2009 American Control Conference
Hyatt Regency Riverfront, St. Louis, MO, USA
June 10-12, 2009

Robust Adaptive Markov Decision Processes in Multi-vehicle Applications

Luca F. Bertuccelli, Brett Bethke, and Jonathan P. How

Aerospace Controls Laboratory
Massachusetts Institute of Technology
{lucab, bbethke, jhow}@mit.edu

Abstract— **This paper presents a new robust and adaptive framework for Markov Decision Processes that accounts for errors in the transition probabilities. Robust policies are typically found off-line, but can be extremely conservative when implemented in the real system. Adaptive policies, on the other hand, are specifically suited for on-line implementation, but may display undesirable transient performance as the model is updated though learning. A new method that exploits the individual strengths of the two approaches is presented in this paper. This robust and adaptive framework protects the adaptation process from exhibiting a worst-case performance during the model updating, and is shown to converge to the true, optimal value function in the limit of a large number of state transition observations. The proposed framework is investigated in simulation and actual flight experiments, and shown to improve transient behavior in the adaptation process and overall mission performance.**

I. Introduction

Many decision processes, such as Markov Decision Processes (MDPs) and Partially Observable MDPs (POMDPs) are modeled as a probabilistic process driven by a known Markov Chain. In practice however, the true parameters of the Markov Chain are frequently unavailable to the modeler, and many researchers have recently addressed the issue of robust performance in these decision systems [1]–[4].

While many authors have studied the problem of MDPs with uncertain transition probabilities [5]–[7], robust counterparts to these MDPs have been obtained only recently. Robust MDP counterparts have been introduced in the work of Bagnell et al [8], Nilim [1] and Iyengar [2]. Bagnell presented a robust value iteration algorithm for solving the robust MDPs. The convergence of robust value iteration was formally proved by Nilim [1] and Iyengar [2]. Both Nilim and Iyengar introduced meaningful uncertainty sets for the transition probabilities that could be efficiently solved by adding an additional, "inner" optimization on the uncertain transition probabilities. One of the methods for finding a robust policy in [1] was to use scenario-based methods, wherein the performance is optimized for different realizations of the transition probabilities. However, it was recently shown that a scenario-based approach may require an extremely large number of realizations to yield a robust policy [4]. This observation motivated the development of a specific scenario selection process using the first two moments of a Bayesian prior to obtain robust policies using much fewer scenarios [4], [21].

Robust methods find robust policies that hedge against errors in the transition probabilities. However, there are many cases when this type of an approach is too conservative. For example, it may be possible to identify the transition probabilities by observing state transitions, and obtain improved estimates, and resolve the optimization to find a less conservative policy. Model-based learning of MDPs is closely related to indirect adaptive control [9] in that the transition probabilities are estimated in real-time using a maximum likelihood estimator. At each time step, certainty equivalence is assumed on the transition probabilities, and a new policy is found with the new model estimate [10]. Jaulmes et al. [11], [12] study this problem in an active estimation context using POMDPs. Marbach [13] considers this problem, when the transition probabilities depend on a parameter vector. Konda and Tsitsiklis [14] consider the problem of slowly-varying Markov Chains in the context of reinforcement learning. Sato [15] considers this problem and shows asymptotic convergence of the probability estimates also in the context of dual control. Kumar [16] also considered the adaptation problem. Ford and Moore [17] consider the problem of estimating the parameters of a non-stationary Hidden Markov Model.

This paper demonstrates the need to account for both robust planning and adaptation in MDPs with uncertainty in their transition probabilities. Just like in control [18] or in task assignment problems [19], adaptation alone is generally not sufficient to ensure reliable operation of the overall control system. This paper shows that robustness is critical to mitigating worst-case performance, particularly during the transient periods of the adaptation.

This paper contributes a new combined robust and adaptive problem formulation for MDPs with errors in the transition probabilities. The key result of this paper shows that *robust and adaptive* MDPs can converge to the truly optimal objective in the limit of a large number of observations. We demonstrate the robust component of this approach by using a Bayesian prior, and finds the robust policy by using scenario-based methods. We then augment the robust approach with an adaptation scheme that is more effective at incorporating new information in the models. The MDP framework is discussed in Section II, the impact of uncertainty is demonstrated in Section III, and then we present the individual components of robustness and adaptation in

978-1-4244-4524-0/09/$25.00 ©2009 AACC

Section IV. The combined robust and adaptive MDP is shown to converge to the true, optimal value function in the limit of a large number of observations. The paper concludes in Section VI with a set of demonstrative numerical simulations and actual flight results on our UAV testbed.

