Multi-objective Reinforcement Learning with Continuous Pareto Frontier Approximation

Matteo Pirotta, Simone Parisi, and Marcello Restelli

Department of Electronics, Information, and Bioengineering, Politecnico di Milano, Piazza Leonardo Da Vinci, 32, I-20133, Milano, Italy

June 16, 2014

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 use 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.

1 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) [1] 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) [2, 3] 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 [4]: 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 [5]. The superiority of multiple-policy methods resides in the ability to represent the Pareto-optimal manifold, allowing a posteriori selection of the solution, a graphical representation of the frontier that can give better insight into the relationships among the objectives, and encapsulate all the trade-offs among the multiple objectives. Building the exact frontier is generally impractical in real-world problems, the goal is thus 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 [6]. Among multiple-policy algorithms it is possible to identify two classes: value-based [7, 8] and gradient approaches [9, 10]. While value-based approaches suffer from the course
of dimensionality problem and, in general, are not able to identify concave frontiers, gradient–based techniques lack of guarantees of uniform covering of the Pareto frontier.

In this paper, we propose a novel gradient–based MORL approach that generates a continuous approximation of the local Pareto–optimal solution manifold in the policy space. 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– (Section 3), how to estimate such gradient from sample trajectories (Section 4), a discussion of frontier quality measures that can be effectively integrated in the proposed gradient approach (Section 5), and an empirical evaluation of its performance in a multi–objective extension of the discrete-time Linear-Quadratic Gaussian regulator and in a water reservoir management domain.

2 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 $\langle S, A, P, R, \gamma, D \rangle$, where $S \subseteq \mathbb{R}^n$ is the continuous state space, $A \subseteq \mathbb{R}^m$ is the continuous action space, $P$ is a Markovian transition model where $P(s'|s,a)$ defines the transition density between state $s$ and $s'$ under action $a$, $R = [R_1, \ldots, R_q]^T$ and $\gamma = [\gamma_1, \ldots, \gamma_q]^T$ are $q$-dimensional column vectors of reward functions $R_i : S \times A \times S \to \mathbb{R}$ 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 $\pi$ is associated to $q$ expected returns $J^\pi = [J_1^\pi, \ldots, J_q^\pi]$, where

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

being $r_i(t+1) = R_i(s_t, a_t, s_{t+1})$ 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 parameterized space of policies $\Pi_\theta = \{ \pi_\theta : \theta \in \Theta \subseteq \mathbb{R}^d \}$ (where $\pi_\theta$ is a compact notation for $\pi_\theta(s, \theta)$) is considered. Given a policy parametrization $\theta$, we assume the policy performance $J : \Theta \to \mathbb{R}$ to be at least $C^2$. $J$ is defined as the expected reward over the space of all possible trajectories $\mathbb{T}$:

$$J(\theta) = \int_{\mathbb{T}} p(\tau|\theta) r(\tau) d\tau,$$

where $\tau$ is a trajectory drawn from density distribution $p(\tau|\theta)$ with reward vector $r(\tau)$ that represents the accumulated expected discounted reward over trajectory $\tau$: $r_i(\tau) = \sum_{t=0}^{T} \gamma_i^t r_i(t+1)$.

In MOMDPs for each policy parameter $\theta$, $q$ gradient directions are defined $[11]$

$$\nabla_\theta J_i(\theta) = \int_{\mathbb{T}} \nabla_\theta \log p(\tau|\theta) r_i(\tau) d\tau = \mathbb{E}_{\tau \in \mathbb{T}} \left[ \nabla_\theta \log p(\tau|\theta) r_i(\tau) \right] = \mathbb{E}_{\tau \in \mathbb{T}} \left[ r_i(\tau) \sum_{i=1}^{T} \nabla_\theta \log \pi(a_i|s_i, \theta) \right],$$

where each direction $\nabla_\theta J_i$ is associated to a particular discount factor–reward function pair $< \gamma_i, R_i >$.

As shown in previous equation, the differentiability of the performance measure is connected to the differentiability of the policy class by: $\nabla_\theta \log p(\tau|\theta) = \sum_{i=1}^{T} \nabla_\theta \log \pi(a_i|s_i, \theta)$.

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 $\pi$ dominates policy $\pi'$, which is denoted by $\pi \succeq \pi'$, if:

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

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

A remark on notation. In the following we will use the symbol $D_{X^T}$ to denote the derivative $^T$ 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: $\nabla_x f = (D_x f)^T$. Finally, the symbol $I_x$ will be used to denote an $x \times x$ identity matrix.

