Plan for the Worst with Advice: Advice-augmented Robust Markov Decision Processes

Tinashe Handina **Kishan Panaganti** **Eric Mazumdar** **Adam Wierman**

Department of Computing + Mathematical Sciences
California Institute of Technology, Pasadena, CA 91125

Abstract

We consider the integration of advice into Robust Markov Decision Processes (RMDPs). While the RMDP formulation aids in modeling ambiguity with respect to transition dynamics, it is overly conservative due to its focus on worst-case instances. To move beyond the worst-case framework, we propose an advice-augmented setting in which the decision maker has access to advice in the form of a predicted transition kernel they seek to leverage to obtain better guarantees. The decision maker in this setting cares about finding a policy that performs well for both the worst case and advice transition dynamics. Thus, we define *robustness* and *consistency* as metrics the decision maker optimizes and propose a family of optimization problems whose solutions are Pareto-optimal with respect to robustness and consistency. Under standard assumptions on the ambiguity set, the optimal solutions are deterministic, Markovian, and stationary. Given a set of Pareto-optimal policies, we then provide a policy selection algorithm that achieves max-min optimality across robustness and consistency.

1 Introduction

Building upon the classical MDP formulation, Robust Markov Decision Processes (RMDPs) [1] have emerged as a principled approach to account for ambiguity with respect to the underlying transition probabilities. In the general MDP setting, the next state the agent finds themselves in is determined by a transition probability function. In many real-world settings this is, however, difficult to know precisely. The RMDP formulation thus considers sets of transition kernels and finds policies that are optimal with respect to the worst case transition kernel. The RMDP formulation is conservative in its approach. In particular, considering the worst-case transition kernel, while helpful in providing guarantees, often leads to suboptimal performance when the worst case is not realized.

The proliferation of machine learning combined with the advancement of domain expertise has resulted in the development of advice-augmented algorithms in a wide range of fields. In this model, a decision maker is given (untrusted) advice for whatever problem instance they are trying to solve. The advice that the agent is presented with, importantly, does not have any guarantee of correctness. The decision maker would thus like to leverage the advice in a manner such that, if the advice indeed is correct, they reap as much benefit from the advice as possible and, if the advice is incorrect, they retain some guarantees obtained from classical advice-agnostic algorithms. To this end, *robustness* and *consistency* [2] have emerged as metrics to understand the performance of advice-augmented algorithms. Robustness is a measure of the advice-augmented algorithm's performance in the worst case whilst consistency measures the algorithm's performance in the cases the advice is correct.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: MLxOR: Mathematical Foundations and Operational Integration of Machine Learning for Uncertainty-Aware Decision-Making.

This paper integrates advice into the RMDP formulation. We assume that the decision maker has advice on the transition kernel, which could be the output of some machine-learned prediction process or domain expertise. We seek to understand if the conservative guarantees obtained by the RMDP formulation can be improved upon through the use of advice in a way that still retains guarantees from the robust formulation if the advice is incorrect. We seek to answer the following question:

Can we design decision-making policies that optimally integrate advice into the Robust Markov Decision Process framework?

Our contributions: We propose an approach to find policies that are *Pareto-optimal* with respect to robustness and consistency. That is, for the policies identified through our approach, no other decision-making policy can dominate the robustness and consistency of this set of policies. Furthermore, we show how these policies are *stationary, Markovian, and deterministic* given assumptions on characterizations of the underlying ambiguity set which defines the RMDP problem for the decision maker. We additionally provide a *max-min formulation* to identify a particular policy that maximizes the minimum of robustness and consistency. Finally, we provide an algorithm for efficiently finding such a policy within the set of Pareto-optimal policies defined by our approach.

2 Preliminaries

We consider a finite-horizon Markov Decision Process (MDP) parametrized by $\{\mathcal{S}, \mathcal{A}, P = \{p_t\}_{t=0}^{T}, R = \{r_t\}_{t=0}^{T}\}$. $\mathcal{S}$ denotes the state space, $\mathcal{A}$ the action space, P the stochastic transition kernel, R the reward functions with T being the time horizon. At each timestep, $t \in \{0, 1, \cdots, T\}$, the decision maker is in some state $s_t \in S$ from which they select $a_t \in \Delta(\mathcal{A})$ denoting a decision on the probability of playing any particular action. The probability transition function $p_t : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ then determines the probability of transitioning into any particular state for the next time step. The decision maker attains a reward determined by the function $r_t : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ which takes as input the current state, the action the decision maker selected, the next state and outputs a real valued number. As a convention, the final rewards at timestep T are only determined by the final state the decision maker finds themselves in (i.e., $r_T : \mathcal{S} \to \mathbb{R}$).

For analysis, it is often helpful to consider the characteristics of a decision process starting from a particular time step. Given a fixed transition kernel P, we define the transition kernel starting from a particular time t i.e., $P_t = \{p_\tau\}_{\tau=t}^{T}$. We note here that, $P_0 = \{p_\tau\}_{\tau=0}^{T} = P$. Let $R_t(\pi, P_t, h_t)$ be a random variable denoting the cumulative reward starting from time step t using a particular policy π given history h_t and transitions determined by P_t i.e., $R_t(\pi, P_t, h_t) = \sum_{\tau=t}^{T-1} r_\tau(s_\tau, d_\tau(h_\tau), s_{\tau+1}) + r_T(s_T)$. We can write the optimization problem the decision maker faces as: $\sup_\pi \mathbb{E}_P^\pi \left[\sum_{t=0}^{T-1} r_t(s_t, d_t(h_t), s_{t+1}) + r_T(s_T) \right] = \sup_\pi \mathbb{E}_P^\pi \left[R_0(\pi, P, h_0) \right].$

Knowing the exact transition probabilities is often not possible in many real-world applications. In order to guarantee performance in instances where the transition probabilities are unknown, the robust formulation of the decision maker's optimization problem seeks to find a policy that performs the best under the worst case instantiation of transition probabilities. In particular, we assume that for the probability transitions, there exists an ambiguity set $\mathcal{P}$ over which the decision maker finds the policy π which satisfies $\sup_\pi \inf_{P \in \mathcal{P}} \mathbb{E}_P^\pi \left[R_0(\pi, P, h_0) \right]$.

As a matter of notation, we take the subset of the ambiguity set restricted to measures at time step t to be $\mathcal{P}_t$. Previous work has shown that the structure of the ambiguity set $\mathcal{P}$ leads itself to different characterizations of the optimal policy. In our work we consider *convex, (s, a)–rectangular* ambiguity sets. Previous results have shown the optimality of deterministic, Markovian policies in this setting [1].

While the robust formulation takes a worst-case optimization view, our work attempts to find a middle ground between worst-case pessimism and best-case optimism. In particular, we consider the case where the decision maker has access to some suggested $\mathrm{P}' \in \mathcal{P}$ which is advice on the true underlying transition kernel. We call this setting the "advice-augmented setting." Advice in this setting is not guaranteed to be correct, and as such, the decision maker would like to balance not being too far off from the optimal in the case that the advice is indeed correct and retaining some notion of robustness

in case the advice is incorrect. As a special case, when the advice is a data-driven model, we recover the traditional trade-off of robustness and specificity settings [3, 4].