II. Markov Decision Process

A. Problem Formulation

The Markov Decision Process (MDP) framework that we consider in this paper consists of a set of states $i \in S$ of cardinality N, a set of control actions $u \in \mathscr{U}$ of cardinality M with a corresponding policy $\mu : S \rightarrow \mathscr{U}$, a transition model given by $A^u_{ij} = \Pr(i_{k+1}|j_k, u_k)$, and a reward model $g(i,u)$. The time-additive objective function is defined as

$$J_\mu = g_N(i_N) + \sum_{k=0}^{N-1} \phi^k g_k(i_k, u_k) \tag{1}$$

where $0 < \phi \leq 1$ is an appropriate discount factor. The goal is to find an optimal control policy, μ^*, that maximizes an expected objective given some known transition model A^u

$$J^* = \max_\mu \ \mathbf{E}\left[J_\mu(i_0)\right] \tag{2}$$

In an infinite horizon setting ($N \rightarrow \infty$), the solution to Eq. 2 can be found by solving the Bellman Equation

$$J^*(i) = \max_u \left[g(i) + \phi \sum_j A^u_{ij} J^*(j) \right], \quad \forall i \tag{3}$$

The optimal control is found by solving

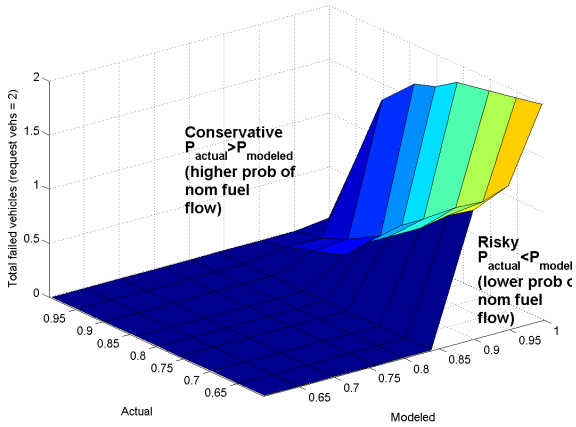
$$u^*(i) \in \arg\max_{u \in \mathscr{U}} \ \mathbf{E}\left[J_\mu(i_0)\right] \qquad \forall i \in S \tag{4}$$

The optimal policy can be found in many different ways using Value Iteration or Policy Iteration, while Linear Programming can be used for moderately sized problems [20].

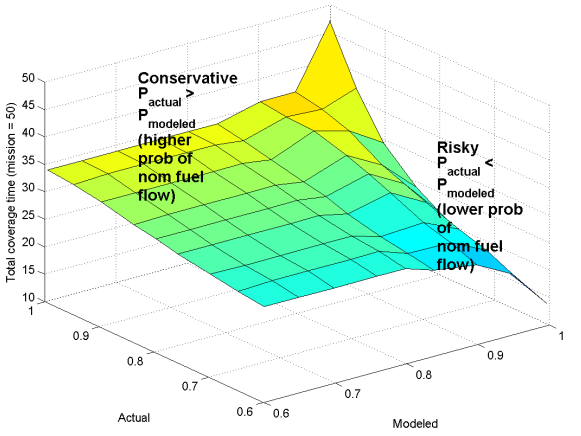
III. Model Uncertainty

It has been shown that the value function can be biased in the presence of small errors in the transition probabilities [3], and that the optimal policy μ^* can be extremely sensitive to small errors in the model parameters. For example, in the context of UAV missions, it has been shown that errors in the state transition matrix, $\tilde{A}^u$, can result in increased UAV crashes when implemented in real systems [21].

An example of this suboptimal performance is reflected in Figure 1, which shows two summary plots for a 2-UAV persistent surveillance mission formulated as an MDP [22] averaged over 100 Monte Carlo simulations. The simulations were performed with modeling errors: the policy was found by using an estimated probability shown on the y-axis ("Modeled"), but implemented on the real system that assumed a nominal probability shown on the x-axis, ("Actual"). Figure 1(a) shows the mean number of failed vehicles in the mission. Note that in the region labeled "Risky", the failure rate is increased significantly such that all the vehicles in the mission are lost due to the modeling error. Figure 1(b) shows the penalty in total coverage time when the transition probability is underestimated (in the area denoted



