Bibliographical survey for MDP with imprecise transition function

Peter Kogan

July 18, 2021

MDP is used to model wide variety of real world problems and search for the optimal behaviour. The parameters of MDPs (reward and transition functions) are usually learned from data or estimated based on experts opinion. Typically we aren't not sure about the exact reward or/and transition measures. This type of uncertainty is called external variation (as opposite to the randomness of the transition given action that is called internal variation).

The exact estimations are usually differ from real world values (if there are definite values at all). The optimal policy we found may perform much worse than it have done on the model. The cases when the policies are precomputed using exact dynamic programming with the estimated transition probabilities, but the system evolves according to different, true transition probabilities are studied by Mannor et al. (2007) Mastin and Jaillet (2012).

Furthermore, optimizing for the estimated parameter realization may risk unacceptable performance under other less likely parameter realizations. The question is: given a small set of data or controversial experts' opinions regarding the model, how does decision maker find policy? The world "optimal" is omitted here on purpose. There are several ways to define "good" decision given imprecise probabilities, Troffaes (2007) mentions admissibility, maximal expected utility, maximality, E-admissibility, Γ-maximax, Γ-maximin.

The most widely studied approach is so called robust MDP and it's related to Γ-maximin reward (or minimax cost). In this context, under the assumption that parameters lay in a given uncertainty set, one considers a dynamic game against nature as equivalent to choosing the best strategy for the worst-case scenario Nilim and El Ghaoui (2003) Wiesemann et al. (2013). There are 3 main formulations of the robust MDP Mannor and Xu (2019) : coupled , uncoupled uncertainty and distributional robustness. The first two are related to independence of uncertainty in reward and transition function across states and decision epochs. The case of uncoupled uncertainty is the most intensively studied. In this formulation researhers represent the uncertainty by ambiguity sets that are constructed from data. While the general case is intractable, many algorithms were proposed for so called rectangular ambiguity sets. Their studies dates back to seventies Satia and Lave Jr (1973); Givan et al. (2000); White III and Eldeib (1994); Bagnell et al. (2001). Different types of ambiguity sets as polytope and ellipsoid also has efficient solutions. The downside of this paradigm is its difficulty in incorporating probabilistic information of unknown parameters into the model Mannor and Xu (2019). Another useful sub-case is bounded-parameter MDP (BMDP) Givan et al. (2000). BMDP is a set of exact MDPs specified by giving upper and lower bounds on transition probabilities and rewards.

On the other hand , some work was done for dealing with dependence of uncertainties Mannor et al. (2012). The settings are similar, but the external variation may be dependent across the cases. Here only the reward case turned out to be tractable, and the coupled external variation in transition function is NP-hard. The attempts to address both reward and transition uncertainty with dependence across the states was done by Ahmed et al. (2017) via sampling algorithms.

In distributionally robust MDPs Xu and Mannor (2010), the uncertain parameters are modeled as random variables following an unknown distribution , where it is assumed to belong to an ambiguity set of distributions constructed from historical data. Here the focus is also on the worst case scenario. Note that external variation could be addressed by encoding the unknown model parameters into the

states of a partially observable MDP (POMDP). However, the optimization of POMDPs becomes challenging even for small state spaces Ghavamzadeh et al. (2016).
Alternative performance criteria have been suggested to address external variation, such as the worst-case expected utility and regret measures Xu and Mannor (2006) Delgado et al. (2011).

Additional thread of research presents a percentile criteria that is conceptually natural and representative of the trade-off between optimistic and pessimistic point of views on the question. Percentile measures are viewed as softer notions of robustness where the goal is to maximize the value achieved for a fixed confidence probability Chen and Bowling (2012) Delage and Mannor (2010). In fact, percentile optimization with general transition uncertainty is NP-hard Delage (2009). But assuming independent Dirichlet distribution over external variation and given enough observations we can find an approximate policy by complicated optimization Delage and Mannor (2010). Percentile criteria can be replaced by different objective as VaR and CVaR from Financial Economics Rockafellar and Uryasev (2002) .

References

Ahmed, A., Varakantham, P., Lowalekar, M., Adulyasak, Y., and Jaillet, P. (2017). Sampling based approaches for minimizing regret in uncertain markov decision processes (mdps). *Journal of Artificial Intelligence Research*, 59:229–264. 1

Bagnell, J. A., Ng, A. Y., and Schneider, J. G. (2001). Solving uncertain markov decision processes. 1

Chen, K. and Bowling, M. (2012). Tractable objectives for robust policy optimization. *Advances in Neural Information Processing Systems*, 25:2069–2077. 2

Delage, E. and Mannor, S. (2010). Percentile optimization for markov decision processes with parameter uncertainty. *Operations research*, 58(1):203–213. 2

Delage, E. H. (2009). *Distributionally robust optimization in context of data-driven problems*. Stanford University. 2

Delgado, K. V., Sanner, S., and De Barros, L. N. (2011). Efficient solutions to factored mdps with imprecise transition probabilities. *Artificial Intelligence*, 175(9-10):1498–1527. 2

Ghavamzadeh, M., Mannor, S., Pineau, J., and Tamar, A. (2016). Bayesian reinforcement learning: A survey. *arXiv preprint arXiv:1609.04436*. 2

Givan, R., Leach, S., and Dean, T. (2000). Bounded-parameter markov decision processes. *Artificial Intelligence*, 122(1-2):71–109. 1

Mannor, S., Mebel, O., and Xu, H. (2012). Lightning does not strike twice: Robust mdps with coupled uncertainty. *arXiv preprint arXiv:1206.4643*. 1

Mannor, S., Simester, D., Sun, P., and Tsitsiklis, J. N. (2007). Bias and variance approximation in value function estimates. *Management Science*, 53(2):308–322. 1

Mannor, S. and Xu, H. (2019). Data-driven methods for markov decision problems with parameter uncertainty. In *Operations Research & Management Science in the Age of Analytics*, pages 101–129. INFORMS. 1

Mastin, A. and Jaillet, P. (2012). Loss bounds for uncertain transition probabilities in markov decision processes. In *2012 IEEE 51st IEEE Conference on Decision and Control (CDC)*, pages 6708–6715. IEEE. 1

Nilim, A. and El Ghaoui, L. (2003). Robustness in markov decision problems with uncertain transition matrices. In *NIPS*, pages 839–846. Citeseer. 1

Rockafellar, R. T. and Uryasev, S. (2002). Conditional value-at-risk for general loss distributions. *Journal of banking & finance*, 26(7):1443–1471. 2

Satia, J. K. and Lave Jr, R. E. (1973). Markovian decision processes with uncertain transition probabilities. *Operations Research*, 21(3):728–740. 1

Troffaes, M. C. (2007). Decision making under uncertainty using imprecise probabilities. *International journal of approximate reasoning*, 45(1):17–29. 1

White III, C. C. and Eldeib, H. K. (1994). Markov decision processes with imprecise transition probabilities. *Operations Research*, 42(4):739–749. 1

Wiesemann, W., Kuhn, D., and Rustem, B. (2013). Robust markov decision processes. *Mathematics of Operations Research*, 38(1):153–183. 1

Xu, H. and Mannor, S. (2006). The robustness-performance tradeoff in markov decision processes. *Advances in Neural Information Processing Systems*, 19:1537–1544. 2

Xu, H. and Mannor, S. (2010). Distributionally robust markov decision processes. In *NIPS*, pages 2505–2513. 1