3 Advice Augmented Planning for MDPs

In the advice augmented setting, the decision maker seeks to balance worst case guarantees from the robust formulation, with performance on the advice transition kernel. We consider *robustness* and *consistency* as metrics to evaluate these different goals. Robustness for a particular policy π is a measure of that policy's performance under a worst-case transition kernel. Consistency, on the other hand, is a measure of a policy's performance on the advice's stochastic transition kernel.

$$\mathrm{Rob}_{\mathcal{P}}(\pi) := \inf_{P \in \mathcal{P}} \mathbb{E}_P^{\pi}[R_0(\pi, P)]. \quad \mathrm{Con}_{\mathtt{P}'}(\pi) := \mathbb{E}_{\mathtt{P}'}^{\pi}[R_0(\pi, \mathtt{P}')].$$

The decision maker thus attempts to maximize both these objectives given a fixed ambiguity set $\mathcal{P}$ and advice transition kernel $\mathtt{P}'$. It is not guaranteed in general to find a particular policy π, that jointly maximizes both these objectives. We thus restrict ourselves to finding a policy that is on the Pareto frontier of robustness and consistency for a particular ambiguity set $\mathcal{P}$ and advice transition probability $\mathtt{P}'$.

Definition 3.1. *Given an ambiguity set and advice transition kernel $\mathcal{P}$, $\mathtt{P}'$, a policy π is Pareto-optimal with respect to robustness and consistency, if there **does not** exist π' such that $\mathrm{Rob}_{\mathcal{P}}(\pi) \leq \mathrm{Rob}_{\mathcal{P}}(\pi')$ and $\mathrm{Con}_{\mathtt{P}'}(\pi) \leq \mathrm{Con}_{\mathtt{P}'}(\pi')$ with one of the inequalities being strict.*

Definition 3.2. *The Pareto frontier, **PF**, is the set of policies such that for all policies $\pi \in$ **PF**, π is Pareto-optimal.*

3.1 Identifying policies on the Pareto frontier

This subsection is dedicated to developing our advice-augmented RMDP framework. We begin by characterizing a set of policies guaranteed to be on the Pareto frontier. Consider the following optimization problem for $\lambda \in (0, 1)$:

$$G_0^{*,\lambda} = \sup_{\pi \in \Pi} \left\{ \lambda \, \mathrm{Rob}_{\mathcal{P}}(\pi) + (1 - \lambda) \, \mathrm{Con}_{\mathtt{P}'}(\pi) \right\}. \tag{1}$$

In the above formulation, the decision maker, for a particular value of λ, tries to find a policy that jointly solves for robustness and consistency. We also present a version of equation 1 that begins at time step t:

$$G_t^{*,\lambda}(h_t) = \sup_{\pi \in \Pi} \Big\{ \lambda \inf_{P_t \in \mathcal{P}_t} \mathbb{E}_P^{\pi} \left[R_t(\pi, P_t, h_t) \right] + (1 - \lambda) \mathbb{E}_{\mathtt{P}'}^{\pi} \left[R_t(\pi, \mathtt{P}'_t, h_t) \right] \Big\}.$$

We now consider the advice augmented setting. We proceed to show the existence of a stationary, deterministic optimal policy for equation 1 given assumptions on $(s, a)-$rectangularity and convexity of the ambiguity set. We leave the proofs of Theorems and Propositions to the Appendix for clarity.

Theorem 3.3. *Let the ambiguity set $\mathcal{P}$ of transition kernels be $(s, a)-$rectangular and convex. The set of functions $\{G_t^{*,\lambda} : t = 0, 1, \ldots, T\}$ satisfy the following equations:*

$$G_T^{*,\lambda}(h_T) = r_T(s_T).$$

$$G_t^{*,\lambda}(h_t) = \sup_{a \in \mathcal{A}(s_t)} \left\{ \inf_{p_t \in \mathcal{P}(s_t, a)} \left\{ \lambda \cdot \mathbb{E}_p[r_t(s_t, a, s)] \;\; + (1 - \lambda) \cdot \mathbb{E}_{\mathtt{p}'_t}[r_t(s_t, a, s)] + \hat{G}_t^{*,\lambda}(h_t, a, p_t) \right\} \right\}.$$

Proposition 3.4. *Let Π_D be the set of all history dependent deterministic policies. Π_D can then fully characterize the advice augmented value function $\{G_t^{*,\lambda} : t = 0, 1, \ldots, T\}$ in that for all t we have: $G_t^{*,\lambda}(h_t) = \sup_{\pi \in \Pi_D} \{G_t^{\pi}(h_t)\}$. Furthermore, consider the optimal policy π^*, for the advice augmented setting, this policy is Markovian*

An important consequence of Theorem 3.3 and Proposition 3.4 is that not only does there exist a solution to equation 1 that is stationary and deterministic, one can also efficiently identify this solution in polynomial time via dynamic programming. This therefore neatly presents an advice-augmented algorithm which incorporates advice into a solution for Robust Markov Decision Processes.

We now move on to show that optimal policies for equation 1 indeed are Pareto optimal with respect to robustness and consistency, which is the best we can hope for. In general, it is known that maximizing the sum of utilities results in points that are Pareto optimal with respect to the summed utility functions. We instantiate this result in a manner that allows us to apply it with respect to consistency and robustness.

Proposition 3.5. *Consider any two functions f_1, f_2 with attainable suprema. The set $S_\lambda = \{x \mid x \in \arg\max_x(\lambda f_1(x) + (1-\lambda) f_2(x))\}$ has elements that lie on the Pareto frontier with respect to f_1, f_2 for $\lambda \in (0,1)$.*

The result above tells us that policies that solve equation 1 are on the Pareto frontier. Concretely, we get this result:

Theorem 3.6. *Consider π^* which is an optimal policy for equation 1, we have that π^* is Pareto optimal with respect to robustness and consistency for $\lambda \in (0,1)$.*

Theorem 3.6 follows from applying Proposition 3.5 onto the optimization problem equation 1.

3.2 Balancing robustness and consistency on the Pareto frontier

In the previous subsection, we found a set of policies that lies on the Pareto frontier which means that the set of policies we identified has a tight trade-off between robustness and consistency. While this is good, the framework above does not provide a means through which one could *select* a point on the Pareto frontier. A natural choice of a policy on the Pareto frontier is one that balances both robustness and consistency. To that end, we consider the following objective:

$$\arg\max_{\pi \in S_\pi} \min\{\operatorname{Rob}_{\mathcal{P}}(\pi), \operatorname{Con}_{\mathtt{P}'}(\pi)\} \tag{2}$$
$$\text{where } S_\pi = \{\pi \mid \pi = \arg\max_{\pi} \lambda \operatorname{Rob}_P(\pi) + (1-\lambda)\operatorname{Con}_{\mathtt{P}'}(\pi) \text{ for some } \lambda \in (0,1)\}.$$

S_π is the set of policies π that are optimal with respect to equation 1 for different values of λ. We know that this set only contains policies that are Pareto optimal with respect to consistency and robustness. We therefore seek to select a Pareto optimal policy $\hat{\pi} \in S_\pi$ which is optimal with respect to equation 2.