(a) Total number of failed vehicles



(b) Mean coverage time vs mismatched fuel flow probabilities

Fig. 1. Impact on modeling error on the overall mission effectiveness

as "Risky" in the figure). In this region, the total coverage time decreases from approximately 40 time steps (out of a 50 time step mission) to only 10 time steps. It is of paramount importance to develop precise mathematical descriptions for these errors and use this information to find robust policies. While there are many methods to describe uncertainty sets [1], [2], our approach relies on a Bayesian description of this uncertainty. This choice is primarily motivated by the need to update estimates of these probabilities in real-time in a computationally tractable manner. This approach assumes a prior Dirichlet distribution on each row of the transition matrix, and recursively updates this distribution with observations. The Dirichlet distribution f_D at time k for a row of the N-dimensional transition model is given by $\mathbf{p}_k = [p_1, p_2, \ldots, p_N]^T$ and positive distribution parameters $\alpha(k) = [\alpha_1, \alpha_2, \ldots, \alpha_N]^T$, is defined as

$$\begin{aligned} f_D(\mathbf{p}_k|\alpha(\mathbf{k})) &= K \prod_{i=1}^{N} p_i^{\alpha_i - 1}, \quad \sum_i p_i = 1 \\ &= K \ p_1^{\alpha_1 - 1} p_2^{\alpha_2 - 1} \ldots (1 - \sum_{i=1}^{N-1} p_i)^{\alpha_N - 1} \end{aligned} \tag{5}$$

where K is a normalizing factor that ensures the probability distribution integrates to unity. Each p_i is the i^{th} entry of

the m^{th} row, that is: $p_i = A^u_{m,i}$ and $0 \le p_i \le 1$ and $\sum_i p_i = 1$. The primary reasons for using the Dirichlet distribution is that the mean $\bar{p}_i$ satisfies the requirements of a probability vector $0 \le \bar{p}_i \le 1$ and $\sum_i \bar{p}_i = 1$ by construction. Furthermore, the parameters α_i can be interpreted as "counts", or times that a particular state transition was observed. This enables computationally tractable updates on the distribution based on new observations. The uncertainty set description for the Dirichlet is known as a credibility region, and can be found by Monte Carlo integration.

IV. Adaptation and Robustness

This section discusses individual methods for adapting to changes in the transition probabilities, as well as methods for accounting for robustness in the presence of the transition probability uncertainty.

A. Adaptation

It is well known that the Dirichlet distribution is conjugate to the multinomial distribution, implying a measurement update step that can be expressed in closed form using the previously observed counts $\alpha(k)$. The posterior distribution $f_D(\mathbf{p}_{k+1}|\alpha(k+1))$ is given in terms of the prior $f_D(\mathbf{p}_k|\alpha(k))$ as

$$\begin{aligned} f_D(\mathbf{p}_{k+1}|\alpha(k+1)) &\propto f_D(\mathbf{p}_k|\alpha(k)) f_M(\beta(k)|\mathbf{p}_k) \\ &= \prod_{i=1}^{N} p_i^{\alpha_i - 1} p_i^{\beta_i} = \prod_{i=1}^{N} p_i^{\alpha_i + \beta_i - 1} \end{aligned}$$

where $f_M(\beta(k)|\mathbf{p}_k)$ is a multinomial distribution with hyper-parameters $\beta(k) = [\beta_1, \ldots, \beta_N]$. Each β_i is the total number of transitions observed from state i to a new state i': mathematically $\beta_{i'} = \sum_i \delta_{i,i'}$ and

$$\delta_{i,i'} = \begin{cases} 1 & \text{if transition} \\ 0 & \text{Otherwise} \end{cases}$$

indicates how many times transitions were observed from state i to state i'. For the next derivations, we assume that only a single transition can occur per time step, $\beta_i = \delta_{i,i'}$.

Upon receipt of the observations $\beta(k)$, the parameters $\alpha(k)$ are updated according to

$$\alpha_i(k+1) = \alpha_i(k) + \delta_{i,i'} \tag{6}$$

and the mean can be found by normalizing these parameters $\bar{p}_i = \alpha_i / \alpha_0$.

