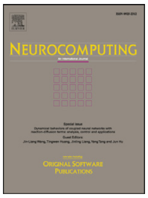
Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom



Manifold-based multi-objective policy search with sample reuse

S. Parisi[a,*], M. Pirotta[b], J. Peters[a,c]

[a] *Technische Universität Darmstadt, Hochschulstr. 10, Darmstadt 64289, Germany*
[b] *Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133 Milano, Italy*
[c] *Max Planck Institute for Intelligent Systems, Spemannstr. 41, Tübingen 72076, Germany*

ARTICLE INFO

Article history:
Received 28 January 2016
Revised 18 September 2016
Accepted 1 November 2016
Available online xxx

Keywords:
Multi-objective
Reinforcement learning
Policy search
Black-box optimization
Importance sampling

ABSTRACT

Many real-world applications are characterized by multiple conflicting objectives. In such problems optimality is replaced by Pareto optimality and the goal is to find the Pareto frontier, a set of solutions representing different compromises among the objectives. Despite recent advances in multi-objective optimization, achieving an accurate representation of the Pareto frontier is still an important challenge. Building on recent advances in reinforcement learning and multi-objective policy search, we present two novel manifold-based algorithms to solve multi-objective Markov decision processes. These algorithms combine episodic exploration strategies and importance sampling to efficiently learn a manifold in the policy parameter space such that its image in the objective space accurately approximates the Pareto frontier. We show that episode-based approaches and importance sampling can lead to significantly better results in the context of multi-objective reinforcement learning. Evaluated on three multi-objective problems, our algorithms outperform state-of-the-art methods both in terms of quality of the learned Pareto frontier and sample efficiency.

1. Introduction

Many real-world problems are characterized by the presence of multiple conflicting objectives, such as economic systems [1], medical treatment [2], control of robots [3,4], water reservoirs [5] and elevators [6]. These applications can be modeled as multi-objective reinforcement learning (MORL) problems, where the standard notion of optimality is replaced by *Pareto optimality*, a concept for representing compromises among the objectives. Despite the increasing interest in multi-objective problems and recent advances in reinforcement learning, MORL is still a relatively young field of research.

MORL approaches can be classified in two main categories [7] based on the number of policies they learn: single policy and multiple policy. While the majority of MORL approaches belong to the former category, in this paper we focus on the latter and aim to learn a set of policies representing the best compromises among the objectives, namely the *Pareto frontier*. Providing an *accurate* and *uniform* representation of the complete Pareto frontier is often beneficial. It encapsulates all the trade-offs among the objectives and gives better insight into the problem, thus helping the a posteriori selection of the most favorable solution.

Following the same line of thoughts of RL, initially MORL researchers have focused on the development of value function-based approaches, where the attention was posed on the recovery of the optimal value function (for more details, we refer to the survey in [8]). Recently[1], policy search approaches have also been extended to multi-objective problems [9,10]. However, the majority of MORL approaches perform exploration in the action space [11]. This strategy, commonly known as *step-based*, requires a different exploration noise at each time step and many studies [12,13] have shown that it is subject to several limitations, primarily due to the high variance in the policy update. Furthermore, common algorithms involve the solution of several (independent) single-objective problems in order to approximate the Pareto frontier [9,14–16]. This approach implies an inefficient use of the samples, as each optimization is usually carried out on-policy, and most of MORL state-of-the-art approaches are inapplicable to large problems, especially in the presence of several objectives.

In this paper, we address these limitations and present the first manifold-based episodic algorithms in MORL literature. First, these algorithms follow an *episodic* exploration strategy (also known as *parameter-based* or *black-box*) in order to reduce the variance dur-

* Corresponding author.
E-mail addresses: parisi@ias.tu-darmstadt.de, simone@robot-learning.de, simone.parisi@mail.polimi.it (S. Parisi), matteo.pirotta@polimi.it (M. Pirotta), mail@jan-peters.net (J. Peters).

[1] The first seminal work dates back to 2001 [1].

http://dx.doi.org/10.1016/j.neucom.2016.11.094
0925-2312/© 2017 Elsevier B.V. All rights reserved.