Since finding an optimal solution for equation 2 essentially is optimizing over a particular set of π's which themselves are solutions to an optimization problem that depends on λ, we can view finding a solution for equation 2 as a problem of finding an optimal value of λ over the set (0,1). Importantly we note that there is a surjection between the set (0,1) and S_π as for each $\lambda \in (0,1)$ we can find a particular policy π^λ that is the optimal policy for the equation equation 1 for that fixed value of λ. We can thus rewrite the optimization problem as: $\hat{\pi} := \arg\max_{\lambda \in (0,1)} \min\{\operatorname{Rob}_{\mathcal{P}}(\pi^\lambda), \operatorname{Con}_{\mathtt{P}'}(\pi^\lambda)\}$.

Ultimately we show that binary search over the set of λ's will suffice in determining the value of λ and, the policy π^λ that optimizes equation 2. First, we show that the functions $\operatorname{Rob}_P(\pi^\lambda)$, $\operatorname{Con}_{\mathtt{P}'}(\pi^\lambda)$ over the set S_π can be nicely characterized as increasing and decreasing with respect to λ. It is this characterization that then allows us to leverage a binary-search-inspired algorithm to identify the optimal policy with respect to equation 2.

Theorem 3.7. *Let $\pi^\lambda \in S_\pi$ for all $\lambda \in (0,1)$. $\operatorname{Rob}_P(\pi^\lambda)$, $\operatorname{Con}_{\mathtt{P}'}(\pi^\lambda)$ for fixed $\mathcal{P}, \mathtt{P}'$ and defined over the domain of π^λ are increasing and decreasing functions in λ respectively.*

Given that we have now established how $\operatorname{Rob}_P(\pi^\lambda)$ and $\operatorname{Con}_{\mathtt{P}'}(\pi^\lambda)$ are increasing and decreasing functions with respect to λ when defined over the set S_π, we can now conclude and devise a mechanism by which we effectively identify which policy in S_π is the optimizes equation 2.

Theorem 3.8. *Assume that the ambiguity set $\mathcal{P}$ is (s,a)-rectangular and convex. There is an algorithm such that for any fixed $\mathcal{P}$, $\mathtt{P}'$, it can identify the policy $\pi^\lambda \in S_\pi$ which optimizes $\arg\max_{\pi \in S_\pi} \min\{\operatorname{Rob}_{\mathcal{P}}(\pi), \operatorname{Con}_{\mathtt{P}'}(\pi)\}$ with $\mathcal{O}(\log(|S_\pi|))$ policy evaluations.*

4 Conclusion

In this work we introduce advice-augmented Robust Markov Decision Processes. In our setting the decision maker has access to advice on the true transition kernel and they seek to leverage this

advice. To that end, we consider robustness and consistency as metrics that a decision maker is seeking to jointly optimize. To that end, Pareto-optimality is our desired property in this optimization landscape given the two objectives. We show how under standard assumptions on the ambiguity set, Pareto-optimal solutions are deterministic, Markovian, and stationary. We show how to identify Pareto-optimal policies, and furthermore, given a set ofPareto-optimal policies, we provide a policy selection algorithm that achieves max-min optimality across robustness and consistency

References

[1] Garud N. Iyengar. Robust dynamic programming. *Mathematics of Operations Research*, 30(2):257–280, 2005.

[2] Thodoris Lykouris and Sergei Vassilvtiskii. Competitive caching with machine learned advice. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3296–3305, 10–15 Jul 2018.

[3] Shixiong Wang, Haowei Wang, and Jean Honorio. Learning against distributional uncertainty: On the trade-off between robustness and specificity. *arXiv preprint arXiv:2301.13565*, 2023.

[4] Man Yiu Tsang and Karmel S Shehadeh. On the trade-off between distributional belief and ambiguity: Conservatism, finite-sample guarantees, and asymptotic properties. *arXiv preprint arXiv:2410.19234*, 2024.

[5] Arnab Nilim and Laurent El Ghaoui. Robust control of Markov decision processes with uncertain transition matrices. *Operations Research*, 53(5):780–798, 2005.

[6] Huan Xu and Shie Mannor. Distributionally robust Markov decision processes. In *Advances in Neural Information Processing Systems*, pages 2505–2513, 2010.

[7] Shie Mannor, Ofir Mebel, and Huan Xu. Robust MDPs with k-rectangular uncertainty. *Mathematics of Operations Research*, 41(4):1484–1509, 2016.

[8] Wolfram Wiesemann, Daniel Kuhn, and Berç Rustem. Robust Markov decision processes. *Mathematics of Operations Research*, 38(1):153–183, 2013.

[9] Kishan Panaganti and Dileep Kalathil. Robust reinforcement learning using least squares policy iteration with provable performance guarantees. In *International Conference on Machine Learning*, pages 511–520. PMLR, 2021.

[10] Zhengqing Zhou, Zhengyuan Zhou, Qinxun Bai, Linhai Qiu, Jose Blanchet, and Peter Glynn. Finite-sample regret bound for distributionally robust offline tabular reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pages 3331–3339. PMLR, 2021.

[11] Kishan Panaganti and Dileep Kalathil. Sample complexity of robust reinforcement learning with a generative model. In *International Conference on Artificial Intelligence and Statistics*, pages 9582–9602. PMLR, 2022.

[12] Pengqian Yu and Huan Xu. Distributionally robust counterpart in Markov decision processes. *IEEE Transactions on Automatic Control*, 61(9):2538–2543, 2015.

[13] Reazul Hasan Russel and Marek Petrik. Beyond confidence regions: Tight bayesian ambiguity sets for robust mdps. *Advances in Neural Information Processing Systems*, 2019.

[14] Peter J Huber. Robust estimation of a location parameter. In *Breakthroughs in statistics: Methodology and distribution*, pages 492–518. Springer, 1992.

[15] Guanya Shi, Kamyar Azizzadenesheli, Michael O'Connell, Soon-Jo Chung, and Yisong Yue. Meta-adaptive nonlinear control: Theory and algorithms. *Advances in Neural Information Processing Systems*, 34:1–15, 2021.

[16] Rohan Sinha, James Harrison, Spencer M Richards, and Marco Pavone. Adaptive robust model predictive control via uncertainty cancellation. *IEEE Transactions on Automatic Control*, 67(3):1386–1401, 2021.

[17] Francisco Garcia and Philip S Thomas. A meta-mdp approach to exploration for lifelong reinforcement learning. *Advances in Neural Information Processing Systems*, 32, 2019.

[18] Anurag Ajay, Abhishek Gupta, Dibya Ghosh, Sergey Levine, and Pulkit Agrawal. Distributionally adaptive meta reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 35, pages 22591–22604, 2022.

[19] Annie Xie, Shagun Sodhani, Chelsea Finn, Joelle Pineau, and Amy Zhang. Robust policy learning over multiple uncertainty sets. In *International Conference on Machine Learning*, pages 24414–24429, 2022.

[20] Jacob Beck, Risto Vuorio, Evan Zheran Liu, Zheng Xiong, Luisa Zintgraf, Chelsea Finn, and Shimon Whiteson. A survey of meta-reinforcement learning. *arXiv preprint arXiv:2301.08028*, 2023.

[21] Mohammad Mahdian, Hamid Nazerzadeh, and Amin Saberi. Online optimization with uncertain information. *ACM Trans. Algorithms*, 8(1), January 2012.

[22] Carl Elkin and Sims Witherspoon. Machine learning can boost the value of wind energy. *DeepMind Blog*, Dec 2019.