Our recent work [23] has shown that the mean $\bar{p}_i$ can be equivalently expressed recursively in terms of the previous mean and variance

$$\bar{p}_i = \alpha_i / \alpha_0 \tag{7}$$

$$\Sigma_{ii} = \frac{\alpha_i(\alpha_0 - \alpha_i)}{\alpha_0^2(\alpha_0 + 1)} \tag{8}$$

by recursively writing these moments as

$$\bar{p}_i(k+1) = \bar{p}_i(k) + \Sigma_{ii}(k) \frac{\delta_{i,i'} - \bar{p}_i(k)}{\bar{p}_i(k)(1 - \bar{p}_i(k))}$$

$$\Sigma_{ii}^{-1}(k+1) = \gamma_{k+1}\ \Sigma_{ii}^{-1}(k) + \frac{1}{\bar{p}_i(k+1)(1 - \bar{p}_i(k+1))}$$

where $\gamma_{k+1} = \frac{\bar{p}_i(k)(1-\bar{p}_i(k))}{\bar{p}_i(k+1)(1-\bar{p}_i(k+1))}$. Furthermore, it was shown that these mean-variance recursions, just as their count-equivalent counterparts, can be slow in detecting changes if the model is non-stationary. Hence, a modified set of recursions was derived that showed that the following recursions provided a much more effective change-detection mechanism.

$$\bar{p}_i(k+1) = \bar{p}_i(k) + 1/\lambda_k \Sigma_{ii}(k) \frac{\delta_{i,i'} - \bar{p}_i(k)}{\bar{p}_i(k)(1 - \bar{p}_i(k))} \tag{9}$$

$$\Sigma_{ii}^{-1}(k+1) = \lambda_k \gamma_{k+1} \Sigma_{ii}^{-1}(k) + \frac{1}{\bar{p}_i(k)(1 - \bar{p}_i(k))} \tag{10}$$

The key change was the addition of an effective process through the use of a discount factor $0 < \lambda_k \le 1$, and this allowed for a much faster estimator response [23].

B. Robustness

While an adaptation mechanism is useful to account for changes in the transition probabilities, the estimates of the transition probabilities are only guaranteed to converge in the limit of an infinite number of observations. While in practice the estimates do not require an unbounded number of observations, simply replacing the uncertain model $\tilde{A}$ with the best estimate $\hat{A}$ may lead to a biased value function [3] and sensitive policies, especially if the estimator has not yet converged to the true parameter A. For the purposes of this paper, the robust counterpart of Eq. (2) is defined as [1], [2]

$$J_R^* = \min_{\tilde{A} \in \mathscr{A}} \max_{\mu} \mathbf{E}\left[J_\mu(i_0)\right] \tag{11}$$

Like the nominal problem, the objective function is maximized with respect to the control policy; however, for the robust counterpart, the objective is minimized with respect to the *uncertainty set* $\mathscr{A}$.

When the uncertainty model $\mathscr{A}$ is described by a Bayesian prior, scenario-based methods can be used to generate realizations of the transition probability model. This gives rise to a scenario-based robust method which can turn out to be computationally intensive, since the total number of scenarios needs to be large [21]. This motivated our work [4] that, given a prior Dirichlet distribution on the transition probabilities, deterministically generates samples of each transition probability row $\mathscr{Y}_i$ (so-called Dirichlet Sigma Points) using the first two statistical moments of each row of the transition probability matrix, $\bar{p}$ and Σ,

$$\mathscr{Y}_0 = \bar{p}$$

$$\mathscr{Y}_i = \begin{cases} \bar{p} + \beta \left(\Sigma^{1/2}\right)_i & \forall i = 1, \ldots, N \\ \bar{p} - \beta \left(\Sigma^{1/2}\right)_i & \forall i = N+1, \ldots, 2N \end{cases}$$

where β is a tuning parameter that depends on the level of desired conservatism, which in turn depends on the size of the credibility region. Here, $\Sigma_i^{1/2}$ denotes the i^{th} row of the matrix square root of Σ. The uncertainty set $\mathscr{A}$ contains the deterministic samples $\mathscr{Y}_i\ \forall i \in \{1, 2, \ldots, 2N\}$.

V. Robust Adaptation

There are many choices for replanning efficiently using model-based methods, such as Real Time Dynamic Programming (RTDP) [24], [25]. RTDP assumes that the transition probabilities are unknown, and are continually updated through an agent's actions in the state space. Due to computational considerations, only a single sweep of the value iteration is performed at each measurement update. The result of Gullapalli [25] shows that if each state and action are executed infinitely often, then the (asynchronous) value iteration algorithm converges to the true value function. An alternative strategy is to perform synchronous value iteration, by using a bootstrapping approach where the old policy is used as the initial guess for the new policy [26].

In this section, we consider the full *robust* replanning problem (see Algorithm 1). The two main steps are an adaptation step, where the Dirichlet distributions (or alternatively, the Dirichlet Sigma Points) for each row and action are updated based on the most recent observations, and a robust replan step. For this paper, we use the Dirichlet Sigma Points to find the robust policy by using scenario-based methods, but we note that the following theoretical results apply to any robust value function. While appealing to account for both robustness and adaptation, it is critical to demonstrate that the proposed algorithm in fact converges to the true, optimal solution in the limit. We show this next.