1As done in [12], we suppose that local Pareto-optimal solutions that are not Pareto-optimal do not exist.

2The derivative operator is well defined for matrices, vectors and scalar functions. Refer to [13] for details.
In this section we first provide a general definition of the optimization problem that we want to solve and then we explain how we can solve it in the MOMDP case using a gradient–based approach.

3.1 Parametric Pareto Front in MOO

It has been shown [5] that local Pareto–optimal solutions locally form a \((q-1)\)-dimensional manifold, assuming \(d > q\). It follows that in two-objective problems, the Pareto–optimal solutions can be described by curves both in decision and objective spaces. The idea behind this work is to parametrize the local Pareto-optimal solution curve in the objective space, in order to produce a continuous representation of the Pareto frontier.

Let \(T\) be open in \(\mathbb{R}^b\) with \(b \leq q\). The high–dimensional analogous of a parameterized curve is a smooth map \(\psi: T \rightarrow \mathbb{R}^q\) of class \(C^l\) (\(l \geq 1\)), where \(t \in T\) and \(\rho \in P \subseteq \mathbb{R}^k\) are the free variable and the parameters, respectively. The set \(F = \psi(\rho)\), together with the map \(\psi\), constitute a parametrized manifold of dimension \(b\), denoted by \(F_\rho(T)\) [14]. This manifold represents our approximation of the true Pareto frontier. The goal is to find the best approximation, i.e., to the parameters \(\rho^*\) that minimize the distance from the real frontier 

\[
\rho^* = \arg \min_{\rho \in P} \mathcal{I}(F_\rho(T)),
\]

where \(\mathcal{I}: \mathcal{F}(T) \rightarrow \mathbb{R}\) is some loss function that measures the discrepancy between the Pareto-optimal frontier and \(F_\rho(T)\). However, since the Pareto frontier is not known, a different indicator function is needed. The definition of such metric is an open problem in literature. Recently [4], 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 in Section 5.

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

\[
J(\rho) = \int_{\mathcal{F}(T)} \mathcal{I}_V dV,
\]

where \(dV\) is a symbol used to denote the integral w.r.t. the volume of the manifold [14] and \(\mathcal{I}: \mathcal{F}(T) \rightarrow \mathbb{R}\) is a continuous indicator function that for each point of \(\mathcal{F}(T)\) measures its Pareto–optimality. A standard way to maximize previous equation is to perform gradient ascent, updating the parameters according to the gradient direction:

\[
\rho_{t+1} = \rho_t + \alpha_t \nabla \mathcal{I}_V(\rho).
\]

3.2 Parametric Pareto Front in MOMDP

Given a continuous indicator function \(\mathcal{I}\), the integral of \(\mathcal{I}\) over \(\mathcal{F}(T)\), with respect to volume, is defined by the equation [14]

\[
\int_{\mathcal{F}(T)} \mathcal{I}_V dV = \int_T (\mathcal{I} \circ \psi_\rho) Vol(D_t \psi_\rho(t)) dt,
\]

provided this integral exists and \(Vol(X) = [\det (X^T \cdot X)]^{\frac{1}{2}}\).
Now, the key point is the definition of the map $\psi$. Notice that the direct map between the parameter space $T$ and the objective space is unknown, but can be easily defined through a reparametrization that involves the policy space $\Theta$, as shown in Figure. In previous section we have mention that there is a tight relationship between the (local) manifold in the objective space and the (local) manifold in the variable space. This mapping is well known and it is defined by the performance function $J(\theta)$, that defines the goodness a policy $\pi^*$ w.r.t. the objectives. This means that, given a set of policy parametrizations, we can define the associated points in the objective space. As a consequence, the optimization problem can be reformulated as the search for the best approximation of the Pareto manifold in the policy parameter space, i.e., to the search of the manifold in the policy parameter space that describe the optimal Pareto frontier.