[23] Russell Lee, Jessica Maghakian, Mohammad Hajiesmaili, Jian Li, Ramesh Sitaraman, and Zhenhua Liu. Online peak-aware energy scheduling with untrusted advice. 1(1):59–77, January 2022.

[24] Tongxin Li, Tinashe Handina, Shaolei Ren, and Adam Wierman. Safe exploitative play with untrusted type beliefs, 2024.

[25] Thomas Lavastida, Benjamin Moseley, R. Ravi, and Chenyang Xu. Learnable and Instance-Robust Predictions for Online Matching, Flows and Load Balancing. In *29th Annual European Symposium on Algorithms (ESA 2021)*, volume 204 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 59:1–59:17, 2021.

[26] Daan Rutten, Nicolas Christianson, Debankur Mukherjee, and Adam Wierman. Smoothed online optimization with unreliable predictions. *Proc. ACM Meas. Anal. Comput. Syst.*, 7(1), March 2023.

[27] Andrew Y. Ng, Daishi Harada, and Stuart J. Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *Proceedings of the Sixteenth International Conference on Machine Learning*, ICML '99, page 278–287, San Francisco, CA, USA, 1999. Morgan Kaufmann Publishers Inc.

[28] Eric Wiewiora, Garrison Cottrell, and Charles Elkan. Principled methods for advising reinforcement learning agents. In *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*, ICML'03, page 792–799, 2003.

[29] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[30] James MacGlashan, Mark K. Ho, Robert Loftin, Bei Peng, Guan Wang, David L. Roberts, Matthew E. Taylor, and Michael L. Littman. Interactive learning from policy-dependent human feedback. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 2285–2294, 06–11 Aug 2017.

[31] Lin Guan, Mudit Verma, Suna (Sihang) Guo, Ruohan Zhang, and Subbarao Kambhampati. Widening the pipeline in human-guided reinforcement learning with explanation and context-aware data augmentation. In *Advances in Neural Information Processing Systems*, volume 34, pages 21885–21897, 2021.

[32] Tongxin Li, Yiheng Lin, Shaolei Ren, and Adam Wierman. Beyond black-box advice: learning-augmented algorithms for mdps with q-value predictions. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, 2023.

[33] Noah Golowich and Ankur Moitra. Can q-learning be improved with advice? In *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 4548–4619. PMLR, 02–05 Jul 2022.

APPENDIX

A1: Related Work

The RMDP framework introduced by [1, 5] considers a distributionally robust optimization formulation of decision-making to alleviate environmental ambiguities. These initial works demonstrated that optimal deterministic stationary policies exist under (s, a)-rectangular assumptions on the ambiguity sets. Furthermore they established that the robust value function satisfy a robust Bellman recursion.

By now, many follow-up works have extensively studied different structural properties of the RMDP framework. Moving beyond traditional rectangular uncertainty sets, [6, 7] introduced nested uncertainty sets with different robust budget levels, accounting for different time-dependent real-world scenarios for parameter uncertainty. [8] further develops a unifying framework for distributionally robust optimization with standardized ambiguity sets, providing tractability conditions for solving such problems. Further, there has been recent progress [9, 10, 11] in data-driven learning settings. More aligned with our goals, there has also been work aimed at overcoming the conservative nature of the traditional robust formulation in RMDPs. [12] incorporates multi-modal distributions with mean and variance information in the uncertainty set constructions. [13] proposes a new paradigm for robustness guarantees by constructing Bayesian ambiguity sets using inference rather than relying on confidence regions. [14, 3, 4] consider the trade-off between traditional robust and data-driven model specificity planning strategies.

More traditional meta-adaptive controllers [15, 16] learn shared non-linear dynamic representations that enable robust adaptation across environments using Lyapunov stability. This paradigm aligns with adaptive meta-MDP methods [17, 18] that learn meta-representations by task exploration or distributional optimization for adaptation. The integration of meta-representations with robust planning [19, 20] demonstrates zero-shot adaptation to new environments in large-scale problems. These hybrid architectures occupy a sweet spot, leveraging the meta-learning goal of adaptation while retaining robust planning in dynamical environments.

Building off work such as [21] that looked at how to design advice-augmented algorithms for problems like online ad allocation and load balancing, advice-augmented algorithms have increased in popularity. Their salience has been fueled by the development and incorporation of machine learning into practical problems for example, wind farm optimization [22]. Advice-augmented algorithms have now been developed for a wide range of settings such as: competitive caching [2], energy scheduling [23], games [24], network flow allocation [25], and smoothed online optimization [26]. In these settings, algorithms are designed to leverage advice which is assumed to come from a black box. Little to no assumptions are made on the correctness of the advice and the goal is usually to develop algorithms that achieve an optimal trade off between robustness and consistency.

There has also been work that looks to understand how to incorporate advice or additional information into decision processes. One particular approach is 'reward shaping' wherein external guidance is provided in the form of additional rewards to guide the decision maker [27], [28]. Feedback has also been incorporated into the general MDP setting from a variety of sources such as human experts [29], [30] and visual explanations [31]. Furthermore, advice has been incorporated in the form of Q-value predictions for general MDPs with robustness and consistency tradeoff results proved for this setting [32], [33].

In this work, motivated by the less conservative goals of the meta-adaptive setting and advice-augmented algorithms, we introduce an advice-augmented RMDP framework for the first time. In this setting, advice is in the form of a transition kernel to aide in planning for decision-making systems.

A2: Missing Proofs

For ease of future exposition, we introduce the following notation:

$$
\begin{aligned}
&V_t^{\pi}(h_t) = \inf_{P_t \in \mathcal{P}_t} \mathbb{E}_{P_t}^{\pi}\left[R_t(\pi, P_t, h_t)\right].\\
&J_t^{\pi}(h_t) = \mathbb{E}_{\mathtt{P}_t'}^{\pi}\left[R_t(\pi, \mathtt{P}_t', h_t)\right].\\
&\hat{G}_t^{\pi,\lambda}(h_t, d_t(h_t), p_t) = \lambda \cdot \mathbb{E}_{p_t}^{\pi}[V_{t+1}^{\pi}(h_t, d_t(h_t), s)] + (1-\lambda)\cdot \mathbb{E}_{\mathtt{p}_t'}^{\pi}[J_{t+1}^{\pi}(h_T, d_t(h_t), s)].\\
&\hat{G}_t^{*,\lambda}(h_t, d_t(h_t), p_t) = \sup_{\pi}\left\{\lambda \cdot \mathbb{E}_{p_t}^{\pi}[V_{t+1}^{\pi}(h_t, d_t(h_t), s)] + (1-\lambda)\cdot \mathbb{E}_{\mathtt{p}_t'}^{\pi}[J_{t+1}^{\pi}(h_t, d_t(h_t), s)]\right\}.
\end{aligned}
$$

Theorem 4.1 (Restatement of Theorem 3.3)**.** *Let the ambiguity set $\mathcal{P}$ of transition kernels be $(s,a)-$rectangular and convex. The set of functions $\{G_t^{*,\lambda} : t = 0, 1, \ldots, T\}$ satisfy the following equations:*

$$
\begin{aligned}
&G_T^{*,\lambda}(h_T) = r_T(s_T).\\
&G_t^{*,\lambda}(h_t) = \sup_{a \in \mathcal{A}(s_t)}\left\{\inf_{p_t \in \mathcal{P}(s_t,a)}\{\lambda \cdot \mathbb{E}_p[r_t(s_t,a,s)]\right.\\
&\left.+(1-\lambda)\cdot \mathbb{E}_{\mathtt{p}_t'}[r_t(s_t,a,s)] + \hat{G}_t^{*,\lambda}(h_t, a, p_t)\right\}\Big\}.
\end{aligned}
$$

Proof. We note that the first equation follows directly from realizing that $\lambda \cdot r_T(s_T) + (1-\lambda)\cdot r_T(s_T) = r_T(s_T)$. This illustrates the satisfaction of the first equation.

