Fairness in Reinforcement Learning

Paul Weng
Shanghai Jiao Tong University, Shanghai, China
University of Michigan-Shanghai Jiao Tong University Joint Institute
paul.weng@sjtu.edu.cn

Abstract

Decision support systems and autonomous systems start to be deployed in real applications. Although their operations often impact many users or stakeholders, no fairness consideration is generally taken into account in their design, which could lead to completely unfair outcomes for some users or stakeholders. To tackle this issue, we advocate for the use of social welfare functions that encode fairness and present this general novel problem in the context of (deep) reinforcement learning, although it could possibly be extended to other machine learning tasks.

1 Introduction

Thanks to the progress in artificial intelligence and machine learning, but also notably to better sensors and increased computing power, decision support systems (DSS) and autonomous systems (AS) have started to become an integral part of our lives. A DSS can help us make better, faster and more informed decisions in complex decision-making problems where generally multiple stakeholders are involved. An AS can offer more efficient, more reactive and more adaptive control than human-operated systems or preprogrammed systems using fixed rules. However, as both DSS and AS are generally deployed among many users and may impact several stakeholders, fairness considerations become crucial for those systems to run successfully and to be accepted by all the different parties. Thus, both systems need to be efficient in their solutions, but also fair to their users or stakeholders.

Traditional (e.g., utilitarian) approaches consist in optimizing a single cumulated cost/utility function (e.g., power consumption, QoS, QoE, financial and/or ecological cost) without any fairness consideration and are therefore insufficient, because in order to reach the overall optimum, the utility of some users/stakeholders could be unjustly sacrificed. In order to take into account the welfare of each user/stakeholder, a multiobjective formulation, where each objective can be interpreted as the cost/utility of one user/stakeholder, is required. However, standard multiobjective methods generally focus on computing the set of Pareto-optimal solutions (solutions that cannot be improved on one objective, without worsening another). This is infeasible in practice because (1) this set may be extremely large, (2) in the case of AS, only one specific solution can be automatically applied and moreover, (3) Pareto-optimality itself does not encode any notion of fairness. An approach specifically designed for selecting a fair solution among the Pareto-optimal ones is therefore necessary.

As applications of artificial intelligence and machine learning start to pervade our everyday life, experts, policy makers and the general public start to realize that questions about fairness, ethics and safety are essential. Indeed, DSS and AS should not discriminate against us, should be designed to really help and not harm us. The problem presented in this paper fits in this new growing trend that proposes to enforce more human and social criteria to measure the quality of artificial systems. To achieve this goal, we describe an interdisciplinary approach that exploits results developed notably in economics (fairness models), applied mathematics (optimization and statistics) and computer science (machine learning). For concreteness, we describe it in sequential decision-making problems.

2 Background

A sequential decision-making problem can be modeled as a Markov Decision Process (MDP). In this section, we first recall this model and the reinforcement learning problem. We then summarize multiobjective optimization approaches in sequential decision-making and underline their insufficiency for tackling fairness. We finish this section with an overview of fairness modeling and fair optimization. To simplify the presentation, we assume without loss of generality that the preferences of users/stakeholders are represented as utility (e.g., reward or payoff) to be maximized.

Markov Decision Process and Reinforcement Learning. In an MDP [Puterman, 1994], an agent repeatedly observes its state, chooses an action, obtains an immediate scalar numeric reward, and moves to a new state. Solving an MDP (i.e., *planning*) amounts to finding a controller (called a *policy*) in order to maximize a standard decision criterion, e.g., the expected discounted reward or the expected average reward. While in planning problems, the model of the environment (e.g., transition and reward functions) is assumed to be known, in reinforcement learning (RL) problems [Sutton and Barto, 1998], this assumption is relaxed: an RL agent learns a best policy while interacting with the unknown environment by trial and error.

Thanks to their generality, those frameworks (MDP and RL) have been successfully applied in many diverse domains. For instance, MDPs and its extensions have been used for data center control [Weng *et al.*, 2018] or ecological conservation [Chadès *et al.*, 2012]. RL have been applied to robotics [Peters *et al.*, 2003] or medicine [Pilarski *et al.*, 2011]. The past few years, research in RL has become very active since the recent successes of the combination of deep learning and RL (called deep RL), notably in video games [Mnih *et al.*, 2015].