Formally, let $\phi_p : T \rightarrow \Theta$ be a smooth map of class $C^l$ ($l \geq 1$) defined on the same domain of $\psi_p$. We think of the map $\phi_p$ as a parameterization of the subset $\phi_p(T)$ of $\Theta$: each choice of a point $t \in T$ gives rise to a point $\phi_p(t)$ in $\phi_p(T)$. This means that only a subset $\Theta_p(T)$ of the space $\Theta$ can be spanned by map $\phi_p$, i.e., $\Theta_p(T)$ is a $b$-dimensional parametrized manifold in the policy space

$$\Theta_p(T) = \{ \theta : \theta = \phi_p(t), \forall t \in T \},$$

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

$$F(\Theta_p(T)) = \{ J(\theta) : \theta \in \Theta_p(T) \}.$$

**Lemma 3.1.** Let $T$ be an open set in $\mathbb{R}^b$, let $F_p(T)$ be a manifold parametrized by a smooth map $\psi_p : T \rightarrow \mathbb{R}^s$. Provided that the map $\psi_p$ can be expressed as composition of maps $J$ and $\phi_p$, i.e., $\psi_p = J \circ \phi_p$, and that $J$ is a continuous function defined at each point of $F_p(T)$, the integral w.r.t. the volume is given by

$$J(\rho) = \int_{F_p(T)} \mathcal{I} dV = \int_T (I \circ (J \circ \phi_p)) \text{Vol} (D_b J(\theta) D_{\theta} \phi_p(t)) dt.$$

The associated gradient w.r.t. the map parameters $\rho$ is given component-wise by

$$\frac{\partial J(\rho)}{\partial \rho_i} = \int_T \frac{\partial}{\partial \rho_i} (I \circ (J \circ \phi_p)) \text{Vol} (T) dt$$

$$+ \int_T (I \circ (J \circ \phi_p)) \text{Vol} (T) \left( \text{vec} \right) \left( T^T T \right)^{-1} N_b (I_b \otimes T^T) D_{\theta} \phi_p(t) dt$$

where $T = D_b J(\theta) D_{\theta} \phi_p(t)$, $\otimes$ is the Kronecker product, $N_b = \frac{1}{2} (I_{b^2} + K_{b^2})$ is a symmetric ($b^2 \times b^2$) idempotent matrix with rank $\frac{1}{2} b(b+1)$ and $K_{b^2}$ is a permutation matrix.

As the reader may have noticed, we have left the term $D_{\theta} T$ 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

$$D_{\theta} T = (D_{\theta} \phi_p(t)^T \otimes I_b) D_b (D J(\theta)) D_{\theta} \phi_p(t) + (I_b \otimes D_b J(\theta)) D_{\theta} (D_{\theta} \phi_p(t))$$

where $D_b (D J(\theta))$ 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, $H J(\theta) = D^T (D J(\theta))^T$. However, the two matrices contain the same elements, but in different order. The following equation relates the Hessian matrix to $D_b (D J(\theta))$:

$$H^m_{\theta} J_i = D^2_{\theta,\theta} J_i(\theta) = \frac{\partial}{\partial \theta_m} \left( \frac{\partial J_i(\theta)}{\partial \theta_n} \right) = D^p_{\theta} (D J(\theta))$$

where $p = i + q(m-1)$, where $q$ 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 where the authors provided a formulation based on the policy gradient theorem. However, we provide a different derivation of the Hessian coming from the trajectory–based definition of the expected discounted reward for episodic MDPs.

**Lemma 3.2.** For any MOMDP, the Hessian $H J(\theta)$ of the expected discounted reward $J$ w.r.t. the policy parameters $\theta$ is a $qd \times d$ matrix obtained by stacking the Hessian of each component

$$H J(\theta) = \frac{\partial}{\partial \theta^T} \text{vec} \left( \frac{\partial J_i(\theta)}{\partial \theta^T} \right)^T = \begin{bmatrix} H J_1(\theta) \\ \vdots \\ H J_q(\theta) \end{bmatrix},$$
\[ H \mathcal{J}_i(\theta) = \int_T p(\tau|\phi) r_i(\tau) \left[ \nabla_\theta \log p(\tau|\theta) \nabla_\theta \log p(\tau|\theta)^T + D_\theta \left( \nabla_\theta \log p(\tau|\theta) \right) \right] d\tau. \] (1)

4 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 \( \nabla \phi, \mathcal{J}(\rho) \) 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 \( \nabla \phi, \mathcal{J}(\rho) \) provided in Lemma 3.1 is composed by terms related to the parameterization of the manifold in the policy space and terms related to the MDP. Since the map \( \phi_\rho \) is free to be designed, the associated terms (e.g., \( D_\phi(t) \)) need to be estimated. On the other hand, terms related to the MDP (\( \mathcal{J}^\theta \), \( D_\theta \mathcal{J}(\theta) \) and \( H \mathcal{J}(\theta) \)) 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 \([15, 17]\), the estimate of the Hessian is not addressed in literature. Recently, the simultaneous perturbation stochastic approximation technique was exploited to estimate the Hessian \([13]\). Here we provide a Hessian estimate from trajectory samples obtained through the current policy, removing the necessity of generating policy perturbations.