Before proceeding with the rest of the proof we note the following: the conditional measure P_{t+1} does not affect the term $r_t(s_t, d_t(h_t), s_{t+1})$; by the $(s,a)-$rectangularity assumption we have that $P_t = (p_t, P_{t+1})$ and that $\mathtt{P}_t' = (\mathtt{p}_t', \mathtt{P}_{t+1}')$. With this notation realize that:

$$
\begin{aligned}
&G_t^{*,\lambda}(h_t)\\
&= \sup_{\pi \in \Pi}\left\{\inf_{P_t \in \{\mathcal{P}_\tau\}_{\tau=t}^{T}} \lambda \mathbb{E}_{P_t}^{\pi}\left[R_t(\pi, P_t, h_t)\right]\right.\\
&\quad\left.+(1-\lambda)\mathbb{E}_{\mathtt{P}_t'}^{\pi}\left[R_t(\pi, \mathtt{P}_t', h_t)\right]\right\}.\\
&= \sup_{\pi \in \Pi}\left\{\inf_{P_t \in \{\mathcal{P}_\tau\}_{\tau=t}^{T}} \mathbb{E}_{(p_t, P_{t+1})}^{\pi}\left[\lambda\left(R_t(\pi, P_t, h_t)\right)\right.\right.\\
&\quad\left.\left.+(1-\lambda)\mathbb{E}_{(\mathtt{p}_t', \mathtt{P}_{t+1}')}^{\pi}\left[R_t(\pi, \mathtt{P}_t', h_t)\right]\right]\right\}.
\end{aligned}
$$

At this point, we consider this equation considering the decision taken at time t as well as the transition distribution function at that time step $p_t(\cdot)$. We denote by $h_{t+1}^{\pi,p_t} = (h_t, d_t(h_t), p_t(s_t, d_t(h_t)))$. This, in essence, is a random version of h_{t+1} given h_t, a probability transition function p_t and policy d_t.

With this we can see that:

$$
\begin{aligned}
&= \sup_{\pi\in\Pi}\Big\{ \inf_{p_t\in\mathcal{P}_t} \mathbb{E}^{\pi}_{p_t}\left[\lambda\cdot r_t(s_t,d_t(h_t),s)+\right.\\
&\quad (1-\lambda)\cdot\mathbb{E}^{\pi}_{\mathtt{p}'_t}\left[r_t(s_t,d_t(h_t),s)\right]\\
&\quad +\lambda \inf_{P_{t+1}\in\{\mathcal{P}_\tau\}_{\tau=t+1}^{T}} \mathbb{E}_{P_{t+1}}\left[R_{t+1}(\pi,P_{t+1},h_{t+1}^{\pi,p_t})\right.\\
&\quad \left.\left.+(1-\lambda)\mathbb{E}^{\pi}_{\mathtt{p}'_t}\left[\mathbb{E}^{\pi}_{\mathtt{P}'_{t+1}}\left[R_{t+1}(\pi,\mathtt{P}'_{t+1},h_t^{\pi,\mathtt{p}'_t})\right]\right]\right]\right]\Big\}\\
&= \sup_{\pi\in\Pi}\Big\{\lambda \inf_{p_t\in\mathcal{P}_t} \mathbb{E}^{\pi}_{p_t}\left[r_t(s_t,d_t(h_t),s)\right]\\
&\quad +(1-\lambda)\cdot\mathbb{E}^{\pi}_{\mathtt{p}'_t}\left[r_t(s_t,d_t(h_t),s)\right]\\
&\quad +\lambda\cdot\mathbb{E}^{\pi}_{p_t}\left[\inf_{P_{t+1}\in\{\mathcal{P}_\tau\}_{\tau=t+1}^{T}} \mathbb{E}^{\pi}_{P_{t+1}}\left[R_{t+1}(\pi,P_{t+1},h_t^{\pi,p_t})\right]\right]\\
&\quad +(1-\lambda)\cdot\mathbb{E}^{\pi}_{\mathtt{p}'_t}\left[\mathbb{E}^{\pi}_{\mathtt{P}'_{t+1}}\left[R_{t+1}(\pi,\mathtt{P}'_{t+1},h_t^{\pi,\mathtt{p}'_t})\right]\right]\Big\}\\
&= \sup_{\pi\in\Pi}\Big\{\lambda \inf_{p_t\in\mathcal{P}_t} \mathbb{E}^{\pi}_{p_t}\left[r_t(s_t,d_t(h_t),s)\right]\\
&\quad +(1-\lambda)\cdot\mathbb{E}^{\pi}_{\mathtt{p}'_t}\left[r_t(s_t,d_t(h_t),s)\right]\\
&\quad +\hat{G}_t^{\pi,\lambda}(h_t,d_t(h_t),p_t)\Big\}.
\end{aligned}
$$

Consider a fixed $P\in\mathcal{P}$ and a particular policy π, we have that $\hat{G}^{\pi,\lambda}(h_t,d_t(h_t),p_t)\leq \hat{G}_T^{*,\lambda}(h_t,d_t(h_t),p_t)$. This then implies that:

$$
\begin{aligned}
&G_t^{*,\lambda}(h_t) && (3)\\
&\leq \sup_{\pi\in\Pi}\Big\{ \inf_{p_t\in\mathcal{P}_T} \mathbb{E}^{\pi}_{p_t}\left[\lambda\cdot r_t(s_t,d_t(h_t),s)\right]\\
&\quad +(1-\lambda)\cdot\mathbb{E}^{\pi}_{\mathtt{p}'_t}\left[r_t(s_t,d_t(h_t),s)\right]\\
&\quad +\hat{G}_t^{*,\lambda}(h_t,d_t(h_t),p_t)\Big\}\\
&= \sup_{d_t\in\mathcal{D}_t}\Big\{ \inf_{p_t\in\mathcal{P}_T} \mathbb{E}^{\pi}_{p_t}\left[\lambda\cdot r_t(s_t,d_t(h_t),s)\right] && (4)\\
&\quad +(1-\lambda)\cdot\mathbb{E}^{\pi}_{\mathtt{p}'_t}\left[r_t(s_t,d_t(h_t),s)\right]\\
&\quad +\hat{G}_t^{*,\lambda}(h_t,d_t(h_t),p_t)\Big\}.
\end{aligned}
$$

We get 3 by making the substitution of the upper bound described above. We then get from 3 to 4 by realizing that the only dependency on π is in $d_t(\cdot)$. We denote $\mathcal{D}_t$ to be the set of all history-dependent decision rules at time t.
To complete the proof consider:

$$
\begin{aligned}
&\hat{G}_t^{*,\lambda}(h_t,d_t(h_t),p)\\
&= \sup_{\pi\in\Pi}\left\{\lambda\cdot\mathbb{E}^{\pi}_{p}[V^{\pi}_{t+1}(h_t,d_t(h_t),s)]\right.\\
&\quad \left.+(1-\lambda)\cdot\mathbb{E}^{\pi}_{\mathtt{p}'_t}[J^{\pi}_{t+1}(h_t,d_t(h_t),s)]\right\}
\end{aligned}
$$