Algorithm 1 Robust Replanning

Initialize uncertainty model: for example, Dirichlet distribution parameters α

while Not finished **do**

Using a statistically efficient estimator, update estimates of the transition probabilities (for each row, action). For example using the discounted estimator of Eq. 9

$$\bar{p}_i(k+1) = \bar{p}_i(k) + 1/\lambda_k \Sigma_{ii}(k) \frac{\delta_{i,i'} - \bar{p}_i(k)}{\bar{p}_i(k)(1-\bar{p}_i(k))}$$
$$\Sigma_{ii}^{-1}(k+1) = \lambda_k \gamma_{k+1} \Sigma_{ii}^{-1}(k) + \frac{1}{\bar{p}_i(k)(1-\bar{p}_i(k))}$$

For each uncertain row of the transition probability matrix, find the robust policy using robust DP

$$\min_{\mathscr{A}} \max_{\mu} \mathbf{E}\left[J_\mu\right] \tag{12}$$

For example, update the Dirichlet Sigma Points (for each row, action),

$$\begin{aligned} \mathscr{Y}_0 &= \bar{p} \\ \mathscr{Y}_i &= \bar{p} + \beta \left(\Sigma^{1/2}\right)_i \ \forall i = 1,\ldots,N \\ \mathscr{Y}_i &= \bar{p} - \beta \left(\Sigma^{1/2}\right)_i \ \forall i = N+1,\ldots,2N \end{aligned} \tag{13}$$

and find new robust policy $\min_\mu \max_{\mathscr{A}\in\mathscr{Y}} \mathbf{E}\left[J_\mu\right]$

Return

end while

A. Convergence

Gullapalli and Barto [25] showed that in an adaptive (but non-robust) setting, an asynchronous version of the Value Iteration algorithm converges to the optimal value function.

Theorem 1: [25] (*Convergence of an adaptive, asynchronous value iteration algorithm*) For any finite state, finite action MDP with an infinite-horizon discounted performance measure, an indirect adaptive asynchronous value iteration algorithm converges to the optimal value function with probability one if:

1) the conditions for convergence of the non-adaptive algorithm are met;
2) in the limit, every action is executed from every state infinitely often;
3) the estimates of the state transition probabilities remain bounded and converge in the limit to their true values with probability one.

Proof: See [25]. ∎

Using the framework of the above theorem, the robust counterpart to this theorem is stated next.

Theorem 2: (*Convergence of a* **robust** *adaptive, asynchronous value iteration algorithm*) For any finite state, finite action MDP with an infinite-horizon discounted performance measure, a robust, indirect adaptive asynchronous value iteration algorithm of Theorem 1

$$J_{k+1}(i) = \begin{cases} \min_\mu \max_{A_k \in \mathscr{A}_k} \mathbf{E}[J_\mu] & \text{if } i \in B_k \in S \\ J_k(i) & \text{Otherwise} \end{cases} \tag{14}$$

converges to the optimal value function with probability one if the conditions of Theorem 1 are satisfied, and the uncertainty set $\mathscr{A}_k$ converges to the singleton $\hat{A}_k$, in other words, $\lim_{k\to\infty} \mathscr{A}_k = \{\hat{A}_k\}$. Here B_k denotes the subset of states that are updated at each time step.

Proof: The key difference between this theorem and Theorem 1 is the maximization over the uncertainty set $\mathscr{A}_k$. However, as additional observations are incurred and by virtue of the convergent, unbiased estimator, the size of the uncertainty set will decrease to the singleton unbiased estimate $\hat{A}_k$. Furthermore, since the robust operator given by $T \doteq \min_\mu \max_{A_k \in \mathscr{A}_k}$ is a contraction mapping [1], [2]. Using both of these arguments, and since this unbiased estimate will in turn converge to the true value of the transition probability, then the robust adaptive asynchronous value iteration algorithm will converge to the true, optimal solution. ∎

Corollary 3: (*Convergence of synchronous version*) The synchronous version of the robust, adaptive MDP will converge to the true, optimal value function.

Proof: In the event that an entire sweep of the state space occurs at each value iteration, then the uncertainty set $\mathscr{A}_k$ will still converge to the singleton $\{\hat{A}_k\}$. ∎

Remark: (*Convergence of robust adaptation with Dirichlet Sigma Points*) For the Dirichlet Sigma Points, the discounted estimator of Eq. 9 converges in the limit of a large number of observations (with appropriate choice of λ_k), and the covariance Σ is eventually driven to 0, then each of the Dirichlet Sigma Points will collapse to the singleton, the unbiased estimate of the true transition probabilities. This means that the model will have converged, and that the robust solution will in fact have converged to the optimal value function.

