ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 22-24 April 2015, i6doc.com publ., ISBN 978-287587014-8.
Available from http://www.i6doc.com/en/.

Multi-objective optimization perspectives on reinforcement learning algorithms using reward vectors

Mădălina M. Drugan[1] *

Artificial Intelligence Lab, Vrije Universiteit Brussels,
Pleinlaan 2, 1050-B, Brussels, Belgium,
e-mail: Madalina.Drugan@vub.ac.be

Abstract. Reinforcement learning is a machine learning area that studies which actions an agent can take in order to optimize a cumulative reward function. Recently, a new class of reinforcement learning algorithms with multiple, possibly conflicting, reward functions was proposed. We call this class of algorithms the *multi-objective reinforcement learning* (MORL) paradigm. We give an overview on multi-objective optimization techniques imported in MORL and their theoretical simplified variant with a single state, namely the *multi-objective multi-armed bandits* (MOMAB) paradigm.

1 Introduction

In some practical applications, such as e.g. control applications, there are often multiple criteria, or objectives, that need to be optimized at the same time. Real-world applications that motivate the usage of *multi-objective reinforcement learning* (MORL) are: 1) the wet clutch from control theory [1], and 2) traffic light control [2]. A generic and tunable problem instance generator [3] proposes large and challenging multi-objective environments. Another benchmark with test problems for MORL is proposed in [4]. The main goal of MORL algorithms is to learn and optimize an action selection in difficult and complex on-line multi-objective environments by enriching *reinforcement learning* (RL) with the intuition and computational efficiency of *multi-objective optimization* (MOO) in handing these environments that could be deterministic, stochastic or adversarial.

In this paper, we are briefing algorithms in the already identified intersecting areas between reinforcement learning and multi-objective optimization. Although they seem very different, both paradigms address decision making problems with the same goal of optimizing the reward obtained through the behaviour of an agent. Reinforcement learning [5] is a well established area of research within machine learning (ML) that solves *sequential* decision problems in an initially unknown environment. Multi-Objective Optimization [6] is considered here a subfield of *multi-criteria* decision making (MCDM) concerned with optimization of more than one objective simultaneously and where a decision

*Madalina M. Drugan was supported by the IWT-SBO project PERPETUAL (gr. nr. 110041) and FWO project "Multi-criteria RL" (gr. nr. G.087814N).

ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 22-24 April 2015, i6doc.com publ., ISBN 978-287587014-8.
Available from http://www.i6doc.com/en/.

maker decides which solutions are important and when to show these solutions to the decision maker. Currently, MOO is seldom used for stochastic optimization problems although there are important application areas, like risk analysis, that could benefit from stochastic MOO algorithms. An overview [7] on the decision problems involving multi-objective and stochastic optimization for operations research concludes that this combination is very promising although an unexplored research area.

The paper is organized as follows. Section 2 gives an introduction in reinforcement learning. Section 3 presents introductory multi-objective optimization terms and techniques used in RL. In Section 4, we present recent developments in the multi-objective reinforcement learning (MORL) paradigm. In Section 5, we focus on a special simplified case of MORL used for theoretical analysis: the multi-objective multi-armed bandits (MOMAB) paradigm. Section 6 concludes the paper.

2 Short introduction in reinforcement learning

Markov decision processes [8] (MDP) are a popular formalism to study decision-making under uncertainty where the objective is to maximize a cumulative reward value. An MDP is characterized by a set of states and a set of actions. For each state - action pair there is a probability of entering a next state, i.e. a transition probability, accompanied by a possible stochastic reward. This process is Markovian since the distribution over the next states is independent of the past through the current state and action. The action-selection mechanism in an MDP is described by a policy that specifies a probability of selecting an action in a specific state. The quality of a policy is measured by the expected discounted sum of future rewards.

Two important techniques to solve MDPs are dynamic programming and reinforcement learning. *Dynamic programming* [8] (DP) solves MDPs by breaking them down in simpler sub-problems using the Bellman equation [9], which expresses that the expected value of a state is defined in terms of an immediate reward and the expected future sum of rewards. Note that DP methods assume the transition model and reward functions are known.