It follows that for all $\varepsilon>0$ there exists a policy $\pi^{\varepsilon}\in\Pi$ such that

$$
\hat{G}_t^{\pi^{\varepsilon},\lambda}(h_t,d_t(h_t),p)\geq \hat{G}_t^{*,\lambda}(h_t,d_t(h_t),p)-\varepsilon,\quad \forall\, h_t\in\mathcal{H}_t.
$$

Consider all decision rules $d_t \in \mathcal{D}_t$, $p_t \in \mathcal{P}_t$. We now have:

$$\begin{aligned}
G_t^{*,\lambda}(h_t) &= \sup_{\pi \in \Pi} \Big\{ \inf_{p_t \in \mathcal{P}_t} \mathbb{E}^{\pi}_{p_t} \left[\lambda \cdot r_t(s_t, d_t(h_t), s)\right] \\
&\quad + (1-\lambda) \cdot \mathbb{E}^{\pi}_{\mathtt{p}'_t} \left[r_t(s_t, d_t(h_t), s)\right] \\
&\quad + \hat{G}_t^{\pi,\lambda}(h_t, d_t(h_t), p_t) \Big\} \\
&\geq \sup_{d_t \in \mathcal{D}_t} \Big\{ \inf_{p_t \in \mathcal{P}_t} \mathbb{E}^{\pi}_{p_t} \left[\lambda \cdot r_t(s_t, d_t(h_t), s)\right] \\
&\quad + (1-\lambda) \cdot \mathbb{E}^{\pi}_{\mathtt{p}'_t} \left[r_t(s_t, d_t(h_t), s)\right] \\
&\quad + \hat{G}_t^{\pi^{\varepsilon},\lambda}(h_t, d_t(h_t), p_t) \Big\} \\
&\geq \sup_{d_t \in \mathcal{D}_t} \Big\{ \inf_{p_t \in \mathcal{P}_t} \mathbb{E}^{\pi}_{p_t} \left[\lambda \cdot r_t(s_t, d_t(h_t), s)\right] \\
&\quad + (1-\lambda) \cdot \mathbb{E}^{\pi}_{\mathtt{p}'_t} \left[r_t(s_t, d_t(h_t), s)\right] \\
&\quad + \hat{G}_t^{*,\lambda}(h_t, d_t(h_t), p_t) \Big\} - \varepsilon.
\end{aligned}$$

Since $\varepsilon > 0$ is arbitrary, we then have that:

$$\begin{aligned}
&G_t^{*,\lambda}(h_t) \\
&= \sup_{d_t \in \mathcal{D}_t} \Big\{ \inf_{p_t \in \mathcal{P}_t} \mathbb{E}^{\pi}_{p_t} \left[\lambda \cdot r_t(s_t, d_t(h_t), s)\right] \\
&\quad + (1-\lambda) \cdot \mathbb{E}^{\pi}_{\mathtt{p}'_t} \left[r_t(s_t, d_t(h_t), s)\right] \\
&\quad + \hat{G}_t^{*,\lambda}(h_t, d_t(h_t), p_t) \Big\}.
\end{aligned}$$

By the definition of $\mathcal{P}_t$ and $\mathcal{D}_t$, we note the following:

$$\begin{aligned}
&G_t^{*,\lambda}(h_t) \\
&= \sup_{q \in \Delta(\mathcal{A}(s_t))} \Bigg\{ \sum_{a \in \mathcal{A}(s_t)} q(a) \Bigg(\lambda \inf_{p \in \mathcal{P}^{(s_t,a)}} \sum_{s \in \mathcal{S}} p(s) r_t(s_t, a, s) \\
&\quad + (1-\lambda) \cdot \sum_{s \in \mathcal{S}} p_t^{'s_t,a}(s) r_t(s_t, a, s) + \hat{G}^{*,\lambda}(h_t, a, p) \Bigg) \Bigg\} \\
&= \sup_{a \in \mathcal{A}(s_t)} \Bigg\{ \lambda \inf_{p \in \mathcal{P}^{(s_t,a)}} \sum_{s \in \mathcal{S}} p(s) r_t(s_t, a, s) \\
&\quad + (1-\lambda) \cdot \sum_{s \in \mathcal{S}} p_t^{'s_t,a}(s) r_t(s_t, a, s) + \hat{G}^{*,\lambda}(h_t, a, p) \Bigg\} \\
&= \sup_{a \in \mathcal{A}(s_t)} \Big\{ \inf_{p_t \in \mathcal{P}(s_t,a)} \Big\{ \lambda \cdot \mathbb{E}_{p_t}[r_t(s_t, a, s)] \\
&+ (1-\lambda) \cdot \mathbb{E}_{\mathtt{p}'_t}[r_t(s_t, a, s)] + \hat{G}_t^{*,\lambda}(h_t, a, p_t) \Big\} \Big\}.
\end{aligned}$$

To clarify notation, here $p_t^{'s_t,a} \in \Delta(\mathcal{S})$ is the distribution according to $\mathtt{P}'$ at time t given current state s and action a. □

□

Proposition 4.2 (Restatement of Proposition 3.4)**.** *Let Π_D be the set of all history dependent deterministic policies. Π_D can then fully characterize the advice augmented value function $\{G_t^{*,\lambda} : t = 0, 1, \ldots, T\}$ in that for all t we have:*

$$G_t^{*,\lambda}(h_t) = \sup_{\pi \in \Pi_D} \left\{ G_t^{\pi}(h_t) \right\}.$$

Furthermore, consider the optimal policy π^*, for the advice augmented setting, this policy is Markovian

Proof. The proposition above follows directly from the realizing that Theorem 3.3 specifies a deterministic policy and thus optimizing over the space of deterministic policies gives us an optimal policy. We get the Markovian property by realizing that the satisfaction of the equations in Theorem 3.3 are such that the optimal policy selects the best action only based on s_t. More concretely, consider the following argument using inductive reasoning. For the base case, at time step T, the advice augmented reward is only a function of the final state. For the inductive case assume, that we have an optimal Markovian policy for time step $t+1$ for all possible histories h_t, we argue that at time step t we have a Markovian policy. Note that $\sup_\pi\{\lambda V^\pi_{t+1}(h_{t+1}) + (1-\lambda)J^\pi_{t+1}(h_{t+1})\} = G^{*,\lambda}_{t+1}(h_{t+1}) = G^{*,\lambda}_{t+1}(s_{t+1})$ by the inductive hypothesis. Realize how s_{t+1} depends only on s_t from the history h_t. This therefore means that $\hat{G}^{*,\lambda}_t(h_t, d_t(h_t), p_t) = \hat{G}^{*,\lambda}_t(s_t, d_t(h_t), p_t)$ substituting this into the equations in Theorem 3.3 we can then deduce that $G^{*,\lambda}_t(h_t) = G^{*,\lambda}_t(s_t)$. In other words, the only dependency on h_t is via s_t which implies the optimal policy is Markovian at time t as well. □ □

Proposition 4.3 (Restatement of Proposition 3.5)**.** *Consider any two functions f_1, f_2 with attainable suprema. The following set has elements that lie on the Pareto frontier with respect to f_1, f_2:*

$$S_\lambda = \{x \mid x \in \arg\max_x(\lambda f_1(x) + (1-\lambda)f_2(x))\}$$

for $\lambda \in (0,1)$.

