Multi–objective Reinforcement Learning with Continuous Pareto Frontier Approximation

Matteo Pirotta and **Simone Parisi** and **Marcello Restelli**
Department of Electronics, Information and Bioengineering, Politecnico di Milano,
Piazza Leonardo da Vinci, 32, 20133, Milan, Italy
matteo.pirotta@polimi.it, simone.parisi@mail.polimi.it, marcello.restelli@polimi.it

Abstract

This paper is about learning a continuous approximation of the Pareto frontier in Multi–Objective Markov Decision Problems (MOMDPs). We propose a policy–based approach that exploits gradient information to generate solutions close to the Pareto ones. Differently from previous policy–gradient multi–objective algorithms, where n optimization routines are used to have n solutions, our approach performs a single gradient–ascent run that at each step generates an improved continuous approximation of the Pareto frontier. The idea is to exploit a gradient–based approach to optimize the parameters of a function that defines a manifold in the policy parameter space so that the corresponding image in the objective space gets as close as possible to the Pareto frontier. Besides deriving how to compute and estimate such gradient, we will also discuss the non–trivial issue of defining a metric to assess the quality of the candidate Pareto frontiers. Finally, the properties of the proposed approach are empirically evaluated on two interesting MOMDPs.

Introduction

Many real-world control problems (e.g., economic systems, water resource problems, robotic systems, just to mention a few) are characterized by the presence of multiple, conflicting objectives. Such problems are often modeled as Multi-Objective Markov Decision Processes (MOMDPs), where the concept of optimality typical of MDPs is replaced by the one of Pareto optimality, i.e., a set of policies providing a compromise among the different objectives. In the last decades, Reinforcement Learning (RL) (Sutton and Barto 1998) has established as an effective and theoretically-grounded framework that allows to solve single-objective MDPs whenever either no (or little) prior knowledge is available about system dynamics, or the dimensionality of the system to be controlled is too high for classical optimal control methods. Despite the successful developments in RL theory and a high demand for multi-objective control applications, Multi-Objective Reinforcement Learning (MORL) (Roijers et al. 2013) is still a relatively young and unexplored research topic.

MORL approaches can be divided into two main categories, based on the number of policies they learn (Vamplew et al. 2011): single policy and multiple policy. Although, the majority of MORL approaches belong to former category, in this paper, we focus on latter approaches. Multiple-policy approaches aim at learning a set of policies in order to approximate the Pareto frontier. When the number d of decision variables (i.e., policy parameters) is greater than or equal to the number q of objectives, the local Pareto–optimal solutions form a $(q-1)$–dimensional manifold (Harada et al. 2007). The superiority of multiple–policy methods resides in their ability to represent the Pareto–optimal manifold, allowing a posteriori selection of the solution, through a graphical representation of the frontier that can give a better insight into the relationships among the objectives, and encapsulate all the trade-offs among the multiple objectives. Since the exact derivation of the Pareto frontier is generally impractical in real-world problems, the goal is to compute an approximation of the Pareto frontier that includes solutions that are accurate, evenly distributed, and covering a range similar to the one of the actual front (Zitzler et al. 2003). Among multiple–policy algorithms it is possible to identify two classes: value–based (Lizotte, Bowling, and Murphy 2012; Castelletti, Pianosi, and Restelli 2013) and gradient approaches (Shelton 2001; Parisi et al. 2014). While value–based approaches suffer from curse of dimensionality and have difficulties with continuous action spaces, gradient–based techniques lack of guarantees about the uniformity of the coverage of the Pareto frontier.

In this paper, we propose a novel gradient–based MORL approach named Policy Manifold Gradient Algorithm (PMGA) that generates a continuous approximation of the local Pareto–optimal solution manifold in the policy space. We exploit a parametric function to generate a manifold in the policy parameter space, which maps to the objective space through the expected return function. The goal is to find the parameters that induce a frontier as close as possible to the Pareto one. Exploiting this approximation it is possible to generate an arbitrarily dense representation of the Pareto frontier. The main contributions of this paper are: the derivation of the gradient approach in the general case – i.e., independent from the metric used to measure the quality of the current solution– , how to estimate such gradient from sample trajectories, a discussion of frontier quality measures that can be effectively integrated in the proposed gradient

approach, and an empirical evaluation of its performance in a multi–objective extension of the Linear-Quadratic Gaussian regulator and in a water reservoir management domain.

Preliminaries

Multi–Objective Markov Decision Processes (MOMDPs) are an extension of the MDP model, where several pairs of reward functions and discount factors are defined, one for each objective. Formally, a MOMDP is described by a tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathbf{R}, \boldsymbol{\gamma}, D \rangle$, where $\mathcal{S} \subseteq \mathbb{R}^n$ is the continuous state space, $\mathcal{A} \subseteq \mathbb{R}^m$ is the continuous action space, $\mathcal{P}$ is a Markovian transition model where $\mathcal{P}(s'|s,a)$ defines the transition density between state s and s' under action a, $\mathbf{R} = [\mathcal{R}_1, \ldots, \mathcal{R}_q]^\intercal$ and $\boldsymbol{\gamma} = [\gamma_1, \ldots, \gamma_q]^\intercal$ are q-dimensional column vectors of reward functions $\mathcal{R}_i : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ and discount factors $\gamma_i \in [0,1)$, respectively, and D is a distribution from which the initial state is drawn. In MOMDPs, any policy π is associated to q expected returns $\mathbf{J}^\pi = \left[J_1^\pi, \ldots, J_q^\pi\right]$:

$$J_i^\pi = \mathbb{E}\left[\sum_{t=0}^{T} \gamma_i^t r_i(t+1) | x_0 \sim D, \pi\right],$$

being $r_i(t+1) = \mathcal{R}_i(s_t, a_t, s_{t+1})$ the i-th immediate reward obtained when state s_{t+1} is reached from state s_t and action a_t, and T the finite or infinite horizon.

In policy–gradient approaches, a parametrized space of policies $\Pi_{\boldsymbol{\theta}} = \left\{\pi_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^d\right\}$ (where $\pi_{\boldsymbol{\theta}}$ is a compact notation for $\pi(a|s,\boldsymbol{\theta})$) is considered. Given a policy parametrization $\boldsymbol{\theta}$, we assume that the policy performance $\mathbf{J} : \Theta \to \mathbb{R}^q$ is at least of class C^2.

