Scalable Solutions for Markov Decision Processes with Weighted State Aggregation

Olivier Tsemogne[1], Alexandre Reiffers-Masson[1], Lucas Drumetz[1]
IMT-Atlantique
{serge-olivier.tsemogne-kamguia, alexandre.reiffers-masson, lucas.drumetz}@imt-atlantique.fr

Keywords : *operations research, optimization.*

1 Introduction

Solving large-scale Markov Decision Processes (MDPs) poses a significant challenge in the field of artificial intelligence. These MDPs are characterized by a vast or even continuous set of states, making their direct resolution impractical. The crux of this challenge lies in the fact that optimization at each step must account for subsequent steps. Various approximations of the optimal value have been proposed, relying on the transfer of solutions from smaller MDPs. Some authors [4, 2] employ distances between transition probabilities, while others [3, 1] aggregate states into classes, solving a smaller MDP and extrapolating the solution to the larger MDP.

In this article, we address this issue by drawing inspiration from a method introduced by [1], which involves grouping states into classes, transferring transition and reward dynamics to these classes, and then solving the resulting, abstract, smaller MDP. The solution to the original MDP is derived from that of the abstract MDP through extrapolation of the optimal policy. While [1] considers a representative state within each class to transpose dynamics to the classes, we propose a more general approach where each state in the class is assigned a weight representing its "contribution" or "importance" to the class. Drawing insights from resource allocation in queues, we experimentally demonstrate that the choice of a representative is not necessarily the best weight distribution within the classes.

2 Methodology

By employing the "state abstraction" dimension reduction technique, we address the general case of weight distributions within the classes, including distributions without density. Specifically, we assume that a function π, referred to as abstraction, is defined from the set of states to a finite set $\mathcal{U}$ of arbitrary choice. Thus, the pre-image $\pi^{-1}(u)$ of any element u in $\mathcal{U}$ forms a class of states in the MDP to be resolved. We then establish that a weight distribution is defined within each class. Subsequently, for each state-action pair, the reward function and the probability measure of transition corresponding to that action taken in a state of that class are defined as weighted averages of the reward functions and probability measures of transition corresponding to that action taken in the states of that class. Thus, an abstract MDP is defined by transferring the dynamics of the MDP to be resolved. We solve it and transfer its optimal policy to the MDP to be resolved through constant extrapolation, bounding the error to the optimal value in both the general case and the case where reward functions, transitions, and the optimal value of the MDP to be resolved are Lipschitz.

We then evaluate this error in a practical simulation carried out in the context of dynamic resource allocation. We focus on the specific scenario where a scheduler receives an unlimited number of jobs and must transmit them one at a time to queues with the same capacity but different service rates. We aggregate states based on queue occupancy, considering various weight distributions within the classes, including representative selection, to assess the distribution that minimizes the error.

3 Results

The weight distribution within the classes generalizes the selection of representatives, which indeed corresponds to a Dirac measure in each class. Therefore, we demonstrate that the bound obtained in [1] for the difference $\left\| V^* - V^{\widetilde{\pi}^*} \right\|_\infty$ between the optimal value V^* and the value of the extrapolation $\widetilde{\pi}^*$ of the optimal policy of the reduced MDP is bounded by $\frac{2}{(1-\gamma)^2}\Delta^{\max} V^*$, where γ is the discount factor and $\Delta^{\max}$ is the operator that returns the maximum difference between the backup iteration of a value function using exact dynamics and the backup iteration using approximate dynamics obtained by extrapolation. However, we do not generalize the result obtained in the case where the dynamics are Lipschitz.

We conducted simulations to assess the relevance of weighting states within classes rather than selecting representatives in each class for solving a resource allocation problem. In our scenario, we have a scheduler responsible for allocating incoming jobs, one at a time, to a fixed number of queues with varying service rates. The scheduler is rewarded for each successfully processed job. To simplify the problem, we group the queues based on their service rates, and we abstract each occupancy state of the queues by considering the total number of jobs in each group of queues. These groupings define our classes of states. Then we weighted the queues based on the occupancies of individual queues. Regarding class representation, we explored three possibilities: (1) Selecting the top-weighted state within each class; (2) Selecting the top-weighted state within each class and giving equal weight to all states in the class; (3) Considering all states within the class without selecting a representative. Our simulations revealed interesting findings. Some weight distributions outperformed the approach of selecting a single representative state within each class, and conversely, some representative selection strategies outperformed certain weight distributions. Table 1 illustrates these comparative results. Due to space constraint, we cannot provide more details.

TAB. 1: Difference between the optimal value and its approximation
Original problem: 5 queues of length 2 each. Abstract problem: 3 groups of queues. Selecting o a single representative provides across the states the lowest maximum error, and the highest average error.

Selected	Difference with $\gamma = 0.4$	Difference with $\gamma = 0.7$
One	1.49	3.54
Top	1.48	3.56
All	1.48	3.56

Other simulations show that certain weight distributions reduce the resolution time of a problem to less than 1/1000 while yielding an almost optimal policy.

4 Conclusion

Overall, our research generalizes existing results and addresses resource allocation problems with multiple weight distributions while evaluating the quality of our approximations. Our results significantly extend the limits in terms of queue capacity and the number of queues in solving resource allocation problems. We hope that this contribution will provide valuable insights into the realm of large-scale MDPs and resource allocation problem resolution. We would be honored to present our work at the Roadef 2024 conference.

Please note that you should complete the title of the article and the list of authors before submitting this abstract to the Roadef 2024 conference.

References

[1] Berk Bozkurt, Aditya Mahajan, Ashutosh Nayyar, and Yi Ouyang. Weighted-norm bounds on model approximation in mdps with unbounded per-step cost.

[2] Norm Ferns, Prakash Panangaden, and Doina Precup. Metrics for finite markov decision processes. In *UAI*, volume 4, pages 162–169, 2004.

[3] Lihong Li, Thomas J Walsh, and Michael L Littman. Towards a unified theory of state abstraction for mdps. In *AI&M*, 2006.

[4] Jinhua Song, Yang Gao, Hao Wang, and Bo An. Measuring the distance between finite markov decision processes. In *Proceedings of the 2016 international conference on autonomous agents & multiagent systems*, pages 468–476, 2016.