Reinforcement learning (RL) is an alternative of DP where the transition and the reward functions are not known apriori. RL solves MDPs by rewarding good actions and punishing bad actions. The reward is a scalar value that can be stochastic, i.e. drawn according to a probability distribution. A negative reward expresses a punishment, and positive values express a reward. Optimizing actions means trying them out, and evaluating their long term reward and this might only be apparent after a large number of actions have been taken. RL is considered a very general learning technique that has been successfully applied to robot control, games, elevator control, etc.

An important aspect of RL is the exploration / exploitation dilemma: 1) the agent should try new, probably sub-optimal, combinations between states and actions with an exploration strategy, and 2) RL is an optimization algorithm

ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 22-24 April 2015, i6doc.com publ., ISBN 978-287587014-8.
Available from http://www.i6doc.com/en/.

that needs to return feasible, close to optimal, solutions to a control problem using an exploitation strategy. Both these strategies are not trivial and actually, in a realistic environment where the resources are limited, there is a trade-off between them because an RL algorithm that explores a lot will have little time left for exploitation and vice-versa.

Q-learning [10] is a popular model-free RL algorithm that incrementally estimates Q-values for actions based on rewards and the current Q-value function. The learner makes a step in the environment from a current state to a future state using an action while receiving a reward value. The update takes place on the Q-value of the action in the state in which the action was executed. Under certain conditions, e.g. the environment is Markovian, and the agent gradually decreases the learning rate, the system converges to the optimal policy of the MDP. From the exploration strategies used to select the next action, the ϵ-greedy exploration method selects most of the time the best action and with a small probability another action chosen at random.

Multi-armed bandits [11] (MABs) is a popular mathematical formalism for sequential decision-making under uncertainty. An agent must choose between N-arms such that the expected reward over time is maximized. The distribution of the stochastic pay-off of the different arms is assumed to be unknown to the agent. A MAB algorithm starts by uniformly exploring the N-arms, and then gradually focuses on the arm with the best observed performance. A number of real world applications can be modelled as MABs, like the famous Yahoo recommending system [12].

3 Short introduction in multi-objective optimization

The goal of MOO is to simultaneously optimize several, and usually conflicting, objective functions. The solution to a MOO-problem is not a single point, but a set of Pareto optimal solutions, i.e., solutions which cannot be further improved in any objective without worsening at least another objective. This set is usually referred to as the Pareto set. Solving a MOO-problem consists of finding the set of Pareto optimal solutions that is the set of Pareto optimal policies for MORL and the set of Pareto optimal arms in MOMABs. The Pareto front is not necessarily finite and its identification requires specific mechanisms for efficient storage, exploitation and exploration of solutions. Thus, optimizing in multi-objective environments is significantly more complex than optimizing in single objective environments also because of the increasing size of Pareto fronts for an increasing number of objectives.

Note that we could use different dominance relations to identify the Pareto front and to explore the MOO environment. The performance of a MOO algorithm could be assessed using the Pareto partial order relation where a solution is considered better than another solution iff there exists at least one objective where the first solution is better and for all other objectives is better or equal than the other solution. To explore the environment, alternative dominance relations might be used like total order relations, i.e. scalarization functions and

ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 22-24 April 2015, i6doc.com publ., ISBN 978-287587014-8. Available from http://www.i6doc.com/en/.

preference based relations.

A general criterion to classify multi-objective approaches considers different order relationships of the "goodness" of solutions:

1) *Pareto-based ranking* methods perform the search directly in the multi-objective space using the Pareto dominance relation. They are called posterior methods and their aim is to produce all the Pareto optimal solutions. Pareto-based ranking is popular in Evolutionary Computation [13] (EC) and it typically manages sets of solutions. A decision maker selects afterwards her/his preferred solution.

2) *Scalarization functions* [14] transform the multi-objective problem into a single objective problem, by combining the different values of the different objectives into a scalar using linear or non-linear functions. The advantage of this approach is that a multi-objective problem is transformed into a single objective problem that can be solved by a standard optimizer. With this approach, the decision maker could have little or no preference on the particular Pareto optimal solution returned. If a decision maker should select afterwards her/his preferred solution, the disadvantage of this method is that a range of weight vectors needs to be generated in order to identify the entire Pareto front, and this can be both time consuming and possibly inaccurate.