$\mathbf{J}$ is defined as the expected reward over the space of all possible trajectories $\mathbb{T}$: $\mathbf{J}(\boldsymbol{\theta}) = \int_{\mathbb{T}} p(\tau|\boldsymbol{\theta})\, \mathbf{r}(\tau) \mathrm{d}\tau$, where $\tau \in \mathbb{T}$ is a trajectory drawn from density distribution $p(\tau|\boldsymbol{\theta})$ and $\mathbf{r}(\tau)$ represents the accumulated expected discounted reward over trajectory τ: $\mathbf{r}_i(\tau) = \sum_{t=0}^{T} \gamma_i^t r_i(t+1)$.

In MOMDPs for each policy parameter $\boldsymbol{\theta}$, q gradient directions are defined (Peters and Schaal 2008)

$$\nabla_{\boldsymbol{\theta}} J_i(\boldsymbol{\theta}) = \int_{\mathbb{T}} \nabla_{\boldsymbol{\theta}} p(\tau|\boldsymbol{\theta})\, \mathbf{r}_i(\tau) \mathrm{d}\tau = \mathop{\mathbb{E}}_{\tau \in \mathbb{T}}\left[\nabla_{\boldsymbol{\theta}} \log p(\tau|\boldsymbol{\theta})\, \mathbf{r}_i(\tau)\right]$$
$$= \mathbb{E}_{\tau \in \mathbb{T}}\left[\mathbf{r}_i(\tau) \sum_{t=1}^{T} \nabla_{\boldsymbol{\theta}} \log \pi(a_t^\tau | s_t^\tau, \boldsymbol{\theta})\right],$$

where each direction $\nabla_{\boldsymbol{\theta}} J_i$ is associated to a particular discount factor–reward function pair $< \gamma_i, \mathcal{R}_i >$. As shown in previous equation, the differentiability of the performance measure is connected to the differentiability of the policy class by: $\nabla_{\boldsymbol{\theta}} \log p(\tau|\boldsymbol{\theta}) = \sum_{k=1}^{T} \nabla_{\boldsymbol{\theta}} \log \pi(a_k|s_k, \boldsymbol{\theta})$.

Despite what happens in MDPs, in MOMDPs a single policy which dominates all the others usually does not exist; in fact, when conflicting objectives are considered, no policy can simultaneously maximize all the objectives. For these reasons, in Multi-Objective Optimization (MOO) a different dominance concept is used. Policy π *dominates* policy π', which is denoted by $\pi \succ \pi'$, if:

$$\forall i \in \{1, \ldots, q\}, J_i^\pi \geq J_i^{\pi'} \ \wedge \ \exists i \in \{1, \ldots, q\}, J_i^\pi > J_i^{\pi'}.$$

If there is no policy π' such that $\pi' \succ \pi$, the policy π is *Pareto–optimal*. In general, there are multiple Pareto-optimal policies. Solving a MOMDP means finding the set of Pareto-optimal policies $\Pi^* = \left\{\pi \,|\, \nexists \pi', \pi' \succ \pi\right\}$, which maps to the so-called Pareto frontier $\mathcal{F}^* = \left\{\mathbf{J}^{\pi^*} | \pi^* \in \Pi^*\right\}$.[1]

A remark on notation. In the following we will use the symbol $D_X F$ to denote the derivative [2] of a generic function $F : \mathbb{R}^{m \times n} \to \mathbb{R}^{p \times q}$ w.r.t. matrix X. Notice that the following relationship holds for scalar functions of vector variable: $\nabla_x f = (D_x f)^\intercal$. Finally, the symbol I_x will be used to denote an $x \times x$ identity matrix.

Gradient on Policy Manifold for Continuous Pareto Front Approximation

Most of MORL approaches proposed so far produce discrete approximations of the Pareto frontier. While such solutions can be effective in finite MOMDPs, where the Pareto frontier can be obtained by considering the convex hull of a finite set of deterministic stationary policies (Roijers et al. 2013), this is not the case for continuous MOMDPs. In this paper we aim to build a continuous approximation of the Pareto frontier exploiting a gradient–based method. Differently from other MORL approaches (Shelton 2001; Parisi et al. 2014), where policy–gradient methods are used to generate a finite set of policies as close as possible to the Pareto frontier, here we consider a parametrized function $\phi_{\boldsymbol{\rho}}$ that– by defining a manifold in the policy parameter space–generates a continuous set of policies. Finally, these points map to a curve in the objective space through the expected return $\mathbf{J}(\boldsymbol{\theta})$ (see Figure 1). The goal is to find the parametrization $\boldsymbol{\rho}$ that provides the best approximation of the Pareto frontier. In the remainder of this section, we will formalize the PMGA approach.

Let $\mathcal{T}$ be open in $\mathbb{R}^b$ with $b \leq q$ and let $\phi_{\boldsymbol{\rho}} : \mathcal{T} \to \Theta$ be a smooth map of class $C^l (l \geq 1)$, where $\mathbf{t} \in \mathcal{T}$ and $\boldsymbol{\rho} \in P \subseteq \mathbb{R}^k$ are the free variables and the parameters, respectively. We think of the map $\phi_{\boldsymbol{\rho}}$ as a parameterization of the subset $\phi_{\boldsymbol{\rho}}(\mathcal{T})$ of Θ: each choice of a point $\mathbf{t} \in \mathcal{T}$ gives rise to a point $\phi_{\boldsymbol{\rho}}(\mathbf{t})$ in $\phi_{\boldsymbol{\rho}}(\mathcal{T})$. This means that only a subset $\Theta_{\boldsymbol{\rho}}(\mathcal{T})$ of the space Θ can be spanned by map $\phi_{\boldsymbol{\rho}}$, i.e., $\Theta_{\boldsymbol{\rho}}(\mathcal{T})$ is a b–dimensional parametrized manifold (Munkres 1997) in the policy parameter space

$$\Theta_{\boldsymbol{\rho}}(\mathcal{T}) = \left\{\boldsymbol{\theta} : \boldsymbol{\theta} = \phi_{\boldsymbol{\rho}}(\mathbf{t}), \forall \mathbf{t} \in \mathcal{T}\right\},$$

and, as a consequence, the associated parametrized Pareto frontier is the b–dimensional open set defined as

$$\mathcal{F}_{\boldsymbol{\rho}}(\mathcal{T}) = \left\{\mathbf{J}(\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta_{\boldsymbol{\rho}}(\mathcal{T})\right\}.$$