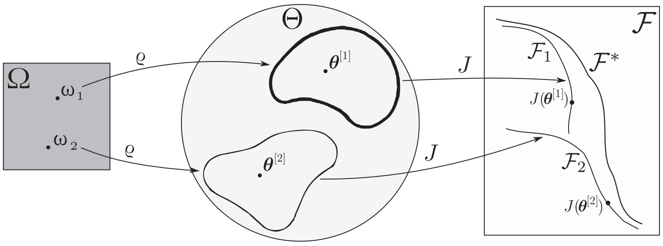


Fig. 1. Transformation map from the high-level distribution ϱ to approximate frontiers in the objective space $\mathcal{F}$. A high-level parameter vector $\boldsymbol{\omega}_i$ maps to a manifold in the policy parameter space $\boldsymbol{\Theta}$. Subsequently, the manifold maps to an approximate frontier $\mathcal{F}_i$, with each vector $\boldsymbol{\theta}^{[j]}$ mapping to a return vector $\boldsymbol{J}(\boldsymbol{\theta}^{[j]})$.

ing the policy update. Second, they perform a *manifold-based* policy search and directly learn a manifold in the policy parameter space to generate infinitely many Pareto-optimal solutions in a single run. By employing *Pareto-optimal* indicator functions, the algorithms are guaranteed to accurately and uniformly approximate the Pareto frontier. Finally, we show how to incorporate *importance sampling* in order to further reduce the sample complexity and to extend these algorithms to the *off-policy* paradigm. To the best of our knowledge, our algorithms are the first ones to tackle all these issues at once.

The remainder of the paper is organized as follows. In Section 2, we introduce the multi-objective problem and discuss related work in MORL literature. Section 3 includes the main contributions of this paper: an episodic manifold-based reformulation of the multi-objective problem, two policy search algorithms and two Pareto-optimal indicatorfunctions to solve it, and an extension to importance sampling for reusing past samples. Section 4 provides a thorough empirical evaluation of the proposed algorithms on three problems, namely a water reservoir control task, a linear-quadratic Gaussian regulator and a simulated robot tetherball game. Finally, in Section 5 we discuss the results of this study and propose possible avenues of investigation for future research.

2. Preliminaries

In this section, we provide the mathematical framework and the terminology as used in this paper. Moreover, we present a categorization of the multi-objective approaches presented in MORL literature and we briefly discuss their advantages and drawbacks.

2.1. Problem statement and notation

Multi-objective Markov decision processes (MOMDPs) are an extension of MDPs in which several pairs of reward functions and discount factors are defined, one for each objective. Formally, a MOMDP is described by a tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \boldsymbol{\mathcal{R}}, \boldsymbol{\gamma}, \mathcal{D} \rangle$: $\mathcal{S} \subseteq \mathbb{R}^{d_s}$ is a continuous state space, $\mathcal{A} \subseteq \mathbb{R}^{d_a}$ is a continuous action space, $\mathcal{P}$ is a Markovian transition model and $\mathcal{P}(s'|s, a)$ defines the transition density between state s and s' under action a, $\boldsymbol{\mathcal{R}} = \left[\mathcal{R}_1, \dots, \mathcal{R}_{d_{\mathrm{R}}}\right]^\top$ and $\boldsymbol{\gamma} = \left[\gamma_1, \dots, \gamma_{d_{\mathrm{R}}}\right]^\top$ are vectors of reward functions $\mathcal{R}_i : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ and discount factors $\gamma_i \in [0, 1)$, respectively, and $\mathcal{D}$ is the initial state distribution.

The *policy* followed by the agent is described by a conditional distribution $\pi(a|s)$ specifying the probability of taking action a in state s. In MOMDPs, a policy π is associated to d_{R} expected returns $\boldsymbol{J}^\pi = \left[J_1^\pi, \dots, J_{d_{\mathrm{R}}}^\pi\right] \in \mathcal{F}$, where $\mathcal{F} \subseteq \mathbb{R}^{d_{\mathrm{R}}}$ is the policy performance space. Using the trajectory-based definition, the ith expected return is

$$J_i^\pi = \mathbb{E}_{\boldsymbol{\tau} \sim p(\cdot|\pi)}[R_i(\boldsymbol{\tau})],$$

where $\boldsymbol{\tau} = \{s_t, a_t\}_{t=1}^{H_\tau} \in \boldsymbol{T}$ is a trajectory (episode) of length $H_{\boldsymbol{\tau}}$ (possibly infinite) drawn from the distribution $p(\boldsymbol{\tau}|\pi)$, with return $R_i(\boldsymbol{\tau}) = \sum_{t=1}^{H_\tau} \gamma_i^{t-1} \mathcal{R}_i(s_t, a_t)$. Since it is not common to have multiple discount factors (the problem becomes NP-complete [17]), we consider a unique value γ for all the objectives.