3) The a-priori techniques assume that the decision maker has preferences for a particular region of the Pareto optimal set and uses, for example, a utility function or a lexicographic order to rank the objectives [15].

4) The interactive methods, used for example in games, permanently interact with the user in order to select a preferred solution.

4 Multi-objective reinforcement learning paradigm

Multi-objective dynamic programming [16] and multi-objective MDP (MOMDP) [17] find their roots in the 80s where the immediate reward values are replaced with reward vectors. In general this leads to multiple optimal policies that are incomparable, each being optimal with respect to at least one of the criteria. A variety of techniques from MOO are incorporated into RL to construct efficient MORL algorithms which can learn all Pareto optimal policies. Multi-objective RL (MORL) algorithms are usually scalarization-based RL. For an overview of these methods we refer to [18].

[19] shows that convex Pareto fronts can be identified always with linear scalarization functions using MOMDPs in continuous environments with two and three objectives. [20] applies this technique in a multi-agent setting that continuously interacts with the decision making system. [21] extends these techniques to non-convex Pareto fronts. In [22] and [23], MORL walks on a continuous state Pareto front using the policy gradient algorithm.

Online MORL algorithms use linear scalarization functions with Q-learning, where Q-values are extended to Q-vectors to identify the Pareto front of policies [4]. MORL has different updating rules for the Q-vectors and the selection of the next action when compared with single objective MDP. For example, reward

ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 22-24 April 2015, i6doc.com publ., ISBN 978-287587014-8.
Available from http://www.i6doc.com/en/.

vectors are scalarized and Q-vectors are updated separately in each objective, and the next action is selected from a list generated with ϵ-greedy exploration or the hypervolume based indicator [24]. The advantage of using linear scalarization functions is that it becomes straightforward to show the convergence to the true Pareto front.

As in MOO, when a set of fixed uniformly spread, linear or non-linear, scalarization functions is used, most probably only a subset of the Pareto front is identified [25].

An adaptive set of scalarization functions [26, 27] identifies a larger number of Pareto optimal solutions [28], and thus this MORL is efficient in two and three objective environments. The correlation between two objectives is exploited in [29]. In [30], multiple copies of a single objective RL policy are used to solve a given problem.

The hypervolume unary indicator [31] commonly assesses the performance of MOO algorithms, and is recently used to guide the search as another transformation of MOO into a single objective problem [32]. The advantage of hypervolume-based search over scalarization functions is that there is no need to search for a set of functions to generate the Pareto front. The disadvantage is that the decision maker has no control over the output Pareto optimal solutions. A successful MORL algorithm uses the hypervolume indicator in the exploration mechanism [24]. In the dynamics of hypervolume-based MORL algorithm there were noticed the same downside as for MOO algorithms that use the hypervolume indicator [33]. Hypervolume-based search is also used by a Monte Carlo tree search method in [34].

In order to speed up the learning process in multi-objective environments, model-based algorithms have been designed. Intuitively, keeping a model should help, because whenever some information is gained from an exploration move, a model-based algorithm can propagate that information throughout the search space, and thus less exploration is needed. However, as pointed out in [35], this is not always the case since the extra information should be adequately stored and used. Most multi-objective DP approaches construct a list to keep track of the Pareto optimal value functions. In MOMDPs, the value function has different cumulative reward components in each objective for a policy. The set of Pareto Q-value functions in each state and the set of Q-values are stored in order to work on this set based multi-objective DP [36, 37]. Compared with single objective DP, the maxim operator is replaced with a Pareto optimal operator for sets.

This approach has severe computational problems for dynamical environments, where the reward vectors change over time. Then, the environment changes over time, and model-based RL might not exit from the learning loops, even for small, but dynamical, environments. [38] proposes a bootstrapping rule for MOMDPs similar with the bootstrapping rule from [8] to ensure convergence of Q-vectors even when the underlying environment is stochastic. [39] uses Pareto local search for optimizing policies in MOMDPs for planning.

Multi-criteria RL [15] uses preference based dominance relations to order two criteria. [40],[41] propose preference based RL algorithms.

ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 22-24 April 2015, i6doc.com publ., ISBN 978-287587014-8.
Available from http://www.i6doc.com/en/.

5 Multi-objective multi-armed bandits paradigm

A variety of techniques from MOO are adapted to MABs with reward vectors for an efficient trade-off between exploration / exploitation in difficult (i.e. complex and large) multi-objective stochastic problems. The exploration (the search for new useful solutions) versus exploitation (the use and propagation of such solutions) trade-off is an attribute of successful adaptation in both MABs and multi-objective optimization for evolutionary computation. In EC, the exploration implies the evaluation of new solutions that could have low fitness values and the exploitation means the usage of already known good solutions. In MOMABs, there is also an exploration / exploration trade-off of the Pareto front. When the set of Pareto optimal solutions in too large, we need to exploit a representative set of policies, whereas we consider an unbiased exploration of policies in this representative Pareto set.

Theoretical properties are important in designing MOMABs since they are considered simplified MORL algorithms with lower and upper bounds on performance measures. In MOMABs, the performance measures are still subject of research and include: 1) regret metrics that measure the total loss of using suboptimal arms, 2) variance in usage of Pareto optimal arms, and 3) estimated sample complexity or the estimated number of arm pulls to bound the probability of erroneously removing one of the best arms. Most MAB algorithms have one of two goals: 1) to optimize the performance, e.g. minimize the total regret resulting from how many times a suboptimal arm is used instead of an optimal arm, 2) to identify the best arm by successively deleting suboptimal arms when their lower quality bound is assigned with enough confidence.

A mechanism to minimize the regret is to identify the set of Pareto optimal arms with the highest expected reward (i.e. the Pareto front). To calculate upper confidence bounds on expected cumulative regret, we have noticed that the Hoeffding inequality [42] and union bound are extensively used in stochastic MABs. Thus, MOMABs' upper bounds use the same Hoeffding inequality that now applies the union bound over the number of dimensions and, for some algorithms only, over the number of Pareto optimal arms.

The Pareto upper confidence bound algorithm (Pareto UCB1) [43] is an infinite horizon MOMAB that extends a very popular stochastic MAB algorithm named UCB1 (upper confidence bound). Pareto UCB1 pulls each iteration one of the Pareto optimal arms according to an index composed of the arm's estimated mean and an exploration term proportional with the number of times the arm was selected. Similarly with single objective UCB, a method to ameliorate the performance of Pareto UCB1 is to increase the number of times each Pareto optimal arm is pulled [44, 45, 46]. UCB is a very general technique that could be applied on arms with any distribution. Naturally, when the distribution of the arms (e.g. Bernoulli) is also integrated in the algorithm, performance of the algorithm is improved [47, 48]. In [49], the Pareto front of continuous arms is identified using similar algorithms as in [50]. The scalarized upper confidence bound algorithm [43] extends the UCB1 to multi-objective optimization using a

ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 22-24 April 2015, i6doc.com publ., ISBN 978-287587014-8. Available from http://www.i6doc.com/en/.

fixed set of linear and non-linear scalarization functions.

An alternative goal is to identify the Pareto front or to select a representative set of arms with a bounded probability of error. Scalarized Pareto front identification [51] with the homologue MAB uses a fixed set of scalarization functions. A common approach in multi-objective optimization selects a number of weight vectors that are uniform randomly spread in the weight space. However, the performance of the algorithm heavily depends on the optimal solutions identified by the weight vectors and these solutions could be non-uniformly distributed. Techniques from evolutionary computation are used to generate new scalarization functions to identify the entire Pareto front [52]. Topological approaches [53] decompose any Pareto front in a hierarchy of convex shapes that can be identified with linear scalarization.

The knowledge gradient policy [54] is an on-line bandit-style learning method in a reinforcement learning setting where the rewards are updated using Bayesian rules. Multi-objective knowledge gradient algorithms use reward vectors and MOO in a multi-armed bandit setting [55, 56, 57].