This manifold represents our approximation of the true Pareto frontier. The goal is to find the best approximation, i.e., the parameters $\boldsymbol{\rho}$ that minimize the distance from the real frontier $\boldsymbol{\rho}^* = \arg\min_{\boldsymbol{\rho} \in P} \ \mathcal{I}^*\left(\mathcal{F}_{\boldsymbol{\rho}}(\mathcal{T})\right)$, where $\mathcal{I}^* : \mathbb{R}^q \to \mathbb{R}$ is some loss function that measures the discrepancy between the Pareto–optimal frontier and $\mathcal{F}_{\boldsymbol{\rho}}(\mathcal{T})$. However, since $\mathcal{I}^*$ requires the knowledge of the Pareto frontier, a different indicator function is needed. The definition of such metric is an open problem in literature. Recently (Vamplew

[1] As in (Harada, Sakuma, and Kobayashi 2006), we assume that only global Pareto-optimal solutions exist.

[2] The derivative operator is well defined for matrices, vectors and scalar functions. Details in (Magnus and Neudecker 1999).

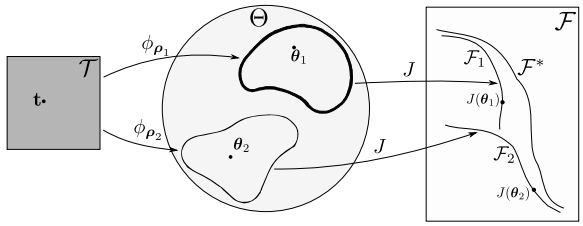


Figure 1: Transformation map corresponding to two different parametrizations $\boldsymbol{\rho}_1$ and $\boldsymbol{\rho}_2$.

et al. 2011), several metrics have been defined, but every candidate presents some intrinsic limits that prevent the definition of a unique superior metric. Furthermore, as we will see in the rest of the paper, the proposed approach needs a metric differentiable w.r.t. policy parameters. We will come back to this topic later.

In general, MOO algorithms compute the value of the frontier as sum of the value of the points composing the discrete approximation. In our scenario, where a continuous frontier approximation is available, it maps to an integration on the Pareto manifold

$$J\left(\boldsymbol{\rho}\right) = \int_{\mathcal{F}(\mathcal{T})} \mathcal{I} \mathrm{d}V,$$

where $\mathrm{d}V$ is a symbol used to denote the integral w.r.t. the volume of the manifold (Munkres 1997) and $\mathcal{I} : \mathcal{F}\left(\mathcal{T}\right) \rightarrow \mathbb{R}$ is a continuous indicator function that for each point of $\mathcal{F}\left(\mathcal{T}\right)$ gives insights about its Pareto–optimality. A standard way to maximize previous equation is to perform gradient ascent, updating the parameters according to the gradient direction: $\boldsymbol{\rho}_{t+1} = \boldsymbol{\rho}_t + \alpha_t \, \nabla_{\boldsymbol{\rho}} J\left(\boldsymbol{\rho}\right)$. The gradient is provided by the following theorem.

Theorem 1. *Let $\mathcal{T}$ be an open set in $\mathbb{R}^b$, let $\mathcal{F}_{\boldsymbol{\rho}}\left(\mathcal{T}\right)$ be a manifold parametrized by a smooth map expressed as composition of maps $\mathbf{J}$ and $\phi_{\boldsymbol{\rho}}$, ($\mathbf{J} \circ \phi_{\boldsymbol{\rho}} : \mathcal{T} \rightarrow \mathbb{R}^q$). Given a continuous function $\mathcal{I}$ defined at each point of $\mathcal{F}_{\boldsymbol{\rho}}(\mathcal{T})$, the integral w.r.t. the volume is given by*

$$J(\boldsymbol{\rho}) = \int_{\mathcal{T}} \left(\mathcal{I} \circ \left(\mathbf{J} \circ \phi_{\boldsymbol{\rho}}\right)\right) Vol\left(D_{\boldsymbol{\theta}}\mathbf{J}(\boldsymbol{\theta}) D_{\mathbf{t}}\phi_{\boldsymbol{\rho}}(\mathbf{t})\right) \mathrm{d}\mathbf{t},$$

provided this integral exists and $Vol\left(X\right) = \left(det\left(X^T \cdot X\right)\right)^{\frac{1}{2}}$. The associated gradient w.r.t. the map parameters $\boldsymbol{\rho}$ is given component–wise by

$$\begin{aligned}\frac{\partial J\left(\boldsymbol{\rho}\right)}{\partial \boldsymbol{\rho}_i} = & \int_{\mathcal{T}} \frac{\partial}{\partial \boldsymbol{\rho}_i} \left(\mathcal{I} \circ \left(\mathbf{J} \circ \phi_{\boldsymbol{\rho}}\right)\right) Vol\left(\mathbf{T}\right) \mathrm{d}\mathbf{t} \\ & + \int_{\mathcal{T}} \left(\mathcal{I} \circ \left(\mathbf{J} \circ \phi_{\boldsymbol{\rho}}\right)\right) Vol\left(\mathbf{T}\right) \left(vec\ \left(\mathbf{T}^T\mathbf{T}\right)^{-T}\right)^T \cdot \\ & \cdot N_b \left(I_b \otimes \mathbf{T}^T\right) D_{\boldsymbol{\rho}_i}\mathbf{T} \, \mathrm{d}\mathbf{t},\end{aligned}$$

where $\mathbf{T} = D_{\boldsymbol{\theta}}\mathbf{J}(\boldsymbol{\theta}) D_{\mathbf{t}}\phi_{\boldsymbol{\rho}}(\mathbf{t})$, $\otimes$ is the Kronecker product, $N_b = \frac{1}{2}\left(I_{b^2} + K_{bb}\right)$ is a symmetric $(b^2 \times b^2)$ idempotent matrix with rank $\frac{1}{2}b(b+1)$ and K_{bb} is a permutation matrix (Magnus and Neudecker 1999).