Unlike in single-objective MDPs, in MOMDPs a single policy dominating all others usually does not exist. When conflicting objectives are considered, no policy can simultaneously maximize all of them. For this reason, in multi-objective optimization a different dominance concept based on Pareto optimality is used. A policy π *strongly dominates* a policy π' (denoted by $\pi \succ \pi'$) if it outperforms π' on all objectives, i.e.,

$$\pi \succ \pi' \iff \forall i \in \{1, \dots, d_{\mathrm{R}}\}, J_i^\pi > J_i^{\pi'}.$$

Similarly, policy π *weakly dominates* policy π' (which is denoted by $\pi \succeq \pi'$) if it is not worse on all objectives, i.e.,

$$\forall i \in \{1, \dots, d_{\mathrm{R}}\}, J_i^\pi \ge J_i^{\pi'} \ \wedge \ \exists i \in \{1, \dots, d_{\mathrm{R}}\}, J_i^\pi = J_i^{\pi'}.$$

If there is no policy π' such that $\pi' \succ \pi$, then the policy π is *Pareto-optimal*. We can also speak of *locally Pareto-optimal* policies, for which the definition is the same as above, except that we restrict the dominance to a neighborhood of π.

Our goal is to determine the set of all Pareto-optimal policies $\Pi^* = \left\{\pi \mid \nexists \pi', \ \pi' \succ \pi\right\}$, which maps to the so-called *Pareto frontier* $\mathcal{F} = \left\{\boldsymbol{J}^{\pi^*} \mid \pi^* \in \Pi^*\right\}$.[2] More specifically, in this paper we consider *parametric* policies $\pi \in \boldsymbol{\Pi}^{\boldsymbol{\theta}} \equiv \{\pi_{\boldsymbol{\theta}} \mid \boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^{d_{\boldsymbol{\theta}}}\}$, where $\boldsymbol{\Theta}$ is the *policy parameters space*. For simplicity, we will use $\boldsymbol{\theta}$ in place of $\pi_{\boldsymbol{\theta}}$ to denote the dependence on the current policy, e.g.$\boldsymbol{J}(\boldsymbol{\theta})$ instead of $\boldsymbol{J}^{\pi_{\boldsymbol{\theta}}}$.

2.2. Related work

MORL approaches can be divided into two categories based on the number of policies they learn [7]. *Single-policy* methods aim to find the best policy satisfying a preference among the objectives. The majority of MORL approaches belong to this category and differ for the way in which preferences are expressed. They are easy to implement, but require a priori decision about the type of the solution and suffer from instability, as small changes on the preferences may result in significant variation in the solution [7]. The most straightforward and common single-policy approach is the scalarization where a function is applied to the reward vector in order to produce a scalar signal. Usually, a linear combination (weighted sum) of the rewards is performed and the weights are used to express the preferences over multiple objective [16,19,20]. Less common is the use of non linear mapping [21]. Although scalarization approaches are simple and intuitive, they may fail in obtaining MOO desiderata, e.g.,a uniform distribution of the weights may not produce accurate and evenly distributed points on the Pareto frontier [22]. On the other hand, several issues of the scalarization are alleviated in RL due to the fact that the Pareto frontier is convex when stochastic policies are considered [8,23]. For example, the convex hull of stochastic policies, each one being optimal w.r.t.a different linear scalarization, represents a viable approximation of the Pareto frontier[3]. Different single-policy approaches are based on thresholds and lexicographic ordering [14] or different kinds of preferences over the objective space [24,25].

Multiple-policy approaches, on the contrary, aim at learning multiple policies in order to approximate the Pareto frontier. Build-

[2] As done in [18], we suppose that locally Pareto-optimal solutions that are not Pareto-optimal do not exist.

[3] In episodic tasks, we can even exploit deterministic optimal policies by constructing mixture policies, i.e., policies stochastically choosing between deterministic policies at the beginning of each episode.