Multiobjective Sequential Decision-making The standard models for sequential decision-making have been extended to the multiobjective (MO) setting [Roijers *et al.*, 2013; Liu *et al.*, 2015] (i.e., the immediate scalar numeric reward is replaced by a vector reward whose components represent objectives) where for instance, an objective can be interpreted in the multicriteria setting as a criterion (e.g., QoS, power consumption, monetary gain) to be optimized, or in the multi-user/stakeholder setting as the welfare of a user/stakeholder (e.g., average waiting times of car in different lanes for the traffic light control problem or QoS for different users for the data center control problem). Most work in MO optimization (MOO) focuses on the multicriteria interpretation. In this paper, we focus on the second interpretation, which naturally leads to fairness considerations (see the next part entitled *Fair Optimization*).

The usual approach in MOO aims at finding the Pareto front, which is the set of all Pareto-optimal solutions [Vamplew *et al.*, 2011]. Unfortunately, computing the Pareto front is in general infeasible because the number of Pareto-optimal solutions can grow exponentially with the size of the problem [Perny *et al.*, 2013]. This observation may justify the computation of an approximation of those sets [Lizotte *et al.*, 2010; Pirotta *et al.*, 2015]. However, even with approximated sets, the approach is not suitable in autonomous systems where only one solution has to be automatically applied.

A solution to this issue relies on using a function that aggregates the objectives into a scalar value in order to select one solution among the set of Pareto-optimal solutions. However, one important point to realize is that the naive approach consisting in aggregating all the objectives with a weighted sum is insufficient. Indeed, such a linear aggregation generally does not provide much control on the trade-offs between the objectives. Moreover, non-supported (i.e., not on the convex hull) Pareto-optimal solutions cannot be obtained whatever the choice of the weights.

More interesting aggregating functions are non-linear and must be strictly increasing (in order to be monotonic with respect to Pareto dominance). Such an approach is generally called compromise programming in the multicriteria setting, which generally consists in minimizing a distance to an ideal point [Steuer, 1986]. Some work has been done for different functions in sequential decision making [Perny and Weng, 2010; Ogryczak *et al.*, 2013]. In the next paragraph, we present an aggregation function for modeling fairness, which we call fair welfare function.

Fair Optimization Fairness is a concept that has conventionally been studied in economics [Moulin, 2004] and political philosophy [Rawls, 1971]. Recently, it has also become an important consideration in other applied fields, such as in applied mathematics [Ogryczak *et al.*, 2014] which focuses on solving fair optimization problems, in artificial intelligence [de Jong *et al.*, 2008; Hao and Leung, 2016] when investigating multi-agent systems, or in computer engineering [Shi *et al.*, 2014] when designing computer networks. As shown by recent surveys [Ogryczak *et al.*, 2014; Luss, 2012], fair optimization is an active and recent research area. Although fairness is a key notion when dealing with multiple parties, it has only recently received attention in machine learning [Busa-Fekete *et al.*, 2017; Speicher *et al.*, 2018; Agarwal *et al.*, 2018; Heidari *et al.*, 2018]. To the best of our knowledge, the only work related to fairness in reinforcement learning investigate this issue in the multi-armed bandit setting [Busa-Fekete *et al.*, 2017].

Fairness can be defined in a theoretically-founded way [Moulin, 2004] and relies on two key principles. The first one (P1) is called "Equal treatment of equals", which states that two users/stakeholders (with identical characteristics with respect to the optimization problem, as assumed in this paper) should be treated the same way. The second one (P2), called the *Pigou-Dalton principle*, is based on the notion of *Pigou-Dalton transfer*, which is a payoff transfer from a richer user/stakeholder to a poorer one without reversing their relative ranking. The Pigou-Dalton principle states that such transfers lead to more equitable distributions. Formally, for any $\boldsymbol{v} \in \mathbb{R}^n$ where $v_i < v_j$ and for any $\epsilon \in (0, v_j - v_i)$ we prefer $\boldsymbol{v} + \epsilon \mathbf{1}_i - \epsilon \mathbf{1}_j$ to $\boldsymbol{v}$ where $\mathbf{1}_i$ (resp. $\mathbf{1}_j$) is the canonical vector, null everywhere except in component i (resp. j) where it is equal to 1. In words, this principle states that, all other things being equal, we prefer more "balanced" distributions (i.e., vectors) of payoffs. Beside those two principles, as we are in an optimization context, an efficiency principle (P3) is also required, which states that given two payoff distributions, if one vector Pareto-dominates another, the former is preferred to the latter.