As the reader may have noticed, we have left the term $D_{\boldsymbol{\rho}_i}\mathbf{T}$ unexpanded. This term represents the rate of expansion/compression of an infinitesimal volume block of the manifold under reparametrization. The derivation of this quantity is not trivial and requires a special focus. Exploiting algebraic tools, we can write

$$\begin{aligned}D_{\boldsymbol{\rho}_i}\mathbf{T} = & \left(D_{\mathbf{t}}\phi_{\boldsymbol{\rho}}(\mathbf{t})^{\mathrm{T}} \otimes I_q\right) D_{\boldsymbol{\theta}}\left(D_{\boldsymbol{\theta}}\mathbf{J}(\boldsymbol{\theta})\right) D_{\boldsymbol{\rho}_i}\, \phi_{\boldsymbol{\rho}}(\mathbf{t}) \\ & + \left(I_b \otimes D_{\boldsymbol{\theta}}\mathbf{J}(\boldsymbol{\theta})\right) D_{\boldsymbol{\rho}_i}\left(D_{\mathbf{t}}\phi_{\boldsymbol{\rho}}(\mathbf{t})\right),\end{aligned}$$

where $D_{\boldsymbol{\theta}}\left(D_{\boldsymbol{\theta}}\mathbf{J}(\boldsymbol{\theta})\right)$ is a transformation of the Hessian matrix of the performance w.r.t. policy parameters, that is, it contains the same elements, but in different order. In fact, the Hessian matrix is defined as the derivative of the transpose Jacobian, that is, $H_{\boldsymbol{\theta}}\mathbf{J}(\boldsymbol{\theta}) = D_{\boldsymbol{\theta}}(D_{\boldsymbol{\theta}}\mathbf{J}(\boldsymbol{\theta}))^{\mathrm{T}}$. The following equation relates the Hessian matrix to $D_{\boldsymbol{\theta}}\left(D_{\boldsymbol{\theta}}\mathbf{J}(\boldsymbol{\theta})\right)$:

$$H^{m,n}_{\boldsymbol{\theta}} J_i = \frac{\partial}{\partial \boldsymbol{\theta}_n}\left(\frac{\partial \mathbf{J}_i}{\partial \boldsymbol{\theta}_m}\right) = D^{p,n}_{\boldsymbol{\theta}}\left(D_{\boldsymbol{\theta}}\mathbf{J}(\boldsymbol{\theta})\right),$$

where $p = i + q(m-1)$ and q is the number of rows of the Jacobian matrix. Up to now, little research has been done on second order methods and in particular on Hessian formulation. A first analysis was performed in (Kakade 2001) where the authors provided a formulation based on the policy gradient theorem (Sutton et al. 1999). However, we provide a different derivation of the Hessian coming from the trajectory–based definition of the expected discounted reward for episodic MDPs (Furmston and Barber 2012).

Theorem 2. *For any MOMDP, the Hessian $H_{\boldsymbol{\theta}}\mathbf{J}(\boldsymbol{\theta})$ of the expected discounted reward $\mathbf{J}$ w.r.t. the policy parameters $\boldsymbol{\theta}$ is a $(qd \times d)$ matrix obtained by stacking the Hessian of each component*

$$H_{\boldsymbol{\theta}}\mathbf{J}(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}^T} vec\left(\frac{\partial \mathbf{J}_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T}\right)^T = \begin{bmatrix} H_{\boldsymbol{\theta}}\mathbf{J}_1(\boldsymbol{\theta}) \\ \vdots \\ H_{\boldsymbol{\theta}}\mathbf{J}_q(\boldsymbol{\theta}) \end{bmatrix},$$

where

$$\begin{aligned}H_{\boldsymbol{\theta}}\mathbf{J}_i(\boldsymbol{\theta}) = & \int_{\mathbb{T}} p\left(\tau|\boldsymbol{\theta}\right) \mathbf{r}_i(\tau) \cdot \\ & \cdot \left(\nabla_{\boldsymbol{\theta}} \log p\left(\tau|\boldsymbol{\theta}\right) \nabla_{\boldsymbol{\theta}} \log p\left(\tau|\boldsymbol{\theta}\right)^T + D_{\boldsymbol{\theta}}\left(\nabla_{\boldsymbol{\theta}} \log p\left(\tau|\boldsymbol{\theta}\right)\right)\right) \mathrm{d}\tau.\end{aligned}$$

Gradient Estimation from Sample Trajectories

In the RL setting, having no prior knowledge about the reward function and the state transition model, we need to estimate the gradient $\nabla_{\boldsymbol{\rho}} J(\boldsymbol{\rho})$ from trajectory samples. In this section we present standard results related to the estimation approaches used in RL literature and we provide a theoretical analysis of the Hessian estimate.

The formulation of the gradient $\nabla_{\boldsymbol{\rho}}\mathbf{J}(\boldsymbol{\rho})$ provided in Theorem 1 is composed by terms related to the parameterization of the manifold in the policy space and terms related to the MDP. Since the map $\phi_{\boldsymbol{\rho}}$ is chosen by the system designer, the associated terms (e.g., $D_{\mathbf{t}}\phi_{\boldsymbol{\rho}}(\mathbf{t})$) can be computed exactly. On the other hand, the terms related to the MDP ($\mathbf{J}(\boldsymbol{\theta})$, $D_{\boldsymbol{\theta}}\mathbf{J}(\boldsymbol{\theta})$ and $H_{\boldsymbol{\theta}}\mathbf{J}(\boldsymbol{\theta})$) need to be estimated. While the estimate of the expected discounted reward and the associated gradient is an old topic in RL literature and several results have been proposed (Kakade 2001; Pirotta, Restelli, and Bascetta 2013), the estimate of the Hessian has not been yet addressed. Recently, the simultaneous perturbation stochastic approximation technique was exploited to estimate the Hessian (Fonteneau and Prashanth 2014). Here we provide a Hessian estimate from trajectory

samples obtained through the current policy, without the need of generating policy perturbations.

