Approximation of Lorenz-Optimal Solutions in Multiobjective Markov Decision Processes

Patrice Perny  
UPMC, LIP6  
patrice.perny@lip6.fr

Paul Weng  
UPMC, LIP6  
paul.weng@lip6.fr

Judy Goldsmith  
University of Kentucky  
goldsmit@cs.uky.edu

Josiah P. Hanna  
University of Kentucky  
jpha226@g.uky.edu

**Introduction**

We consider the problem of finding small representative sets of Lorenz-optimal policies for MOMDPs (Multiobjective Markov Decision Processes). MDPs model planning under uncertainty; most work on MOMDPs finds sets of Pareto-optimal policies, i.e., policies whose expected discounted total reward vectors cannot be improved on one objective without being downgraded on another objective. Unfortunately, if we allow randomized policies, the Pareto set of policies may be infinite; even if we restrict to deterministic policies, there are families of MOMDPs where the sizes of the Pareto sets grow exponentially in the number of MDP states (Ogryczak, Perny, and Weng 2011).

Earlier work looks at polynomial-sized representative samples of the Pareto set (Papadimitriou and Yannakakis 2000; Chatterjee, Majumdar, and Henzinger 2006). Here, we seek a stronger notion than Pareto optimality. It is based on Lorenz dominance, a partial preference order refining Pareto dominance while including an idea of fairness in preferences. It is used for the measurement of inequalities in mathematical economics (Shorrocks 1983), for example to compare income distributions over a population. In our context, it can be used to compare reward vectors by inspecting how they distribute rewards over objectives. We describe algorithms for finding small, representative subsets of the Lorenz-optimal policies for MOMDPs.

**Background**

A *Markov Decision Process (MDP)* is a tuple \((S, A, p, r)\) where: \(S\) is a finite set of states, \(A\) is a finite set of actions, \(p : S \times A \times S \to [0, 1]\) is a transition function giving, for each state and action the probability of reaching a next state, and \(r : S \times A \to \mathbb{R}_+\) is a reward function giving the immediate reward for executing a given action in a given state (Puterman 1994).

Solving an MDP amounts to finding a policy, i.e., determining which action to choose in each state, which maximizes the expected discounted total reward.

A *Multiobjective MDP (MOMDP)* is defined as an MDP with the reward function replaced by \(r : S \times A \to \mathbb{R}^n_{+}\), where \(n\) is the number of criteria, \(r(s, a) = (r_1(s, a), \ldots, r_n(s, a))\) and \(r_i(s, a)\) is the immediate reward for objective \(i\). Now, a policy \(\pi\) is valued by a value function \(V^\pi : S \to \mathbb{R}^n_{+}\), which gives the expected discounted total reward vector in each state.

**Weak Lorenz dominance** is defined as: \(\forall v, v' \in \mathbb{R}^n_{+}, v \succ_P v' \iff \forall i = 1, \ldots, n, v_i \geq v'_i\) where \(v = (v_1, \ldots, v_n)\) and \(v' = (v'_1, \ldots, v'_n)\) and Pareto dominance as: \(v \succ_P v' \iff v \succ_P v'\) and not \((v' \succ_P v)\). For \(X \subseteq \mathbb{R}^n_{+}\), a vector \(v \in X\) is said to be *P-dominated* if there is another vector \(v' \in X\) such that \(v' \succ_P v\); vector \(v\) is said to be *P-optimal* if there is no vector \(v'\) such that \(v' \succ_P v\). For \(X \subseteq \mathbb{R}^n_{+}\), the Pareto-optimal vectors, called the *Pareto set*, is PND\((X) = \{v \in X : \forall v' \in X, \text{ not } v' \succ_P v\}\).

In MOMDPs, for a given probability distribution \(\mu_s\) over initial states, a policy \(\pi\) is preferred to a policy \(\pi'\) if \(\sum s \mu_s V^\pi(s) \succ_P \sum s \mu_s V^{\pi'}(s)\). Standard methods for MDPs can be extended to solve MOMDPs by finding Pareto-optimal policies (White 1982; Furukawa 1980; Viswanathan, Aggarwal, and Nair 1977). The idea of fairness can be introduced using the so-called transfer principle: any utility transfer from an objective to a worse-off objective leads to an improvement. Combining this idea with Pareto dominance amounts to considering Lorenz dominance (for more details, see e.g., (Marshall and Olkin 1979; Shorrocks 1983)).