Those three principles imply that a fair welfare function that aggregates the payoffs of the users/stakeholders need to satisfy three properties. They have to be symmetric (i.e., independent to the order of its arguments for P1), strictly Schur-concave (i.e., monotonic with respect to Pigou-Dalton transfers for P2) and strictly increasing (i.e., monotonic with respect to Pareto dominance for P3). The elementary approach based on maximin (or Egalitarian approach), where one aims at maximizing the worse-off user/stakeholder, does not satisfy the last two properties. A better approach [Rawls, 1971] is based on the lexicographic maximin, which consists in comparing first the worse-off user/stakeholder when comparing two vectors, then in case of a tie, comparing the second worse-off and so on. However, due to the non-compensatory nature of the min operator, vector $(1, 1, \ldots, 1)$ would be preferred to $(0, 100, \ldots, 100)$, which may be debatable.

Many fair welfare function have been proposed. In practice, the choice of a suitable function depends on the application domain. For illustration, we present the fair welfare function based on the Generalized Gini Index (GGI) [Weymark, 1981] $G_{\boldsymbol{w}} : \mathbb{R}^n \to \mathbb{R}$:

$$G_{\boldsymbol{w}}(\boldsymbol{v}) = \sum_i w_i v_i^{\uparrow} \qquad (1)$$

where $\boldsymbol{w} \in [0,1]^n$ is a weight vector such as $w_1 > w_2 > \ldots > w_n$, and $(v_1^{\uparrow}, v_2^{\uparrow}, \ldots, v_n^{\uparrow})$ is the payoff vector $\boldsymbol{v}$ reordered in an increasing fashion.

Functions $G_{\boldsymbol{w}}$ contains the welfare function induced by the classic Gini index or the Bonferroni index [Tarsitano, 1990]. It tends to the Egalitarian approach when $w_2 \to 0, \ldots, w_n \to 0$ and to the lexicographic maxmin when differences between weights tends to infinity.

GGI has been exploited in different MO (continuous and combinatorial) optimization problems. To cite a few, it was used in capital budgeting [Kostreva *et al.*, 2004], allocation problems [Nguyen and Weng, 2017], or flow optimization in wireless mesh networks [Hurkala and Sliwinski, 2012].

3 Problem Formulation

At a high-level, a fair sequential decision-making problem can be understood as solving a non-linear convex optimization problem[1], where the welfare function, which encodes both efficiency and fairness, aggregates the utility of each user/stakeholder:

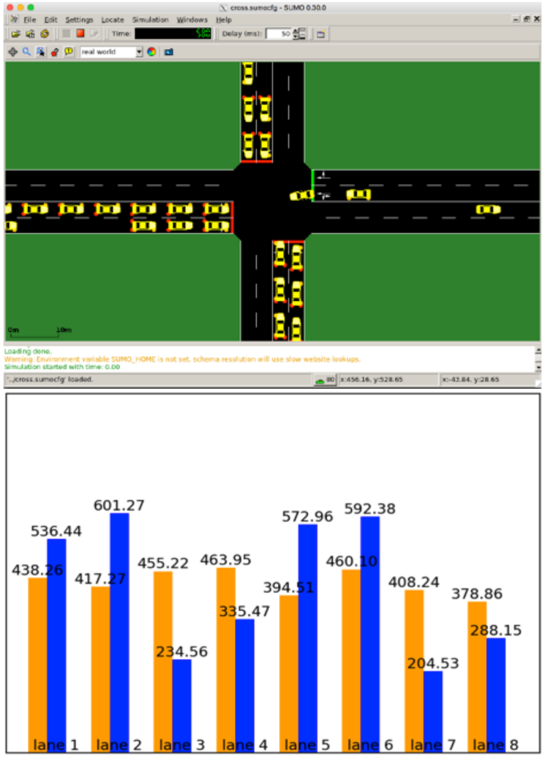
$$\max_{\pi} J(\pi) = H\Big(\sum_s \mu(s) \boldsymbol{V}^{\pi}(s)\Big) \qquad (2)$$

where π is a policy, H is a fair welfare function (e.g., GGI), μ is a probability distribution over initial states, and $\boldsymbol{V}^{\pi}$ is the multiobjective value function of π (e.g., expected discounted or average reward).