Suppose to have access to a set of N trajectories of T steps, since $p(\tau|\boldsymbol{\theta})$ is unknown, the expectation is approximated by the empirical average:

$$\widehat{H}_{\boldsymbol{\theta}}\mathbf{J}_i(\boldsymbol{\theta}) = \frac{1}{N}\sum_{n=1}^{N}\left(\sum_{l=1}^{T}\gamma_i^{l-1}\mathbf{r}_{i,l}^n\right)\left(\sum_{k=1}^{T}\nabla_{\boldsymbol{\theta}}\log\pi_{\boldsymbol{\theta}}(a_k^n|s_k^n)\cdot\right.$$
$$\left.\cdot\left(\sum_{k=1}^{T}\nabla_{\boldsymbol{\theta}}\log\pi_{\boldsymbol{\theta}}(a_k^n|s_k^n)\right)^{\mathrm{T}} + \sum_{k=1}^{T}H_{\boldsymbol{\theta}}\log\pi_{\boldsymbol{\theta}}(a_k^n|s_k^n)\right), \quad (1)$$

where $\left\{s_k^n, a_k^n, \mathbf{r}_{:,k}^n\right\}_{k=1}^T$ denotes the n-th trajectory. This formulation resembles the definition of REINFORCE estimate given in (Williams 1992) for the gradient $\nabla_{\boldsymbol{\theta}}\mathbf{J}(\boldsymbol{\theta})$. Such kind of estimate, known as likelihood ratio methods, overcome the problem of controlling the perturbation of the parameters in finite–difference methods.

In order to simplify the theoretical analysis we make the following assumptions.

Assumption 3 (Uniform boundedness)**.** The reward function, the log–Jacobian, and the log–Hessian of the policy are uniformly bounded: $\forall i = 1,\dots,q,\ \forall m = 1,\dots,d,\ \forall n = 1,\dots,d,\ (s,a,s') \in \mathcal{S}\times\mathcal{A}\times\mathcal{S},\ \boldsymbol{\theta}\in\Theta$

$$\left|\mathbf{R}_i(s,a,s')\right| \le \overline{R}_i, \qquad \left|D_{\boldsymbol{\theta}}^m \log\pi_{\boldsymbol{\theta}}(a|s)\right| \le \overline{D},$$
$$\left|H_{\boldsymbol{\theta}}^{m,n}\log\pi_{\boldsymbol{\theta}}(a|s)\right| \le \overline{G}.$$

Lemma 4. *Given a parametrized policy $\pi(a|s,\boldsymbol{\theta})$, under Assumption 3, the i–th component of the log–Hessian of the expected return can be bounded by*

$$\left\|H_{\boldsymbol{\theta}}\mathbf{J}_i(\boldsymbol{\theta})\right\|_{\max} \le \frac{\overline{R}_i T\gamma^T}{1-\gamma}\left(T\overline{D}^2 + \overline{G}\right).$$

Note that the max norm of a matrix is defined as $\|A\|_{\max} = \max_{i,j}\{a_{ij}\}$. Previous results can be used to derive a bound on the sample complexity of the Hessian estimate.

Theorem 5. *Given a parametrized policy $\pi(a|s,\boldsymbol{\theta})$, under Assumption 3, using the following number of T–step trajectories*

$$N = \frac{1}{2\epsilon_i^2}\left(\frac{\overline{R}_i T\gamma^T}{(1-\gamma)}\left(T\overline{D}^2 + \overline{G}\right)\right)^2 \log\frac{2}{\delta},$$

the gradient estimate $\widehat{H}_{\boldsymbol{\theta}}\mathbf{J}_i(\boldsymbol{\theta})$ generated by Equation (1) *is such that with probability $1-\delta$:*

$$\left\|\widehat{H}_{\boldsymbol{\theta}}\mathbf{J}_i(\boldsymbol{\theta}) - H_{\boldsymbol{\theta}}\mathbf{J}_i(\boldsymbol{\theta})\right\|_{\max} \le \epsilon_i.$$

Finally, the estimate of the integral can be computed using standard Monte–Carlo techniques. Several statistical bounds have been proposed in literature, we refer to (Robert and Casella 2004) for a survey on Monte–Carlo methods.

Metrics for Multi–objective Optimization

In this section we review some indicator functions proposed in literature underlying advantages and drawbacks and we propose some alternatives.

Recently, MOO has focused on the use of performance indicators to turn a multi–objective optimization problem into a single-objective one by optimizing the indicator itself. The indicator function is used to assign to every point a single–objective measure, or, in other words, to give an approximate measure of the discrepancy between the candidate frontier and the Pareto one. Since, instead of optimizing the objective function directly, indicator–based algorithms aim at finding a solution set that maximizes the indicator metric, a natural question arises about the correctness of this change in the optimization procedure and on the properties the indicator functions enjoy.

For instance, hypervolume indicator and its weighted version are among the most widespread metrics in literature. These metrics have gained popularity because they are refinements of the Pareto dominance relation (Zitzler, Thiele, and Bader 2010). Recently, several works have been proposed in order to theoretically investigate the properties of hypervolume indicator (Friedrich, Horoba, and Neumann 2009). Nevertheless, it has been argued that the hypervolume indicator may introduce a bias in the search. From our perspective, the main issue of this metric is the high computational complexity[3] and, above all, the non differentiability. Several other metrics have been defined in the field of MOO, we refer to (Okabe, Jin, and Sendhoff 2003) for an extensive survey. However, MOO literature has not been able to provide a superior metric and among the candidates no one is suited for this scenario. Again the main problems are the non differentiability and the possibility of evaluating only discrete representations of the Pareto frontier.

In order to overcome these issues we have tried to mix different indicator concepts in order to obtain a metric with the desired properties. The insights that have guided our metric definition are related to the MOO desiderata. Recall that the goal of MOO is to compute an approximation of the frontier that includes solutions that are *accurate*, *evenly distributed*, and *covering* a range similar to the actual one (Zitzler et al. 2003). Note that the uniformity of the frontier is intrinsically guaranteed by the continuity of the approximation we have introduced. Having in mind these concepts we need to impose accuracy and extension of the frontier through the indicator function.