VI. Numerical Results

This section presents actual flight demonstrations of the proposed robust and adaptive algorithm on a persistent surveillance mission in the RAVEN testbed [22].

The UAVs are initially located at a base location, which is separated by some (possibly large) distance from the surveillance location. The objective of the problem is to maintain a specified number r of requested UAVs over the surveillance location at all times. The base location is denoted by Y_b, the surveillance location is denoted by Y_s, and a discretized set of intermediate locations are denoted by $\{Y_0, \ldots, Y_s - 1\}$. Vehicles can move between adjacent locations at a rate of one unit per time step.

The UAVs have a specified maximum fuel capacity F_{max}, and we assume that the rate $\dot{F}_{burn}$ at which they burn fuel may vary randomly during the mission: the probability of nominal fuel flow is given by p_{nom}. This uncertainty in the fuel flow may be attributed to aggressive maneuvering that may be required for short time periods, for example. Thus, the total flight time each vehicle achieves on a given flight is a random variable, and this uncertainty must be accounted for in the problem. If a vehicle runs out of fuel while in flight, it crashes and is lost. The vehicles can refuel (at a rate $\dot{F}_{refuel}$) by returning to the base location.

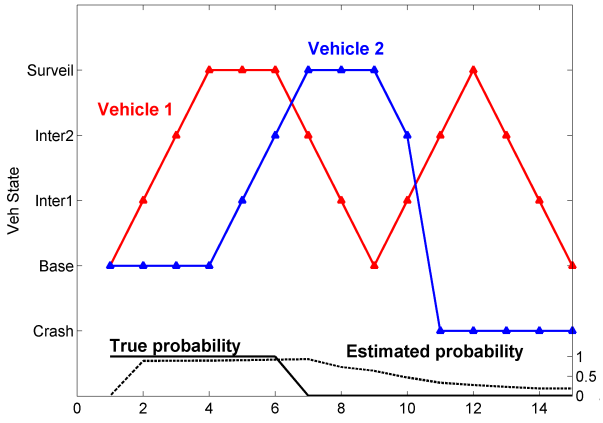
In this section, the adaptive replanning was implemented by explicitly accounting for the uncertainty in the probability of nominal fuel flow, $\tilde{p}_{nom}$. The replanning architecture updates both the mean and variance of the fuel flow transition probability, which is then passed to the online MDP solver, which computes the robust policy. This robust policy is then passed to the policy executer, which implements the control decision on the sytem. The Dirichlet Sigma Points were formed using updated mean and variance

$$\begin{aligned} \mathscr{Y}_0 &= \hat{p}_{nom} \\ \mathscr{Y}_1 &= \hat{p}_{nom} + \beta\ \sigma_p \\ \mathscr{Y}_2 &= \hat{p}_{nom} - \beta\ \sigma_p \end{aligned}$$

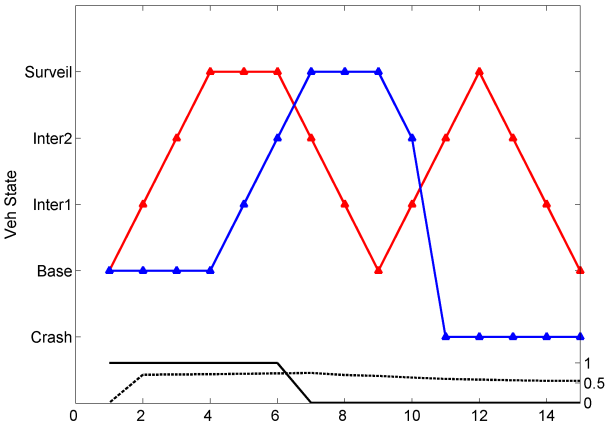
and used to find the robust policy. Using the results from the earlier chapters, appropriate choices of β could range from 1 to 5, where $\beta \approx 3$ corresponds to a 99% certainty region for the Dirichlet (in this case, the Beta density). For this scalar problem, the robust solution of the MDP corresponds to using a value of $\hat{p}_{nom} - \beta\sigma_p$ in place of the nominal probability estimate $\hat{p}_{nom}$.