The difficulty of this new problem lies in the non-linearity of the objective function, which changes the properties of optimal policies and prevents the direct application of dynamic programming or temporal difference methods. However, the properties (e.g., concavity, Schur-concavity, decomposability, etc) of fair welfare functions and those (e.g., temporal structure) of sequential decision-making problems can be exploited to design efficient methods to find fair policies.

4 Preliminary Experimental Results

To demonstrate the potential usefulness of our proposition, we conducted some initial experiments in a traffic light control problem, because such environments are relatively easy to simulate. We use SUMO[2] (see on the top of Fig. 1 for an illustration) to simulate one intersection with a total of 8 lanes under varying traffic conditions. Standard approaches to solve this problem usually minimize the expected waiting times over all lanes. In our formulation, we learn a traffic controller that attempts to minimize the expected waiting times of each lane, while ensuring some notion of fairness over each lane is enforced. In our experiments, we used the generalized Gini index and adapted the DQN algorithm [Mnih *et al.*, 2015] to approximately optimize it. Although we illustrate the approach on the traffic light domain, the method could be applied to diverse other sequential-decision-making problems.



```
cross.sumocfg - SUMO 0.30.0
Loading done.
Warning: Environment variable SUMO_HOME is not set, schema resolution will use slow website lookups.
Simulation started with time: 0.00
'../cross.sumocfg' loaded.
```

Figure 1: Left: Screenshot of the SUMO simulator; Right: Average waiting times for standard DQN (blue) vs GGI-DQN (orange).

Fig. 1 (bottom) shows some initial results (averaged over 20 runs) where we compare our proposed approach (GGI-DQN in orange) with the standard approach (DQN in blue) that minimizes the expected waiting times over all lanes. As expected DQN obtains a lower average waiting times over all lanes (as it optimizes this criterion) than GGI-DQN: 420.72 vs 427.05 (in timesteps in the simulator). However, the average waiting times *in* each lane for the standard approach have an unequal distribution, while our approach provides a much fairer distribution of waiting times.

5 Conclusion

In this paper, we argued for the use of fair welfare functions in machine learning tasks and demonstrated it more specifically in reinforcement learning. We believe that the topic of fair optimization is novel in machine learning and is of great significance, as it naturally provides solutions that take into account the welfare of all the involved parties. As future work, we plan to develop more efficient algorithms in the deep RL setting for optimizing different fair welfare functions, and possibly extend the approach to other machine learning tasks.

[1] The objective function is convex when minimizing costs and concave when maximizing utilities. As customary in the optimization literature, we may refer to both problems as convex optimization problems.

[2] http://sumo.dlr.de/index.html

Acknowledgments

We would like to thank the anonymous reviewers for their helpful feedback. This work is supported by the Shanghai NSF (No. 19ZR1426700) and in part by the program of National Natural Science Foundation of China (No. 61872238).

References

A. Agarwal, A. Beygelzimer, M. Dudík, J. Langford, and H. Wallach. A reductions approach to fair classification. In *ICML*, 2018.

R. Busa-Fekete, B. Szörenyi, P. Weng, and S. Mannor. Multi-objective bandits: Optimizing the generalized Gini index. In *ICML*, 2017.

I. Chadès, J. Carwardine, T.G. Martin, S. Nicol, R. Sabbadin, and O. Buffet. MOMDPs: a solution for modelling adaptive management problems. In *AAAI*, 2012.

S. de Jong, K. Tuyls, and K. Verbeeck. Fairness in multi-agent systems. *The Knowledge Engineering Review*, 23(2):153–180, 2008.

J. Hao and H. Leung. *Fairness in Cooperative Multiagent Systems*, pages 27–70. Springer, Berlin, Heidelberg, 2016.

H. Heidari, C. Ferrari, K.P. Gummadi, and A. Krause. Fairness behind a veil of ignorance: A welfare analysis for automated decision making. In *NeurIPS*, 2018.