Given a reference point $\mathbf{p}$, a simple indicator can be obtained by computing the distance between every point of the frontier $\mathcal{F}$ and the reference point

$$\mathcal{I}_1(\mathbf{J},\mathbf{p}) = \|\mathbf{J}-\mathbf{p}\|_2^2.$$

As shown in the hypervolume indicator, the choice of the reference point may be critical. However, a natural choice is the utopia (ideal) point ($\mathbf{p}_u$), i.e., the point that optimizes all the objective functions. In this case the goal is the minimization of such indicator function. Since any dominated policy is farther from the utopia than at least one Pareto optimal solution, the accuracy can be easily guaranteed. On the other hand, it is also easy to show that this measure forces the solution to collapse into a single point. If the extension of the frontier is the primary concern, maximizing the distance from the antiutopia ($\mathbf{p}_{au}$) results in a metric that grows with the frontier dimension. However, since we are trying to maximize a possibly unbounded function that is not related

[3]The computation of the hypervolume indicator is a #P–hard problem (Friedrich, Horoba, and Neumann 2009).

to the Pareto optimality, this measure does not provide any guarantees about accuracy.

Concerning the accuracy of the frontier, from a theoretical perspective, it is possible to define a metric based on the Pareto optimality. A point $\overline{\boldsymbol{\theta}}$ is Pareto optimal when

$$\mathbf{l}(\overline{\boldsymbol{\theta}}, \boldsymbol{\alpha}) = \sum_{i=1}^{q} \alpha_i \nabla_{\boldsymbol{\theta}} \mathbf{J}_i(\overline{\boldsymbol{\theta}}) = \mathbf{0}, \quad \sum_{i=1}^{q} \alpha_i = 1, \quad \boldsymbol{\alpha} \in \mathbb{R}_+^q,$$

this means that it is not possible to identify an ascent direction that simultaneously improves all the objectives. As a consequence, any point on the Pareto frontier nullifies the norm of direction $\mathbf{l}$. Formally, a metric that reflects the Pareto–optimality can be defined as follows

$$\mathcal{I}_2(\mathbf{J}) = \min_{\boldsymbol{\alpha}} \|\mathbf{l}(\boldsymbol{\theta}, \boldsymbol{\alpha})\|_2^2, \qquad \sum_i \alpha_i = 1, \boldsymbol{\alpha} \in \mathbb{R}_+^q.$$

As for the utopia–based metric, the extent of the frontier is not taken into account. To summarize, all the mentioned indicators provide only one of the desiderata, but we deserve more since achieving only one property may result in a frontier arbitrary far from the actual one. In order to consider all the desiderata we have decided to mix previous concepts into a single indicator

$$\mathcal{I}_3(\mathbf{J}) = \mathcal{I}_1(\mathbf{J}, \mathbf{p}_{au}) \cdot w(\mathbf{J}),$$

where $w(\mathbf{J})$ is a penalization term, i.e., it is a monotonic function that decreases as $\mathcal{I}_2(\mathbf{J})$ increases, e.g., $w(\mathbf{J}) = 1 - \lambda \mathcal{L}_2(\mathbf{J})$. Metric $\mathcal{I}_3$ takes advantage of the expansive behavior of the antiutopia–based indicator and the accuracy of the optimality–based indicator $\mathcal{I}_2$. In this way all the desiderata can be considered by a single scalar measure, that is also C^l $(l \geq 1)$ differentiable.

Experiments

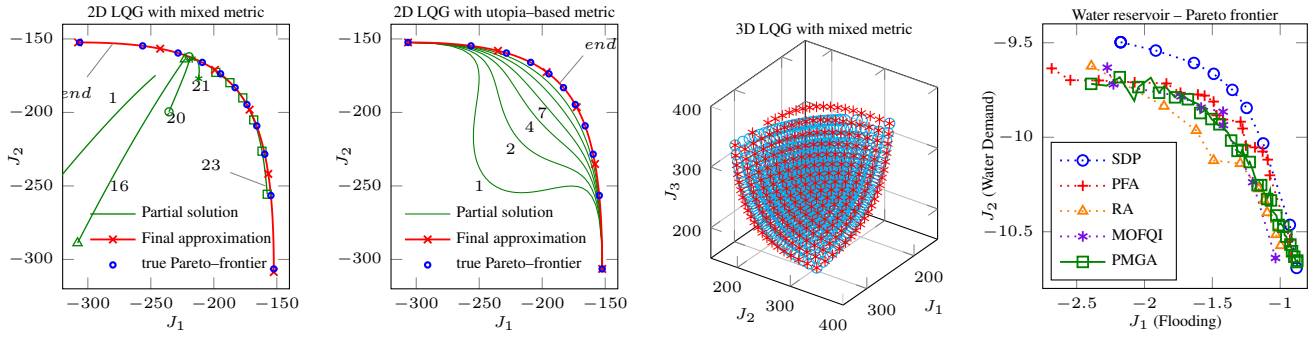
In this section, results related to the numerical simulations of the PMGA algorithm, in continuous domains, are presented. Performances are compared against value–based and gradient algorithms: Stochastic Dynamic Programming (SDP), Multi-Objective Fitted Q–Iteration (MOFQI) (Castelletti, Pianosi, and Restelli 2013), Pareto Following Algorithm (PFA) and Radial Algorithm (RA) (Parisi et al. 2014). The comparison has been performed by selecting MORL algorithms that can operate in continuous domains. In all the experiments the learning rate α was set by hand-tuning.