**Definition 1** For all \(v \in \mathbb{R}^n_{+}\), the Lorenz Vector associated to \(v\) is the vector:

\[
L(v) = (v(1), v(1) + v(2), \ldots, v(1) + v(2) + \ldots + v(n))
\]

where \(v(1) \leq v(2) \leq \ldots \leq v(n)\) represents the components of \(v\) sorted by increasing order.

Hence, the Lorenz dominance relation (L-domination for short) on \(\mathbb{R}^n_{+}\) is defined by:

\[
\forall v, v' \in \mathbb{R}^n_{+}, v \succeq_L v' \iff L(v) \succeq_P L(v').
\]

Its asymmetric part is defined by:

\[
v \succ_L v' \iff L(v) \succ_P L(v').
\]

For example if \(x = (10, 10)\) and \(y = (20, 0)\) then \(L(x) = (10, 20)\) and \(L(y) = (0, 20)\). Hence \(x \succ_L y\) since \(L(x) \succ_P L(y)\). Within a set \(X\), any element \(v\) is said to be L-dominated when \(v' \succ_L v\) for some \(v'\) in \(X\), and L-optimal when there is no \(v'\) in \(X\) such that \(v' \succ_L v\). The set
of L-optimal elements in X, called the Lorenz set, is denoted $LND(X)$. The number of Lorenz-optimal tradeoffs is often significantly smaller than the number of Pareto-optimal tradeoffs.

**Approximation of PND & LND**

**Definition 2** For any $\varepsilon > 0$, the $\varepsilon$-dominance relation is defined on value vectors of $\mathbb{R}^n_+$ as follows:

$$x \succeq_{P} y \iff \forall i \in N, (1 + \varepsilon)x_i \geq y_i.$$ 

For any $\varepsilon > 0$ and any set $X \subseteq \mathbb{R}^n_+$ of bounded value vectors, a subset $Y \subseteq X$ is said to be an $\varepsilon$-P-covering of $PND(X)$ if $\forall x \in PND(X), \exists y \in Y : y \succeq_{P} x$.

For any $\varepsilon > 0$ and any set $X \subseteq \mathbb{R}^n_+$ of bounded value vectors, a subset $Y \subseteq X$ is said to be an $\varepsilon$-L-covering of $LND(X)$ if $\forall x \in LND(X), \exists y \in Y : y \succeq_{L} x$, i.e., $L(y) \succeq_{L} L(x)$.

Given a set $X \subseteq \mathbb{R}^n_+$, an indirect procedure for constructing an $\varepsilon$-L-covering of $LND(X)$ (of bounded size) is first to construct an $\varepsilon$-P-covering of $PND(X)$ denoted $PND_\varepsilon(X)$ and then to compute $LND(PND_\varepsilon(X))$, the set of L-optimal elements in $PND_\varepsilon(X)$. The first step of this procedure can be achieved using a FPTAS proposed in (Chatterjee, Majumdar, and Henzinger 2006). We now present a direct procedure for constructing an $\varepsilon$-L-covering of $LND(X)$ (of bounded size), denoted $LND_\varepsilon(X)$ hereafter, without first approximating the Pareto set. Our procedure relies on the following observation: $Y$ is an $\varepsilon$-L-covering of $LND(X)$ if $L(Y) = \{L(y) : y \in Y\}$ is an $\varepsilon$-P-covering of $PND(L(X))$. Therefore we just have to construct an $\varepsilon$-P-covering of $PND(L(X))$. We now present the construction.

To any feasible tradeoff $x \in X$, we can assign a vector $\psi(x)$ where $\psi(x)_i = \log L_i(x) / \log(1 + \varepsilon)$. Function $\psi$ defines a logarithmic hypergrid on $L(X)$ rather than on $X$. Any hypercube defined by $\psi$ in the hypergrid represents a class of value vectors that all have the same image through $\psi$. Any Lorenz vector $L(x)$ belonging to a given hypercube covers any other Lorenz vector $L(y)$ of the same hypercube in terms of $\succeq_{P}$. Hence, the original vectors $x, y$ are such that $x \succeq_{P} y$. Moreover we have: $\forall x, y \in X, \psi(x) \succeq_{P} \psi(y) \Rightarrow x \succeq_{L} y$. Thus, we can use $P$-optimal $\psi$ vectors to construct an $\varepsilon$-L-covering of the Lorenz set.