Proof. Assume towards a contradiction that there existed a λ, x such that $x \in S_\lambda$ yet x was not Pareto optimal with respect to f_1 and f_2. This would mean that there would necessarily exist some x' such that $f_1(x') \geq f_1(x)$ and $f_2(x') \geq f_2(x)$ with one of the inequalities being strict. From this we have that $\lambda f_1(x') \geq \lambda f_1(x)$ and $(1-\lambda)f_2(x') \geq (1-\lambda)f_2(x)$ which implies that $\lambda f_1(x') + (1-\lambda)f_2(x') > \lambda f_1(x) + (1-\lambda)f_2(x) \quad \forall\lambda \in (0,1)$. The strictness in the last inequality follows from the necessity of having one strict inequality in the Pareto sub-optimality condition. This would contradict the assumption of x's membership in S_λ. □

Theorem 4.4 (Restatement of Theorem 3.7)**.** *Let $\pi^\lambda \in S_\pi$ for all $\lambda \in (0,1)$. $\mathrm{Rob}_P(\pi^\lambda)$, $\mathrm{Con}_{P'}(\pi^\lambda)$ for fixed $\mathcal{P}, \mathcal{P}'$ and defined over the domain of π^λ are increasing and decreasing functions in λ respectively.*

Proof. We begin by considering λ_1, λ_2 such that $\lambda_1 > \lambda_2$. Assume towards a contradiction that $\mathrm{Rob}_P(\pi^{\lambda_2}) > \mathrm{Rob}_P(\pi^{\lambda_1})$. We have the following cases:

Case 1: $\mathrm{Con}_{P'}(\pi^{\lambda_2}) \geq \mathrm{Con}_{P'}(\pi^{\lambda_1})$.

In this instance we have that $\mathrm{Rob}_P(\pi^{\lambda_2}) > \mathrm{Rob}_P(\pi^{\lambda_1})$ and $\mathrm{Con}_{P'}(\pi^{\lambda_2}) \geq \mathrm{Con}_{P'}(\pi^{\lambda_1})$ which implies that:

$$\begin{aligned}&\lambda\,\mathrm{Rob}_P(\pi^{\lambda_2}) + (1-\lambda)\,\mathrm{Con}_{P'}(\pi^{\lambda_2}) > \\ &\lambda\,\mathrm{Rob}_P(\pi^{\lambda_1}) + (1-\lambda)\,\mathrm{Con}_{P'}(\pi^{\lambda_1}) \quad \forall\lambda \in (0,1).\end{aligned}$$

This contradicts π^{λ_1}'s membership in the set S_π.

Case 2: $\mathrm{Con}_{P'}(\pi^{\lambda_2}) < \mathrm{Con}_{P'}(\pi^{\lambda_1})$**.**

We have by definitions and construction of π^λ's that:

$$\begin{aligned}&\lambda_2\,\mathrm{Rob}_P(\pi^{\lambda_2}) + (1-\lambda_2)\,\mathrm{Con}_{P'}(\pi^{\lambda_2}) \geq \\ &\lambda_2\,\mathrm{Rob}_P(\pi^{\lambda_1}) + (1-\lambda_2)\,\mathrm{Con}_{P'}(\pi^{\lambda_1})\end{aligned}$$

Which by the assumptions that $\lambda_1 > \lambda_2$, $\text{Rob}_P(\pi^{\lambda_2}) > \text{Rob}_P(\pi^{\lambda_1})$ and the case that $\text{Con}_{\mathtt{P}'}(\pi^{\lambda_2}) < \text{Con}_{\mathtt{P}'}(\pi^{\lambda_1})$ we can deduce that:

$$\lambda_1 \text{Rob}_P(\pi^{\lambda_2}) + (1-\lambda_1)\text{Con}_{\mathtt{P}'}(\pi^{\lambda_2}) > \lambda_1 \text{Rob}_P(\pi^{\lambda_1}) + (1-\lambda_1)\text{Con}_{\mathtt{P}'}(\pi^{\lambda_1}).$$

This is essentially because, λ_1 moves weight to the function Rob_P, for which we have $\text{Rob}_P(\pi^{\lambda_2}) > \text{Rob}_P(\pi^{\lambda_1})$. This again contradicts π^{λ_1}'s membership in the set S_π.
With these two cases, we can conclude that Rob_P is increasing over the domain S_π with respect to λ. Similar logic follows through to establish how $\text{Con}_{\mathtt{P}'}$ is a decreasing function over the domain S_π with respect to λ. □

Theorem 4.5 (Restatement of Theorem 3.8)**.** *Assume that the ambiguity set $\mathcal{P}$ is (s,a)-rectangular and convex. There is an algorithm such that for any fixed $\mathcal{P}$, $\mathtt{P}'$, it can identify the policy $\pi^\lambda \in S_\pi$ which optimizes the following*

$$\arg\max_{\pi \in S_\pi} \min\{\text{Rob}_{\mathcal{P}}(\pi), \text{Con}_{\mathtt{P}'}(\pi)\}$$

with $\mathcal{O}(\log(|S_\pi|))$ policy evaluations.

Proof. We begin by making the connection that the optimization problem above amounts to selecting an appropriate value of λ as we are optimizing over the set of policies that can be parametrized by the selection of λ. In particular, the optimization problem above can be rewritten as equation 3.2. Consider the following:

$$\lambda^*_{Rob} = \arg\max_\lambda \{\lambda | \text{Rob}_P(\pi^\lambda) \leq \text{Con}_{\mathtt{P}'}(\pi^\lambda)\}.$$

$$\lambda^*_{Con} = \arg\min_\lambda \{\lambda | \text{Con}_{\mathtt{P}'}(\pi^\lambda) \leq \text{Rob}_P(\pi^\lambda)\}.$$

We note that if $\text{Rob}_P(\pi^\lambda) > \text{Con}_{\mathtt{P}'}(\pi^\lambda)\ \forall \lambda \in (0,1)$ then $\lambda^*_{Rob} = 1$ and similarly if $\text{Con}_{\mathtt{P}'}(\pi^\lambda) > \text{Rob}_P(\pi^\lambda)\ \forall \lambda \in (0,1)$ then $\lambda^*_{Con} = 0$. Realize that the solution to equation 3.2 lies in the set $\{\lambda^*_{Rob}, \lambda^*_{Con}\}$. We get this from Theorem 3.7 and note that we have Rob_P as an increasing function and $\text{Con}_{\mathtt{P}'}$ as a decreasing function of λ. We can simply run two binary search algorithms to find policies π^λ that correspond to $\{\lambda^*_{Rob}, \lambda^*_{Con}\}$ respectively. We note that the set S_π contains a finite, discrete, number of policies by the convexity and $(s,a)-$rectangularity of the ambiguity set $\mathcal{P}$ and hence we can find these policies in $\mathcal{O}(\log(|S_\pi|))$ time. After this, we simply pick the policy that results in a higher value for the optimization problem above. □

A3: Experimental Results

In this section, we move to provide an illustration of the tradeoff between robustness and consistency for a grid world navigation problem. In our experimental setup, we have a single agent navigating a grid world from which they attempt to reach a target given a particular starting point. The agent only has 4 actions available to them: Up, Down, Left, Right. The transitions into the next state depend on their current state, their action, and some stochasticity from the general environment. We think of this stochasticity as the likelihood of being transferred into another arbitrary state despite the agent's actions.

