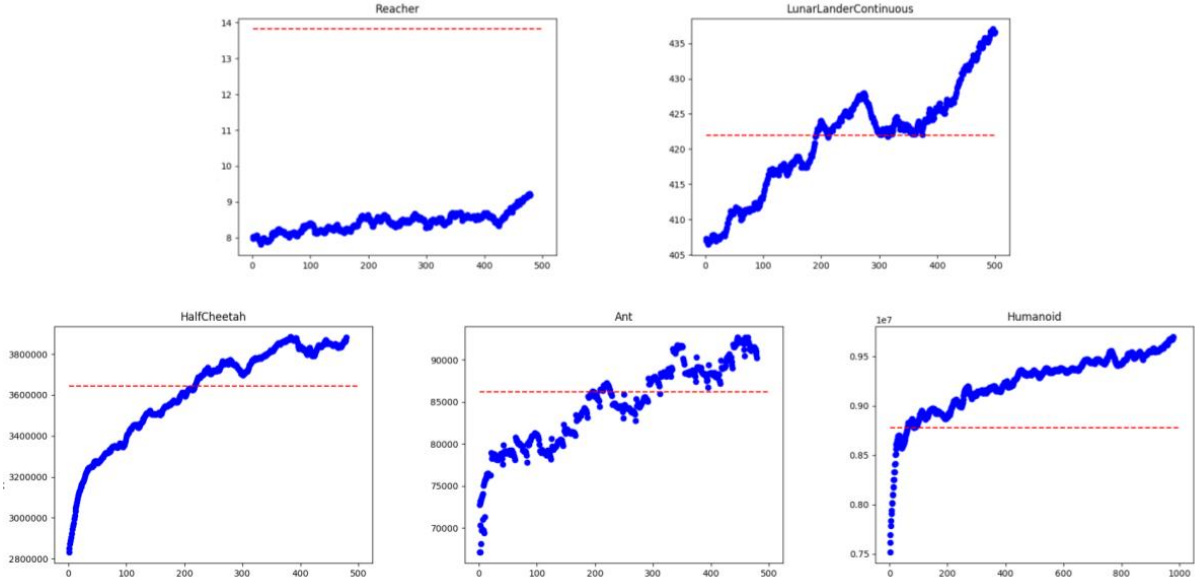


Fig. 4: The improvements of the hypervolume indicator (vertical axis) with respect to the iteration of fine-tuning the meta policy (horizontal axis). The blue curve denotes the hypervolume and the red line denotes the final hypervolume of the Pareto front estimated by RA.

baseline method. Furthermore, meta-MORL keeps improving the results with more updates, i.e., almost 10% improvement is achieved by 30K more iterations, while no more improvement can be observed by further training using the baseline method.

3) *The performance of individual policies:* We also studied the performance of each policy w.r.t. its corresponding single objective RL problem given by the scalarization function. For different scalarization functions, we compared meta-MORL and the baseline by enumerating the valid policies obtained by each method. Also, for each method, we counted the number of the policies that outperform the other method for the same scalar objective. In this respect, the validity of a policy indicates whether the policy can accomplish a task and the accumulated scalar return indicates its performance for that specific task. These measures are illustrated in Fig. 2b and Fig. 2c, respectively. As it is shown, in both cases, meta-MORL outperform the baseline, with greater margins for more complicated control tasks.

It is not surprising that meta-MORL outperforms the baseline in this case. Assume that for the Humanoid task, we would like to learn a policy that minimizes the energy cost as the most important objective. In this case, training a randomly initialized policy will most likely result in a policy that always causing the agent to fall down. Since, this is the most likely local optimal behavior when the policy is initialized randomly. However, using meta-learning, the meta-policy is found such that the combination of different objectives can be achieved by few more training iterations. Therefore, in the above example, the policy can find a more optimal behavior that minimizes the energy consumption cost and still walks forward. In that sense, we can say that the

TABLE I: Hypervolume indicator of the final Pareto front estimated by the two methods

Environment	Meta-MORL	RA
Reacher	9.95	13.82
LunarLander	453	422
HalfCheetah	3.94×10^6	3.64×10^6
Ant	10.4×10^4	8.62×10^4
Humanoid	9.61×10^6	8.78×10^4

meta-learning method results in a more efficient exploration strategy to acquire optimal behaviors.

VII. CONCLUSIONS

In this work, we introduced a novel meta-MORL approach based on model-agnostic meta-learning [8] to solve deep-MORL problems. We proposed to convert a MORL problem to a number of single-objective RL tasks using a parametric scalarization function. Then, using meta-learning, a meta-policy is obtained such that the performance on each task can be improved by few more training iterations. The meta-policy is finally used as the optimal initial policy to train a set of Pareto optimal policies with different objectives.

We evaluated our method in several simulated continuous control tasks and demonstrated that it scales well to high-dimensional control problems. Furthermore, we demonstrated that it outperforms the only baseline method that can be applied to train deep policies to construct the Pareto front for reward vectors with several dimensions.