6 Discussion

We gave an overview of current multi-objective optimization techniques used in reinforcement learning using reward vectors. Note that MORL follows closely the latest research in MOO to generate algorithms with an efficient exploration / exploitation trade-off. Although a lot of work has been done to develop this research field, there are still fundamental questions that cannot be immediately answered using knowledge from either RL or MOO. For example, what is a good (unified) performance measure for these algorithms? How will the analytical performance of MORL algorithms be affected by the inclusion of MOO techniques? There is also a need of a set of real-world applications that highlights the strengths and the weaknesses of these algorithms.

To conclude, reinforcement learning with reward vectors is a novel and promising research area with an initial slow development because of severe computational problems, which is extensively developed in the last years because of the advanced MOO techniques they incorporate. Given the rapid increase in computational power of computer systems, we predict that MORL will follow the popularity trend of RL and will be increasingly applied in applications like robot control and the internet of things.

References

[1] T. Brys, K. Van Moffaert, K. Van Vaerenbergh, and A. Nowé. On the behaviour of scalarization methods for the engagement of a wet clutch. In *12th International Conference on Machine Learning and Applications, (ICMLA)*, pages 258–263, 2013.

[2] T. Brys, T. T. Pham, and M. E. Taylor. Distributed learning and multi-objectivity in traffic light control. *Connect. Sci.*, 26(1):65–83, 2014.

[3] D. Garrett, J. Bieger, and K. R. Thórisson. Tunable and generic problem instance generation for multi-objective reinforcement learning. In *2014 IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning, (ADPRL)*, pages 1–8, 2014.

ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 22-24 April 2015, i6doc.com publ., ISBN 978-287587014-8. Available from http://www.i6doc.com/en/.

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

[5] M. Wiering and M. van Otterlo, editors. *Reinforcement Learning: State-of-the-Art*. Springer, 2012.

[6] C. A. Coello Coello. Evolutionary multi-objective optimization: A critical review. In *Evolutionary Optimization*, volume 48 of *International Series in Operations Research & Management Science*, pages 117–146. Springer US, 2002.

[7] W.J. Gutjahr and A. Pichler. Stochastic multi-objective optimization: a survey on non-scalarizing methods. *Ann Oper Res*, 2013.

[8] R. S. Sutton and A. G. Barto. *Reinforcement Learning: an introduction*. MIT Press, 1998.

[9] R. E. Bellman. *Dynamic Programming*. Princepton University Press, 1957.

[10] C. J. C. H. Watkins and P. Dayan. Technical note q-learning. *Machine Learning*, 8:279–292, 1992.

[11] S. Bubeck and N. Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5(1):1–122, 2012.

[12] Lihong Li, Wei Chu, J. Langford, and R. E. Schapire. A contextual-bandit approach to personalized news article recommendation. *CoRR*, abs/1003.0146, 2010.

[13] J.H. Holland. *Adaptation in Natural and Artificial Systems*. University of Michigan Press, 1975.

[14] G. Eichfelder. *Adaptive Scalarization Methods in Multiobjective Optimization*. Springer, 2008.

[15] Z. Gábor, Z. Kalmár, and C. Szepesvári. Multi-criteria reinforcement learning. In *Proceedings of the Fifteenth International Conference on Machine Learning (ICML)*, pages 197–205, 1998.

[16] N. Furukama. Characterization of optimal policies in vector-valued markovian decision processes. *Math. Oper. Res.*, 5(2):271–279, 1980.

[17] C. C. White and K. W. Kim. Solution procedures for vector criterion markov decision processes. *Large Scale Systems*, 1:129–140, 1980.

[18] D. M. Roijers, P. Vamplew, S. Whiteson, and R. Dazeley. A survey of multi-objective sequential decision-making. *J. Artif. Intell. Res. (JAIR)*, 48:67–113, 2013.

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

[20] D. M. Roijers, S. Whiteson, and F. A. Oliehoek. Computing convex coverage sets for multi-objective coordination graphs. In *Algorithmic Decision Theory - Third International Conference, (ADT)*, pages 309–323, 2013.

[21] D. M. Roijers, S. Whiteson, and F. A. Oliehoek. Linear support for multi-objective coordination graphs. In *International conference on Autonomous Agents and Multi-Agent Systems, (AAMAS)*, pages 1297–1304, 2014.