The robust formulation begins with transition dynamics that are such that the action taken by an agent completely determines the next state the agent finds themselves in. If the agent selects an action that leads them out of the grid world, they simply remain in the same square they took that particular action from (i.e., if the agent moves up, they go up unless they are at the edge of the grid world, in which case they just remain in the same square). Let this nominal transition distribution be P. From this nominal distribution, the robust formulation considers a χ^2-ball of radius ρ around each of the distributions specified by the state and action the agent takes from that state. Concretely, we define the ambiguity set $\mathcal{P}$ to be:

$$\mathcal{P} = \{\mathtt{P}' : \forall\ p_t^{'(s,a)} \in \mathtt{P}',\ D_{\chi^2}(p_t^{'(s,a)}, p_t^{(s,a)}) \leq \rho\}.$$

We note that we make use of the symmetric definition of χ^2 distance:

$$D_{\chi^2}(p_t^{'(s,a)}, p_t^{(s,a)}) = \frac{1}{2}\sum_{\bar{s}\in\mathcal{S}} \frac{(p_t^{'(s,a)}(\bar{s}) - p_t^{(s,a)}(\bar{s}))^2}{p_t^{'(s,a)}(\bar{s}) + p_t^{(s,a)}(\bar{s})}.$$

For the advice, we make use of structural constraints in order to better aid with the transition distributions. In particular, we assume advice which is such that, the agent is only perturbed into some states that are adjacent to their current state. In particular, we find a distribution such that the agent only is likely to slip into "obstacle" cells and only if they are adjacent to such cells.

In Fig. 1, we showcase an instance of this grid world problem. In this instance, the agent starts from one corner of the grid world and seeks to make it to the 'Goal' located on the opposite corner of the grid world. In our evaluations, we assume that the agent is equally likely to start from any cell in the grid world.

Numerically, we show how policies that solve equation 1 trade-off for different values of λ. We make use of dynamic programming [1] to identify the optimal policy for equation 1 as per Theorem 3.3. In our plots, we see how there are only a discrete number of policies that characterize the Pareto Frontier as our theory suggests. Finally, our numerics illustrate the monotonicity suggested by Theorem 3.7. In particular, we see how robustness is an increasing function with respect to λ whilst, consistency is a decreasing function. We also note that as ρ increases, the robust value decreases, incentivizing the selection of λ values closer to 0 in the case that the decision-maker wants to obtain any benefit from the predicted transition kernel.

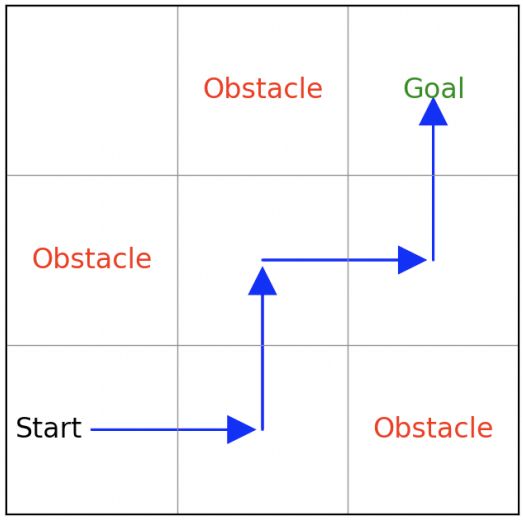

Figure 1: 3X3 Grid world problem instance with optimal trajectory.

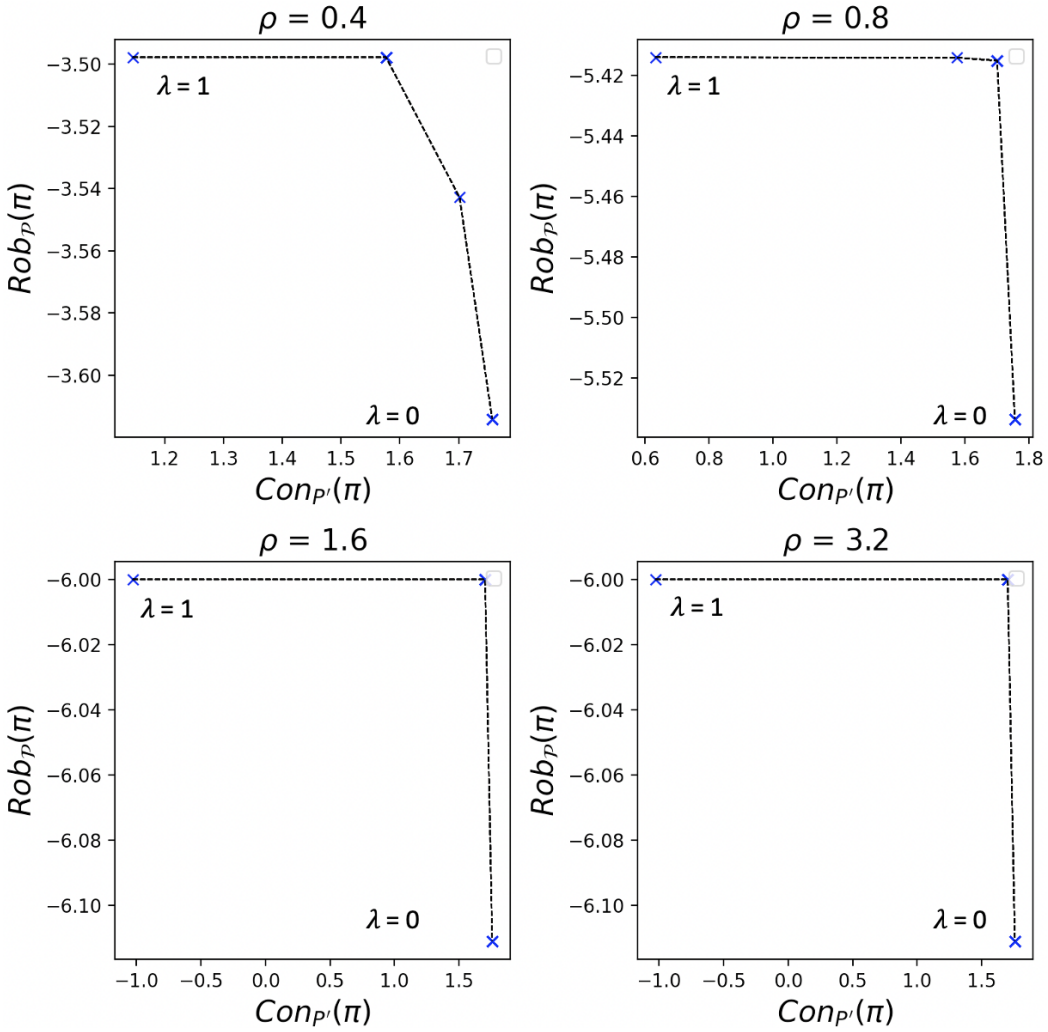

Figure 2: Robustness, consistency trade-off for different values of ρ which specifies the size of the ambiguity set.