VIII. ACKNOWLEDGMENTS

This work is supported by the European Unions Horizon 2020 research and innovation program, the CENTAURO

project (under grant agreement No. 644839), the soc-SMCs project (H2020-FETPROACT-2014), and also by the Academy of Finland through the DEEPEN project.

REFERENCES

- [1] M. Riedmiller, R. Hafner, T. Lampe, M. Neunert, J. Degraeve, T. Van de Wiele, V. Mnih, N. Heess, and J. T. Springenberg, "Learning by playing-solving sparse reward tasks from scratch," *arXiv preprint arXiv:1802.10567*, 2018.
- [2] A. Ghadirzadeh, A. Maki, D. Kragic, and M. Björkman, "Deep predictive policy training using reinforcement learning," in *Intelligent Robots and Systems (IROS), 2017 IEEE/RSJ International Conference on*. IEEE, 2017, pp. 2351–2358.
- [3] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.
- [4] Y. Duan, X. Chen, R. Houthoofd, J. Schulman, and P. Abbeel, "Benchmarking deep reinforcement learning for continuous control," in *International Conference on Machine Learning*, 2016, pp. 1329–1338.
- [5] C. Liu, X. Xu, and D. Hu, "Multiobjective reinforcement learning: A comprehensive overview," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 45, no. 3, pp. 385–398, 2015.
- [6] P. Vamplew, R. Dazeley, A. Berry, R. Issabekov, and E. Dekker, "Empirical evaluation methods for multiobjective reinforcement learning algorithms," *Machine learning*, vol. 84, no. 1-2, pp. 51–80, 2011.
- [7] D. M. Roijers, P. Vamplew, S. Whiteson, and R. Dazeley, "A survey of multi-objective sequential decision-making," *Journal of Artificial Intelligence Research*, vol. 48, pp. 67–113, 2013.
- [8] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," *arXiv preprint arXiv:1703.03400*, 2017.
- [9] S. Parisi, M. Pirota, N. Smacchia, L. Bascetta, and M. Restelli, "Policy gradient approaches for multi-objective sequential decision making," in *Neural networks (ijcnn), 2014 international joint conference on*. IEEE, 2014, pp. 2323–2330.
- [10] K. Van Moffaert and A. Nowé, "Multi-objective reinforcement learning using sets of pareto dominating policies," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 3483–3512, 2014.
- [11] T. Brys, A. Harutyunyan, P. Vrancx, A. Nowé, and M. E. Taylor, "Multi-objectivization and ensembles of shapings in reinforcement learning," *Neurocomputing*, vol. 263, pp. 48–59, 2017.
- [12] D. J. Lizotte, M. H. Bowling, and S. A. Murphy, "Efficient reinforcement learning with multiple reward functions for randomized controlled trial analysis," in *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*. Citeseer, 2010, pp. 695–702.
- [13] I. Das and J. E. Dennis, "A closer look at drawbacks of minimizing weighted sums of objectives for pareto set generation in multicriteria optimization problems," *Structural optimization*, vol. 14, no. 1, pp. 63–69, 1997.
- [14] K. Van Moffaert, M. M. Drugan, and A. Nowé, "Scalarized multi-objective reinforcement learning: Novel design techniques," in *AD-PRL*, 2013, pp. 191–199.
- [15] K. Van Moffaert, M. M. Drugan, and A. Nowé, "Hypervolume-based multi-objective reinforcement learning," in *International Conference on Evolutionary Multi-Criterion Optimization*. Springer, 2013, pp. 352–366.
- [16] S. Natarajan and P. Tadepalli, "Dynamic preferences in multi-criteria reinforcement learning," in *Proceedings of the 22nd international conference on Machine learning*. ACM, 2005, pp. 601–608.
- [17] M. Pirota, S. Parisi, and M. Restelli, "Multi-objective reinforcement learning with continuous pareto frontier approximation," in *29th AAAI Conference on Artificial Intelligence, AAAI 2015 and the 27th Innovative Applications of Artificial Intelligence Conference, IAAI 2015*. AAAI Press, 2015, pp. 2928–2934.
- [18] S. Parisi, M. Pirota, and J. Peters, "Manifold-based multi-objective policy search with sample reuse," *Neurocomputing*, vol. 263, pp. 3–14, 2017.
- [19] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz, "Trust region policy optimization," in *International Conference on Machine Learning*, 2015, pp. 1889–1897.
- [20] "Roboschool: open-source software for robot simulation," <https://blog.openai.com/roboschool/>.
- [21] M. Plappert, M. Andrychowicz, A. Ray, B. McGrew, B. Baker, G. Powell, J. Schneider, J. Tobin, M. Chociej, P. Welinder, V. Kumar, and W. Zaremba, "Multi-goal reinforcement learning: Challenging robotics environments and request for research," 2018.