We start considering a multi–objective version of the standard discrete-time Linear-Quadratic Gaussian regulator (LQG) with multidimensional and continuous state and action spaces (Peters and Schaal 2008). For a complete description of the LQG problem and for the settings, we refer to (Parisi et al. 2014). This scenario is particular instructive since all the terms can be computed exactly, so that we can focus on the effects of using different parametrizations and metrics, while we demand the analysis of the gradient estimation to the water reservoir domain. Initially we present the results for a 2–dimensional LQG problem. The LQG is a problematic domain since it is defined only for control actions in the range $[-1, 0]$, controls outside this range leads to divergence of the system. Our primary concern is related to the boundedness of the control actions, leading to the following parametrization of the manifold in the policy space: $\phi^1_{\boldsymbol{\rho}}(\mathbf{t}) = {}^{-1}/_{1+\exp(\rho_1+\rho_2 t)}$ and $\phi^2_{\boldsymbol{\rho}}(\mathbf{t}) = {}^{-1}/_{1+\exp(\rho_3+\rho_4 t)}$ with $t \in [0, 1]$. While metrics $\mathcal{I}_1$ and $\mathcal{I}_2$ suffer from the problems described in the previous section that prevent PMGA to obtain a good approximation of the Pareto frontier, mixed metric ($\mathcal{I}_3$) achieves both accuracy and coverage. An example of the learning process obtained setting λ to 2.5 (a sensitive analysis w.r.t. λ is available in (Pirotta, Parisi, and Restelli 2014)) and starting from $\boldsymbol{\rho}^{(0)} = [1, 2, 0, 3]^\intercal$ is shown in Figure 2(a). First the accuracy is increased by pushing the parametrization onto the Pareto frontier, then the partial solution is expanded toward the extrema thus improving coverage.

An alternative approach consists in the computation of the optimal parametrizations of the single objectives, for instance through policy gradient techniques, exploiting such information for constraining the policy manifold to pass through these points. Recall that, in general, this information is required to compute the utopia and antiutopia points. Following such approach, two improvements can be easily obtained. First, the number of free parameters decreases and, as a consequence, the learning process simplifies. Second, the approximate frontier is forced to have a sufficiently large area to cover all the extrema. In this way, the coverage problem shown by indicators $\mathcal{I}_1$ and $\mathcal{I}_2$ can be alleviated or, in some cases, completely solved. For instance, forcing the parametrization to cover the extrema, has permitted to achieve both accuracy and coverage using metric $\mathcal{I}_1$ (utopia) and $\mathcal{I}_2$ in the 2–dimensional LQG problem. Figure 2(b) shows the learning process obtained through metric $\mathcal{I}_2$ under these settings. Clearly, no advantages have been found using the antiutopia–based metric. Although, this approach is effective for almost all 2–objective problems, it does not generalize to higher dimensions as shown in (Pirotta, Parisi, and Restelli 2014) for a 3–dimensional LQG.

Consider the 3–dimensional LQG domain described in (Pirotta, Parisi, and Restelli 2014). Despite the parametrization was forced through the single objective optima, the solution obtained with the utopia–based metric tends to concentrate on the center of the frontier, i.e., toward the points that minimize the distance from the utopia. It is important to underline that all the obtained solutions belong to the Pareto frontier, i.e., no dominant solutions are found. The same happens with metric $\mathcal{I}_2$. Mixing the antiutopia with the Pareto optimality, i.e., using metric $\mathcal{I}_3$, provides a way to obtain both accuracy and coverage through the tuning of the λ parameter. Figure 2(c) compares the Pareto frontier with the approximation obtained using $\mathcal{I}_3$ with $\lambda = 135$.

Concerning the approximate framework, we consider the water reservoir problem, a continuous MOMDP that, differently from the LQG, does not admit a closed–form solution. In order to compare the PMGA frontier with the ones obtained by other algorithms, we consider the domain, settings and policy parametrization as described in (Parisi et al. 2014). A second–order polynomial in $\mathbf{t} \in [0, 1]$ with 5 parameters has been used to parametrize the policy manifold. The number of parameters is 5 since we have constrained the policy manifold to pass through the optimal points. The reader may refer to (Pirotta, Parisi, and Restelli 2014) for details. In order to show the capability of PMGA we have de-



(a) (b) (c) (d)

Figure 2: Experimental results. Figures (a) and (b) show some candidate frontiers obtained by PMGA during the learning process in the 2D LQG problem without and with constraints, respectively. For sake of visualization we have decided to use a discrete representation of the true Pareto frontier since it overlaps the result of PMGA . Only few iterations have been reported, each one with the associated iteration number, where end denotes the frontier obtained when the terminal condition is reached. Figure (c) compares the Pareto frontier with its approximation obtained with metric $\mathcal{I}_3$ in the 3D LQG. Figure (d) is related to the water reservoir domain and represent frontiers obtained with different algorithms.

Table 1: 2D LQG: empirical sample complexity. Number of iterations and the evaluations of the model (averaged over 10 runs) required to reach a loss from the weighted sum approximation less than $5 \cdot 10^{-4}$ within 1000 iterations. The symbol $\perp$ is used to denote the hit of such threshold.

		PMGA Parameters (#Episodes = 30, #steps = 30)				
		1	10	30	40	50
#$\mathbf{t}$	#Iterations	390.6 ± 50.0	221.3 ± 32.9	157.0 ± 20.4	149.2 ± 21.9	111.5 ± 9.7
	#Samples (10^5)	$\mathbf{3.5 \pm 0.5}$	19.9 ± 3.0	42.4 ± 5.5	53.7 ± 7.9	50.2 ± 4.4
		PMGA Parameters (#$\mathbf{t}$ = 30, #steps = 30)				
		1	10	30	40	50
#Episodes	#Iterations	$\perp$	$\perp$	127.6 ± 11.0	71.0 ± 4.5	68.4 ± 5.0
	#Samples (10^5)	$\perp$	$\perp$	34.5 ± 3.0	$\mathbf{25.6 \pm 1.6}$	30.8 ± 2.2

cided to test the simplest metric, that is, the utopia–based indicator. The integral estimate was performed using a Monte–Carlo algorithm fed with only 100 random points. For each value of $\mathbf{t}$, 100 trajectory by 100 steps were used to estimate the gradient and Hessian of the policy performance. We start the learning from an arbitrary parametrization with all the parameters $\boldsymbol{\rho}_i$ set to -20. Figure 2(d) reports the final frontier obtained with different algorithms. The approximation obtained by PMGA is comparable to the other results, but PMGA produces a continuous frontier approximation.