Due to the scaling and rounding operations, the number of different possible values for $\psi$ is bounded on the $i^{th}$ axis by $\lceil \log K / \log(1 + \varepsilon) \rceil$, where $K$ is an upper bound such that $0 < x_i < K$. Then we have shown that the cardinality of set $\psi(X) = \{\psi(x) : x \in X\}$ is upper bounded by $\Pi_{i=1}^n \lceil \log K / \log(1 + \varepsilon) \rceil \leq \lceil \log K / \log(1 + \varepsilon) \rceil^n$. Hence, by choosing one representative in each of these hypercubes, we cover the entire set $L(X)$ with a number of solutions polynomial in the size of the problem and $1/\varepsilon$, for any fixed $n$. The resulting covering set can further be reduced in polynomial time to $PND(\psi(X))$. Moreover, using an approach based on Linear Programming, we have shown that any hypercube can be inspected in polynomial time, so this direct approach based on the grid defined in the Lorenz space provides a FPTAS for the set of L-optimal value vectors in MOMDPs.

This procedure provides an $\varepsilon$-covering of L-optimal randomized policies. Whenever we want to restrict the search to deterministic policies, a similar procedure applies after adding integrity constraints as proposed in (Dolgov and Durfee 2005). In this case, we need to solve a mixed integer linear program for every hypercube in the hypergrid. Moreover, in the bi-objective case, we have refined this approach to construct $\varepsilon$-coverings of minimal cardinality, for $P$-optimal and L-optimal elements as well, by adapting a greedy algorithm proposed in (Diakonikolas and Yannakakis 2009).

We tested the different methods presented (or implied) in this short paper on random instances of MOMDPs. The rewards on each objective were randomly drawn from $\{0, 1, \ldots, 99\}$. All the experiments were run on standard PCs with 8GB of memory and an IntelCore 2 Duo 3.33GHz GHz processor. All LPs were solved using Gurobi 5.0. All the experimental results are averaged over 10 runs with discount factor $\gamma = 0.9$.

First, we show, in the bi-objective case, the size of $\varepsilon$-coverings being reduced using the greedy approach for Pareto and Lorenz. We set $|S| = 200$, $|A| = 5$, and $\varepsilon = 0.01$. Figure 1 shows value vectors in the objective space for one random instance of MOMDP. $PND_\varepsilon$ (resp. $LND_\varepsilon$) is an $\varepsilon$-P-covering of $PND_\varepsilon$ (resp. $\varepsilon$-L-covering of $LND_\varepsilon$), min $PND_\varepsilon$ and min $LND_\varepsilon$ are the minimal $\varepsilon$-cover sets.

In the next experiments, we give the computation times (in seconds) for computing the $\varepsilon$-coverings. Here, $|S| = 50$, $|A| = 5$, the number of objectives is 3, and $\varepsilon \in \{0.05, 0.1, 0.15, 0.2\}$. The results in Table 1 show that the direct approach is the most efficient.

**Table 1: Computation times of $\varepsilon$-covers**

| $\varepsilon$ | 0.05 | 0.1  | 0.15 | 0.2  |
|--------------|------|------|------|------|
| $LND(PND_\varepsilon)$ | 265.7 | 169.4 | 126.7 | 101.7 |
| $LND_\varepsilon$ | 5.4   | 4.8   | 4.4   | 4.2   |

**Acknowledgments**

Funded by the French National Research Agency under grant ANR-09-BLAN-0361; The Duncan E. Clarke Memorial Innovation Award, and NSF grants NSF EF-0850237 and CCF-1049360.
References

Chatterjee, K.; Majumdar, R.; and Henzinger, T. 2006. Markov decision processes with multiple objectives. In STACS.

Diakonikolas, I., and Yannakakis, M. 2009. Small approximate pareto sets for biobjective shortest paths and other problems. SIAM Journal on Computing 39(4):1340–1371.

Dolgov, D., and Durfee, E. 2005. Stationary deterministic policies for constrained MDPs with multiple rewards, costs, and discount factors. In IJCAI.

Furukawa, N. 1980. Characterization of optimal policies in vector-valued Markovian decision processes. Mathematics of operations research 5(2):271–279.

Marshall, A., and Olkin, I. 1979. Inequalities: Theory of Majorization and its Applications. Academic Press.

Ogryczak, W.; Perny, P.; and Weng, P. 2011. On minimizing ordered weighted regrets in multiobjective Markov decision processes. In Int. Conf. on Algorithmic Decision Theory.

Papadimitriou, C., and Yannakakis, M. 2000. On the approximability of trade-offs and optimal access of web sources. In FOCS, 86–92.

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

Shorrocks, A. 1983. Ranking income distributions. Economica 50:3–17.

Viswanathan, B.; Aggarwal, V.; and Nair, K. 1977. Multiple criteria Markov decision processes. TIMS Studies in the Management Sciences 6:263–272.

White, D. 1982. Multi-objective infinite-horizon discounted Markov decision processes. J. Math. Anal. Appl. 89:639–647.