Flight experiments were performed for a case when the probability estimate $\hat{p}_{nom}$ was varied in mid-mission, and three different replanning strategies were compared

- **Adaptive only:** The first replan strategy involved only an adaptive strategy, with $\lambda = 0.8$, and using only the estimate $\hat{p}_{nom}$
- **Robust replan, undiscounted adaptation:** This replan strategy used the undiscounted mean-variance estimator $\lambda = 1$, and set $\beta = 4$ for the Dirichlet Sigma Points
- **Robust replan, discounted adaptation:** This replan strategy used the undiscounted mean-variance estimator $\lambda = 0.8$, and set $\beta = 4$ for the Dirichlet Sigma Points



(a) Fast adaptation ($\lambda = 0.8$) with no robustness ($\beta = 0$)



(b) High robustness ($\beta = 4$) but slow adaptation ($\lambda = 1$)

Fig. 2. Experimental results showing vehicle trajectories (red and blue), and probability estimate used in the planning (black)

In all cases, the vehicle takes off from base, travels through 2 intermediate areas, and then reaches the surveillance location. In the nominal fuel flow setting losing 1 unit of fuel per time step, the vehicle can safely remain at the surveillance region for 4 time steps, but in the off-nominal fuel flow setting (losing 2 units), the vehicle can only remain on surveillance for only 1 time step. The main results are shown in Figure 2, where the transition in p_{nom} occurred at $t = 7$ time steps. At this point in time, one of the vehicles is just completing the surveillance, and is initiating the return to base to refuel, as the second vehicle is heading to the surveillance area. The key to the successful mission, in the sense of avoiding vehicle crashes, is to ensure that the change is detected sufficiently quickly, and that the planner maintains some level of cautiousness in this estimate by embedding robustness. The successful mission will detect this change rapidly, and leave the UAVs on target for a shorter time. The result of Figure 2(a) ignores any uncertainty in the estimate but has a fast adaptation (since it uses the factor $\lambda = 0.8$). However, by not embedding the uncertainty, the estimator while detecting the change in p_{nom} quickly, nonetheless allocates the second vehicle to remain at the surveillance region. Consequently, one of the vehicles runs out of fuel

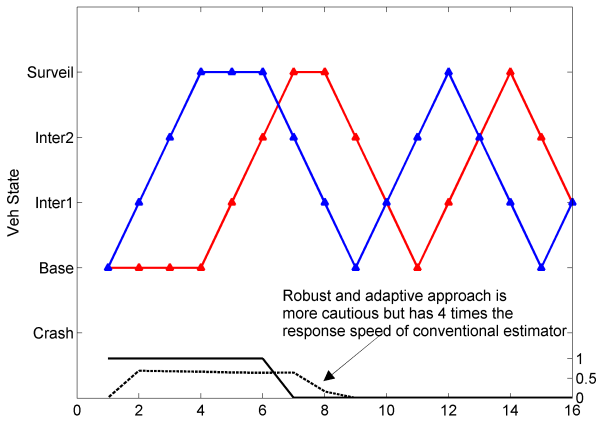


Fig. 3. Fast adaptation ($\lambda = 0.8$) with robustness ($\beta = 4$)

and crashes. At the second cycle of the mission, the second vehicle remains at the surveillance area for only 1 time step.

The result of Figure 2(b) accounts for uncertainty in the estimate but has a slow adaptation (since it uses the factor $\lambda = 1$). However, while embedding the uncertainty, the replanning is not done quickly, and for this different reason from the adaptive, non-robust example, one of the vehicle runs out of fuel, and crashes. At the second cycle of the mission, the second vehicle remains at the surveillance area for only 1 time step.

Figure 3 shows the robustness and adaptation acting together to cautiously allocate the vehicles, while responding quickly to changes in p_{nom}. The second vehicle is allocated to perform surveillance for only 2 time steps (instead of 3), and safely returns to base with no fuel remaining. At the second cycle, both vehicles only stay at the surveillance area for 1 time step. Hence, the robustness and adaptation have together been able to recover mission efficiency by bringing in their relative strengths: the robustness by accounting for uncertainty in the probability, and the adaptation by quickly responding to the changes in the probability.

VII. Conclusions

This paper has presented a combined robust and adaptive framework that accounts for errors in the transition probabilities. This framework is shown to converge to the true, optimal value function in the limit of a large number of observations. The proposed framework has been verified both in simulation and actual flight experiments, and shown to improve transient behavior in the adaptation process and overall mission performance. Our current work is addressing a more active learning mechanism for the transition probabilities, by the use of exploratory actions specifically taken to reduce the uncertainty in the transition probabilities. Our future work will consider the problem of decentralization of the robust adaptive framework across multiple vehicles, specifically addressing the issues of model consensus in a multi-agent system, and the impact of any disagreement on the robust solution.