J. Hurkala and T. Sliwinski. Fair flow optimization with advanced aggregation operators in wireless mesh networks. In *Federated Conference on Computer Science and Information Systems*, pages 415–421, 2012.

M.M. Kostreva, W. Ogryczak, and A. Wierzbicki. Equitable aggregations and multiple criteria analysis. *Eur. J. Operational Research*, 158:362–367, 2004.

C. Liu, X. Xu, and D. Hu. Multiobjective reinforcement learning: A comprehensive overview. *IEEE Trans. on Systems, Man and Cybernetics*, 45(3):385–398, 2015.

D.J. Lizotte, M. Bowling, and S.A. Murphy. Efficient reinforcement learning with multiple reward functions for randomized controlled trial analysis. In *ICML*, 2010.

H. Luss. *Equitable Resource Allocation*. Wiley, 2012.

V. Mnih, K. Kavukcuoglu, D. Silver, A.A. Rusu, J. Veness, M.G. Bellemare, A. Graves, M. Riedmiller, A.K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518:529–533, 2015.

H. Moulin. *Fair Division and Collective Welfare*. MIT Press, 2004.

V.H. Nguyen and P. Weng. An efficient primal-dual algorithm for fair combinatorial optimization problems. In *COCOA*, 2017.

W. Ogryczak, P. Perny, and P. Weng. A compromise programming approach to multiobjective Markov decision processes. *International Journal of Information Technology & Decision Making*, 12:1021–1053, 2013.

W. Ogryczak, H. Luss, M. Pióro, D. Nace, and A. Tomaszewski. Fair optimization and networks: A survey. *Journal of Applied Mathematics*, 2014, 2014.

P. Perny and P. Weng. On finding compromise solutions in multiobjective Markov decision processes. In *ECAI (short paper)*, 2010.

P. Perny, P. Weng, J. Goldsmith, and J. Hanna. Approximation of Lorenz-optimal solutions in multiobjective Markov decision processes. In *UAI*, 2013.

J. Peters, S. Vijayakumar, and S. Schaal. Reinforcement learning for humanoid robotics. In *IEEE-RAS international conference on humanoid robots*, pages 1–20, 2003.

P.M. Pilarski, M.R. Dawson, T. Degris, F. Fahimi, J.P. Carey, and R.S. Sutton. Online human training of a myoelectric prosthesis controller via actor-critic reinforcement learning. In *IEEE International Conference on Rehabilitation Robotics*, 2011.

M. Pirotta, S. Parisi, and M. Restelli. Multi-objective reinforcement learning with continuous pareto frontier approximation. In *AAAI*, 2015.

M.L. Puterman. *Markov decision processes: discrete stochastic dynamic programming*. Wiley, 1994.

J. Rawls. *The Theory of Justice*. Havard university press, 1971.

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

H. Shi, R.V. Prasad, E. Onur, and I.G.M.M. Niemegeers. Fairness in wireless networks:issues, measures and challenges. *IEEE Communications Surveys & Tutorials*, 16(1):5–24, 2014.

T. Speicher, H. Heidari, N. Grgic-Hlaca, K.P. Gummadi, A. Singla, A. Weller, and M.B. Zafar. A unified approach to quantifying algorithmic unfairness: Measuring individual & group unfairness via inequality indices. In *KDD*, pages 2239–2248, 2018.

R.E. Steuer. *Multiple criteria optimization*. John Wiley, 1986.

R.S. Sutton and A.G. Barto. *Reinforcement learning: An introduction*. MIT Press, 1998.

A. Tarsitano. *Income and Wealth Distribution, Inequality, and Poverty*, chapter The Bonferroni Index of Income Inequality. Springer-Verlag, 1990.

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.

P. Weng, Z. Qiu, J. Costanzo, X. Yin, and B. Sinopoli. Optimal threshold policies for robust data center control. *Journal of Shanghai Jiao Tong University*, 23(1):52–60, 2018.

J.A. Weymark. Generalized Gini inequality indices. *Mathematical Social Sciences*, 1:409–430, 1981.