Suppose to have access to a set of \( N \) trajectory, since \( p(\tau|\theta) \) is unknown, the expectation is approximated by the empirical average:

\[ \tilde{H} \mathcal{J}_i(\theta) = \frac{1}{N} \sum_{n=1}^N \left( \sum_{k=1}^T \nabla \phi \log \pi_n^{\theta,s} \cdot \pi_n^{\theta,s} \right)^T \left( \sum_{k=1}^T \nabla \phi \log \pi_n^{\theta,s} \cdot \pi_n^{\theta,s} \right) + \sum_{k=1}^T H \log \pi_n^{\theta,s} \right] \] (2)

where \( \{s_n^a, a_n^s, r_n^a\}^T \) denotes the \( n \)-th trajectory. This formulation resemble the definition of REINFORCE estimate given in \([19]\) for the gradient \( \nabla_\phi \mathcal{J}(\theta) \). Such types of estimate, known as likelihood ratio methods, overcome the problem of finite-difference methods, that is, the problem of control the perturbation of the parameters.

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

Assumption 4.1 (Uniform boundedness). The reward function, the log-Jacobian and the log-Hessian of the policy are uniformly bounded: \( \forall i = 1, \ldots, q, \forall m = 1, \ldots, d, \forall n = 1, \ldots, d, \{s, a, s'\} \in S \times A \times S \in \Theta \)

\[ |R_i(s, a, s')| \leq R_i, \quad |D^{(m)}_\phi \log \pi(a|s, \theta)| \leq D_i, \quad |H^{(m)}_\phi \log \pi(a|s, \theta)| \leq \tilde{C}. \]

Lemma 4.2. Given a parametrized policy \( \pi(a, s, \theta) \), under the assumption Assumption 4.1, the \( i \)-th component of the log-Hessian of the expected return can be bounded by

\[ \|H \mathcal{J}_i(\theta)\|_{\text{max}} \leq \frac{R_i T \gamma^T}{1 - \gamma} \left( T D_i^2 + \tilde{C} \right). \]

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

Theorem 4.3. Given a parametrized policy \( \pi(a, s, \theta) \), under Assumption 4.1 using the following number of \( H \)-step trajectories

\[ N = \frac{1}{2\kappa_\theta^2} \left( \frac{R_i T \gamma^T}{(1 - \gamma)} \left( T D_i^2 + \tilde{C} \right) \right)^2 \log \frac{2}{\delta} \]

the gradient estimate \( \tilde{H} \mathcal{J}_i(\theta) \) generated by Equation (2) is such that with probability \( 1 - \delta \):

\[ \|\tilde{H} \mathcal{J}_i(\theta) - H \mathcal{J}_i(\theta)\|_{\text{max}} \leq \epsilon_i. \]

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 \([20]\) for a survey on Monte-Carlo methods.
5 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 functions 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 [21]. Several works have been proposed in order to theoretically investigate the properties of hypervolume indicator [22]. 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 and, above all, the non differentiability. Several other metrics have been defined in the field of MOO, we refer to [23] 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 capability of evaluate only discrete representations of the Pareto frontier.

In order to overcome these issues we have try to mix different indicator concepts in order to obtain a metric with the desired properties. The insights that have guided or 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 [3]. 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 \( p \), a simple indicator can be obtained by computing the distance between every point of the frontier \( F \) and the reference point

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

As shown in the hypervolume indicator, the choice of the reference point may be critical. However, a natural choice is the utopia (ideal) point \( (p_u) \), 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 antutopia \( (p_{au}) \) 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 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 \( \overline{\theta} \) is Pareto optimal when

\[
I(\overline{\theta}, \alpha) = \sum_{i=1}^{q} \alpha_i \nabla^\alpha J_i(\overline{\theta}) = 0, \quad \sum_{i=1}^{q} \alpha_i = 1, \quad \alpha \in \mathbb{R}_+^q,
\]