To conclude, another advantage of PMGA is the reduced number of samples trajectories to approximate the Pareto frontier. Let us consider the 2D LQR domain and let vary the number of policies used to estimate the integral and the number of episode for each policy evaluation. From Table 1 results that the most relevant parameter is the number of episodes used to estimate MDP terms: $\mathbf{J}(\boldsymbol{\theta})$, $D_{\boldsymbol{\theta}}\mathbf{J}(\boldsymbol{\theta})$ and $H\mathbf{J}(\boldsymbol{\theta})$. This parameter controls the variance in the estimate, i.e., the accuracy of gradient estimate $\nabla_{\boldsymbol{\theta}}\mathbf{J}(\boldsymbol{\rho})$. By increasing the number of episodes, the estimation process is less prone to generate misleading directions, as happens, for instance, in the 1–episode case where parameters move into wrong direction. On the contrary, the number of points used to estimate the integral (denoted in table by #$\mathbf{t}$) seems to have no significant impact on the final performance of the algorithm, but influences the number of model evaluations needed to reach the prescribed accuracy.

Conclusions

In this paper we have proposed PMGA, a novel gradient–based approach to learn a continuous approximation of the Pareto frontier in MOMDPs. The idea is to define a parametric function $\phi_{\boldsymbol{\rho}}$ that describes a manifold in the policy–parameter space, that maps to a manifold in the objective space. Given a metric that measures the quality of the manifold in the objective space (i.e., the candidate frontier), we have shown how to compute (and estimate from trajectory samples) its gradient w.r.t. the parameters of $\phi_{\boldsymbol{\rho}}$. Updating the parameters along the gradient direction generates a new policy manifold associated to an improved (w.r.t. the chosen metric), continuous frontier in the objective space. Although we have provided a derivation that is independent from the specific metric used to measure the quality of the candidate solutions, the choice of such metric strongly influences the final result. We have presented different alternatives, discussed pros and cons of each one, and shown their properties through an empirical analysis.

Future research will further address the study of metrics that can produce good results in general settings. Another interesting research direction consists in using importance sampling techniques for reducing the sample complexity in the gradient estimate. Since the frontier is composed of a continuum of policies, it is likely that a trajectory generated by a specific policy can be partially used also for the estimation of quantities related to similar policies.

References

Castelletti, A.; Pianosi, F.; and Restelli, M. 2013. A multiobjective reinforcement learning approach to water resources systems operation: Pareto frontier approximation in a single run. *Water Resources Research* 49(6):3476–3486.

Fonteneau, R., and Prashanth, L. A. 2014. Simultaneous Perturbation Algorithms for Batch Off-Policy Search. *ArXiv:1403.4514*.

Friedrich, T.; Horoba, C.; and Neumann, F. 2009. Multiplicative approximations and the hypervolume indicator. In *GECCO '09*, 571–578. New York, NY, USA: ACM.

Furmston, T., and Barber, D. 2012. A unifying perspective of parametric policy search methods for markov decision processes. In *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. December 3-6, 2012, Lake Tahoe, Nevada, United States.*, 2726–2734.

Harada, K.; Sakuma, J.; Kobayashi, S.; and Ono, I. 2007. Uniform sampling of local pareto-optimal solution curves by pareto path following and its applications in multi-objective ga. In *Proceedings of GECCO '07*, 813–820. New York, NY, USA: ACM.

Harada, K.; Sakuma, J.; and Kobayashi, S. 2006. Local search for multiobjective function optimization: pareto descent method. In *GECCO*, 659–666.

Kakade, S. 2001. Optimizing average reward using discounted rewards. In *COLT/EuroCOLT*, 605–615. Springer.

Lizotte, D. J.; Bowling, M.; and Murphy, S. A. 2012. Linear fitted-q iteration with multiple reward functions. *Journal of Machine Learning Research* 13:3253–3295.

Magnus, J., and Neudecker, H. 1999. *Matrix Differential Calculus with Applications in Statistics and Econometrics*. Wiley Ser. Probab. Statist.: Texts and References Section. Wiley.

Munkres, J. 1997. *Analysis On Manifolds*. Adv. Books Classics Series. Westview Press.

Okabe, T.; Jin, Y.; and Sendhoff, B. 2003. A critical survey of performance indices for multi-objective optimisation. In *CEC '03.*, volume 2, 878–885 Vol.2.

Parisi, S.; Pirotta, M.; Smacchia, N.; Bascetta, L.; and Restelli, M. 2014. Policy gradient approaches for multi-objective sequential decision making. In *IJCNN 2014, Beijing, China, July 6-11, 2014*, 1–7. IEEE.

Peters, J., and Schaal, S. 2008. Reinforcement learning of motor skills with policy gradients. *Neural Networks* 21(4):682–697.

Pirotta, M.; Parisi, S.; and Restelli, M. 2014. Multi-objective reinforcement learning with continuous pareto frontier approximation supplementary material. *CoRR* abs/1406.3497.

Pirotta, M.; Restelli, M.; and Bascetta, L. 2013. Adaptive step-size for policy gradient methods. In *NIPS 26*. Curran Associates, Inc. 1394–1402.

Robert, C. P., and Casella, G. 2004. *Monte Carlo statistical methods*, volume 319. New York: Springer.

Roijers, D. M.; Vamplew, P.; Whiteson, S.; and Dazeley, R. 2013. A survey of multi-objective sequential decision-making. *JAIR* 48:67–113.

Shelton, C. R. 2001. *Importance Sampling for Reinforcement Learning with Multiple Objectives*. Ph.D. Dissertation, Massachusetts Institute of Technology.

Sutton, R. S., and Barto, A. G. 1998. *Reinforcement Learning: An Introduction*. The MIT Press.

Sutton, R. S.; McAllester, D. A.; Singh, S. P.; and Mansour, Y. 1999. Policy gradient methods for reinforcement learning with function approximation. In *NIPS*, 1057–1063. The MIT Press.

Vamplew, P.; Dazeley, R.; Berry, A.; Issabekov, R.; and Dekker, E. 2011. Empirical evaluation methods for multiobjective reinforcement learning algorithms. *Machine Learning* 84(1-2):51–80.

Williams, R. J. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning* 8(3-4):229–256.

Zitzler, E.; Thiele, L.; Laumanns, M.; Fonseca, C.; and da Fonseca, V. 2003. Performance assessment of multiobjective optimizers: an analysis and review. *Evolutionary Computation, IEEE Transactions on* 7(2):117–132.

Zitzler, E.; Thiele, L.; and Bader, J. 2010. On set-based multiobjective optimization. *Trans. Evol. Comp* 14(1):58–79.