[22] S. Parisi, M. Pirotta, N. Smacchia, L. Bascetta, and M. Restelli. Policy gradient approaches for multi-objective sequential decision making: A comparison. In *2014 IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning, (ADPRL)*, pages 1–8, 2014.

[23] S. Parisi, M. Pirotta, N. Smacchia, L. Bascetta, and M. Restelli. Policy gradient approaches for multi-objective sequential decision making. In *2014 International Joint Conference on Neural Networks, (IJCNN)*, pages 2323–2330, 2014.

[24] K. Van Moffaert, M. M. Drugan, and A. Nowé. Hypervolume-based multi-objective reinforcement learning. In *Evolutionary Multi-Criterion Optimization - 7th International Conference, (EMO)*, pages 352–366, 2013.

[25] K. Van Moffaert, M. M. Drugan, and A. Nowé. Scalarized multi-objective reinforcement learning: Novel design techniques. In *Proceedings of the 2013 IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning, (ADPRL)*, pages 191–199, 2013.

[26] J. Dubois-Lacoste, M. López-Ibáñez, and T. Stützle. Improving the anytime behavior of two-phase local search. *Ann. Math. Artif. Intell.*, 61(2):125–154, 2011.

[27] M. M. Drugan. Sets of interacting scalarization functions in local search for multi-objective combinatorial optimization problems. In *2013 IEEE Symposium on Computational Intelligence in Multi-Criteria Decision-Making, (MCDM)*, pages 41–47, 2013.

[28] K. Van Moffaert, T. Brys, A. Chandra, L. Esterle, P. R. Lewis, and A. Nowé. A novel adaptive weight selection algorithm for multi-objective multi-agent reinforcement learning. In *2014 International Joint Conference on Neural Networks, (IJCNN)*, pages 2306–2314, 2014.

[29] T. Brys, A. Nowé, D. Kudenko, and M. E. Taylor. Combining multiple correlated reward and shaping signals by measuring confidence. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence.*, pages 1687–1693, 2014.

[30] T. Brys, A. Harutyunyan, P. Vrancx, M. E. Taylor, D. Kudenko, and A. Nowé. Multi-objectivization of reinforcement learning problems by reward shaping. In *2014 International Joint Conference on Neural Networks, (IJCNN)*, pages 2315–2322, 2014.

[31] C. M. Fonseca, J. D. Knowles, L. Thiele, and E. Zitzler. A tutorial on the peformance assessment of stochastic multi-objective optimizers. In *Evolutionary Multi-Criterion Optimization Conference, (EMO)*, 2005.

[32] J. M. Bader. *Hypervolume-Based Search for Multiobjective Optimization: Theory and Methods.* PhD thesis, ETH Zurich, 2009.

[33] E. Zitzler, D. Brockhoff, and L. Thiele. The hypervolume indicator revisited: On the design of pareto-compliant indicators via weighted integration. In *Evolutionary Multi-Criterion Optimization, 4th International Conference, (EMO)*, pages 862–876, 2006.

[34] W. Wang and M. Sebag. Multi-objective Monte Carlo tree search. In *Asian conference on Machine Learning*, pages 1–16, 2012.

[35] I. Szita and C. Szepesvári. Model-based reinforcement learning with nearly tight exploration complexity bounds. In *International Conference on Machine Learning (ICML)*, pages 1031–1038, 2010.

[36] M.A. Wiering and E.D. de Jong. Computing optimal stationary policies for multi-objective markov decision processes. In *Proc of Approximate Dynamic Programming and Reinforcement Learning (ADPRL)*, pages 158–165. IEEE, 2007.

[37] M. A. Wiering, M. Withagen, and M. M. Drugan. Model-based multi-objective reinforcement learning. In *2014 IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning, (ADPRL)*, pages 1–6, 2014.

[38] K. Van Moffaert, M. M. Drugan, and A. Nowé. Learning sets of pareto optimal policies. In *International Conference on Autonomous Agents and Multiagent Systems - Adaptive Learning Agents Workshop (ALA)*, 2014.