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 frontier minimizes the norm of direction \( I \). Formally, a metric that respect the Pareto–optimality can be defined as follows

\[
\mathcal{I}_2(J) = \min_{\alpha} \|I(\theta, \alpha)\|_2^2, \quad \sum_{i} \alpha_i = 1, \alpha \in \mathbb{R}_+^l.
\]

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 both the desiderata we have decided to mix previous concepts into a single indicator

\[
\mathcal{I}_3(J) = \mathcal{I}_1(J, p_{au}) \cdot w(J)
\]

\footnote{The computation of the hypervolume indicator is a \#P–hard problem. [22]}
where $w(J)$ is a penalization term, i.e., it is a monotonic function that decreases as $I_2(J)$ increases, e.g., $w(J) = 1 - \lambda I_2(J)$. Metric $I_3$ takes advantage of the expansive behavior of the antutopia–based indicator and the accuracy of the optimality–based indicator $I_2$. In this way all the desiderata can be met by a single scalar measure, that is also $C^t (t \geq 1)$ differentiable.

6 Experiments

In this section, results related to the numerical simulations of the proposed algorithms, in continuous and discrete domains, are presented. In particular, the performance are compared against some existing algorithms [8][10]. In all the experiments the learning rate $\alpha$ 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 [11]. For a complete description of the LQG problem and for the settings, we refer to [10]. This scenario is particular instructive since all the terms can be computed exactly, we focus on showing the properties of the parametrization and we demand the analysis of the estimate phase 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 was related to the boundedness of the control actions, leading to the following parametrization of the manifold in the policy space: $\phi^*_I(t) = \frac{1}{1 + \exp(\rho_1 + \rho_2 t)}$ and $\phi^*_J(t) = \frac{1}{1 + \exp(\rho_3 + \rho_4 t)}$ with $t \in [0, 1]$. While metrics $I_1$ and $I_2$ suffer from the problems described in Section 5 that prevent the algorithm to obtain a correct approximation of the Pareto frontier, mixed metric $(I_3)$ is able to achieve both accuracy and covering. An example of the learning process obtained setting $\lambda$ to 2.5 and starting from $\rho^{(0)} = [1, 2, 0, 3]^T$ is shown in Figure 2(a). Notice that first accuracy is obtained by pushing the parametrization onto the Pareto frontier, then the frontier is expanded toward the extrema in order to attain covering.

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 antutopia 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 problem of covering shown by indicators $I_1$ and $I_2$ can be alleviated or, in some cases, completely eliminated. For instance, forcing the parametrization to cover the extrema, has permitted to achieve both accuracy and covering using metric $I_1$ (utopia) and $I_2$ in the 2–dimensional LQG problem. Figure 2(b) shows the learning process obtained through metric $I_2$ under these settings. Clearly, no advantages have been found using antutopia–based metric. Although, this approach is effective for almost all 2–objective problems, it does not generalize to higher dimensions as shown in the supplementary material for a 3–dimensional LQG.

Consider the 3–dimensional LQG domain described in the supplementary material. Despite the parametrization was forced through the single objective optimums, the solution obtained with the utopia–based metric tends to concentrate on the centre 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 indicator function $I_2$. Mixing the antutopia with the Pareto optimality, i.e., using indicator function $I_3$, provides a way to obtain both accuracy and covering. Figure 2(c) compares the true Pareto frontier with the parametric approximation obtained using $I_3$ with $\lambda = 1.35$.

Concerning the approximate framework, we consider the water reservoir problem, a continuous MOMDP that, differently from the LQG, do not admit a closed–form solution. In order to compare our frontier with the one obtained by other algorithms, we consider the domain, settings and policy parametrization as described in [10]. A simple second–order polynomial in $t \in [0,1]$ with 5 parameters has been used to parametrize the policy manifold. The parameters drop to 5 since we have constrained the policy manifold to pass through the optimal points. The reader may refer to the supplementary material for details. In order to show the capability of the approximate algorithm we have decided 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 instance of variable $t$, 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 $\rho$, set to $-20$. Figure 2(d) reports the final frontier obtained with different algorithms. The approximation obtained by our algorithm is
Figure 2: Experiment results. Figures (a) and (b) show some frontier obtained by the algorithm during the learning process in the 2D LQG problem without and with constraints, respectively. Only few iterations have been reported, each one with the associated iteration number, where end denote the frontier obtained when the terminal condition is reached. Figure (c) compares the true Pareto frontier and the one obtained with metric $I_3$ in the 3D LQG. Other figures ((d) and (e)) are related to the water reservoir domain and represent the Pareto frontiers obtained with different algorithms and the trend of $J(\rho)$ over iterations.