Acknowledgements

Research supported by AFOSR grant FA9550-08-1-0086.

References

[1] A. Nilim and L. E. Ghaoui, "Robust Solutions to Markov Decision Problems with Uncertain Transition Matrices," *Operations Research*, vol. 53, no. 5, 2005.

[2] G. Iyengar, "Robust Dynamic Programming," *Math. Oper. Res.*, vol. 30, no. 2, pp. 257–280, 2005.

[3] S. Mannor, D. Simester, P. Sun, and J. Tsitsiklis, "Bias and Variance Approximation in Value Function Estimates," *Management Science*, vol. 52, no. 2, pp. 308–322, 2007.

[4] L. F. Bertuccelli and J. P. How, "Robust Decision-Making for Uncertain Markov Decision Processes Using Sigma Point Sampling," *IEEE American Controls Conference*, 2008.

[5] D. E. Brown and C. C. White., "Methods for reasoning with imprecise probabilities in intelligent decision systems," *IEEE Conference on Systems, Man and Cybernetics*, pp. 161–163, 1990.

[6] J. K. Satia and R. E. Lave., "Markovian Decision Processes with Uncertain Transition Probabilities," *Operations Research*, vol. 21, no. 3, 1973.

[7] C. C. White and H. K. Eldeib., "Markov Decision Processes with Imprecise Transition Probabilities," *Operations Research*, vol. 42, no. 4, 1994.

[8] A. Bagnell, A. Ng, and J. Schneider, "Solving Uncertain Markov Decision Processes," *NIPS*, 2001.

[9] K. J. Astrom and B. Wittenmark, *Adaptive Control*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 1994.

[10] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction (Adaptive Computation and Machine Learning)*. The MIT Press, 1998.

[11] R. Jaulmes, J. Pineau, and D. Precup., "Active Learning in Partially Observable Markov Decision Processes," *European Conference on Machine Learning (ECML)*, 2005.

[12] R. Jaulmes, J. Pineau, and D. Precup., "Learning in Non-Stationary Partially Observable Markov Decision Processes," *ECML Workshop on Reinforcement Learning in Non-Stationary Environments*, 2005.

[13] P. Marbach, *Simulation-based methods for Markov Decision Processes*. PhD thesis, MIT, 1998.

[14] V. Konda and J. Tsitsiklis, "Linear stochastic approximation driven by slowly varying Markov chains," *Systems and Control Letters*, vol. 50, 2003.

[15] M. Sato, K. Abe, and H. Takeda., "Learning Control of Finite Markov Chains with Unknown Transition Probabilities," *IEEE Trans. on Automatic Control*, vol. AC-27, no. 2, 1982.

[16] P. R. Kumar and W. Lin., "Simultaneous Identification and Adaptive Control of Unknown Systems over Finite Parameters Sets.," *IEEE Trans. on Automatic Control*, vol. AC-28, no. 1, 1983.

[17] J. Ford and J. Moore, "Adaptive Estimation of HMM Transition Probabilities," *IEEE Transactions on Signal Processing*, vol. 46, no. 5, 1998.

[18] P. A. Ioannou and J. Sun, *Robust Adaptive Control*. Prentice-Hall, 1996.

[19] M. Alighanbari and J. P. How, "A Robust Approach to the UAV Task Assignment Problem," *International Journal of Robust and Nonlinear Control*, vol. 18, no. 2, 2008.

[20] M. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley, 2005.

[21] L. F. Bertuccelli, *Robust Decision-Making with Model Uncertainty in Aerospace Systems*. PhD thesis, MIT, 2008.

[22] B. Bethke, J. How, and J. Vian., " Group Health Management of UAV Teams With Applications to Persistent Surveillance," *IEEE American Controls Conference*, 2008.

[23] L. F. Bertuccelli and J. P. How, "Estimation of Non-Stationary Markov Chain Transition Models," *IEEE Conference on Decision and Control*, 2008.

[24] A. Barto, S. Bradtke, and S. Singh., " Learning to Act using Real-Time Dynamic Programming," *Artificial Intelligence*, vol. 72, pp. 81–138, 1993.

[25] V. Gullapalli and A. Barto., "Convergence of Indirect Adaptive Asynchronous Value Iteration Algorithms," *Advances in NIPS*, 1994.

[26] B. Bethke, L. Bertuccelli, and J. P. How, "Experimental Demonstration of MDP- Based Planning with Model Uncertainty," in *AIAA Guidance Navigation and Control Conference*, Aug 2008. AIAA-2008-6322.