[39] C. Kooijman, M. de Waard, M. Inja, D. Roijers, and S. Whiteson. Pareto local search for momdp planning. In *22th European Symposium on Artificial Neural Networks, (ESANN)*, 2015.

[40] J. Fürnkranz, E. Hüllermeier, W. Cheng, and Sang-Hyeun Park. Preference-based reinforcement learning: a formal framework and a policy iteration algorithm. *Machine Learning*, 89(1-2):123–156, 2012.

[41] R. Busa-Fekete, B. Szörényi, P. Weng, W. Cheng, and E. Hüllermeier. Preference-based reinforcement learning: evolutionary direct policy search using a preference-based racing algorithm. *Machine Learning*, 97(3):327–351, 2014.

ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 22-24 April 2015, i6doc.com publ., ISBN 978-287587014-8. Available from http://www.i6doc.com/en/.

[42] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

[43] M. M. Drugan and A. Nowé. Designing multi-objective multi-armed bandits algorithms: A study. In *The 2013 International Joint Conference on Neural Networks, (IJCNN)*, pages 1–8, 2013.

[44] M. M. Drugan, A. Nowé, and B. Manderick. Pareto upper confidence bounds algorithms: An empirical study. In *2014 IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning, (ADPRL)*, pages 1–8, 2014.

[45] A. Durand, C Bordet, and C. Gagne. Improving the pareto ucb1 algorithm on the multi-objective multi-armed bandit. In *Workshop of the 27th Neural Information Processing (NIPS) on Bayesian Optimization*, 2014.

[46] M. M. Drugan and M. Bernard. Exploration versus exploitation trade-off in infinite horizon pareto multi-armed bandits algorithms. In *ICAART 2015 - Proceedings of the 6th International Conference on Agents and Artificial Intelligence*, 2015.

[47] S. Q. Yahyaa, M. M. Drugan, and B. Manderick. Annealing-pareto multi-objective multi-armed bandit algorithm. In *2014 IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning, (ADPRL)*, pages 1–8, 2014.

[48] S. Q. Yahyaa and B. Manderick. Thompson sampling for multi-objective multi-armed bandits problem. In *22th European Symposium on Artificial Neural Networks, (ESANN)*, 2015.

[49] K. Van Moffaert, K. Van Vaerenbergh, and P. Vrancx andA. Nowé. Multi-objective χ-armed bandits. In *2014 International Joint Conference on Neural Networks, (IJCNN)*, pages 2331–2338, 2014.

[50] R. Munos. Optimistic optimization of a deterministic function without the knowledge of its smoothness. In *Advances in Neural Information Processing Systems 24, (NIPS)*, pages 783–791, 2011.

[51] M. M. Drugan and A. Nowé. Scalarization based pareto optimal set of arms identification algorithms. In *2014 International Joint Conference on Neural Networks, (IJCNN)*, pages 2690–2697, 2014.

[52] S. Q. Yahyaa, M. M. Drugan, and B. Manderick. Thompson sampling in the adaptive linear scalarized multi objective multi armed bandit. In *ICAART 2015 - Proceedings of the 6th International Conference on Agents and Artificial Intelligence*, 2015.

[53] M. M. Drugan. Linear scalarization for pareto front identification in stochastic environments. In *EMO Evolutionary Multiobjective optimization*, 2015.

[54] P. Frazier, W. B. Powell, and S. Dayanik. A knowledge-gradient policy for sequential information collection. *SIAM J. Control and Optimization*, 47(5):2410–2439, 2008.

[55] S. Q. Yahyaa, M. M. Drugan, and B. Manderick. Knowledge gradient for multi-objective multi-armed bandit algorithms. In *ICAART 2014 - Proceedings of the 6th International Conference on Agents and Artificial Intelligence*, pages 74–83, 2014.

[56] S. Q. Yahyaa, M. M. Drugan, and B. Manderick. Linear scalarized knowledge gradient in the multi-objective multi-armed bandits problem. In *22th European Symposium on Artificial Neural Networks, (ESANN)*, 2014.

[57] S. Q. Yahyaa, M. M. Drugan, and B. Manderick. Multivariate normal distribution based multi-armed bandit pareto algorithm. In *European conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (EMCL/PKDD), PhD track*, 2014.