As shown in Figure 2(e), despite the low number of exploited samples, the algorithm presents a almost monotonic trend during the learning process.

7 Conclusions

In this paper we have proposed a novel gradient–based approach to learn a continuous approximation of the Pareto frontier in MOMDPs. The idea is to define a parametric function $\phi_\rho$ 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 $\phi_\rho$. 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, thus decreasing the number of samples need for the Monte Carlo estimate of the integral.
References

[1] Richard S. Sutton and Andrew G. Barto. Reinforcement Learning: An Introduction. The MIT Press, March 1998.

[2] C. Liu, X. Xu, and D. Hu. Multiobjective reinforcement learning: A comprehensive overview. IEEE Transactions on Systems Man and Cybernetics Part C, PP(99):1–13, 2013.

[3] Diederik M. Roijers, Peter Vamplew, Shimon Whiteson, and Richard Dazeley. A survey of multi-objective sequential decision-making. JAIR, 48:67–113, 2013.

[4] Peter Vamplew, Richard Dazeley, Adam Berry, Rustam Issabekov, and Evan Dekker. Empirical evaluation methods for multiobjective reinforcement learning algorithms. Machine Learning, 84(1-2):51–80, 2011.

[5] Ken Harada, Jun Sakuma, Shigenobu Kobayashi, and Isao Ono. Uniform sampling of local pareto-optimal solution curves by pareto path following and its applications in multi-objective ga. In Proceedings of GECCO ’07, pages 813–820, New York, NY, USA, 2007. ACM.

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

[7] Daniel J. Lizotte, Michael Bowling, and Susan A. Murphy. Linear fitted-q iteration with multiple reward functions. Journal of Machine Learning Research, 13:3253–3295, 2012.

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

[9] Christian Robert Shelton. Importance Sampling for Reinforcement Learning with Multiple Objectives. PhD thesis, Massachusetts Institute of Technology, August 2001.

[10] Simone Parisi, Matteo Pirotta, Nicola Smacchia, Luca Bascetta, and Marcello Restelli. Policy gradient approaches for multi-objective sequential decision making. In IJCNN 2014, Beijing, China, July 6-11, 2014, pages 1–7. IEEE, 2014.

[11] Jan Peters and Stefan Schaal. Reinforcement learning of motor skills with policy gradients. Neural Networks, 21(4):682–697, 2008.

[12] Ken Harada, Jun Sakuma, and Shigenobu Kobayashi. Local search for multiobjective function optimization: pareto descent method. In GECCO, pages 659–666, 2006.

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

[14] J.R. Munkres. Analysis On Manifolds. Adv. Books Classics Series. Westview Press, 1997.

[15] Sham Kakade. Optimizing average reward using discounted rewards. In COLT/EuroCOLT, pages 605–615. Springer, 2001.

[16] Richard S. Sutton, David A. McAllester, Satinder P. Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In NIPS, pages 1057–1063. The MIT Press, 1999.

[17] Matteo Pirotta, Marcello Restelli, and Luca Bascetta. Adaptive step-size for policy gradient methods. In NIPS 26, pages 1394–1402. Curran Associates, Inc., 2013.

[18] Raphael Fonteneau and Prashanth L. A. Simultaneous perturbation algorithms for batch off-policy search. 2014.

[19] Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. Machine Learning, 8(3-4):229–256, May 1992.

9
[20] Christian P Robert and George Casella. *Monte Carlo statistical methods*, volume 319. New York: Springer, 2004.

[21] Eckart Zitzler, Lothar Thiele, and Johannes Bader. On set-based multiobjective optimization. *Trans. Evol. Comp.*, 14(1):58–79, February 2010.

[22] Tobias Friedrich, Christian Horoba, and Frank Neumann. Multiplicative approximations and the hypervolume indicator. In *GECCO ’09*, pages 571–578, New York, NY, USA, 2009. ACM.

[23] T. Okabe, Y. Jin, and B. Sendhoff. A critical survey of performance indices for multi-objective optimisation. In *CEC ’03.*, volume 2, pages 878–885 Vol.2, Dec 2003.