2008 American Control Conference
Westin Seattle Hotel, Seattle, Washington, USA
June 11-13, 2008

Robust Markov Decision Processes Using Sigma Point Sampling

L. F. Bertuccelli and J. P. How

Abstract— **This paper presents a new robust decision-making algorithm that accounts for model uncertainty in finite state/action, Markov Decision Processes (MDPs). In particular we generate robust and optimal control policies using Sigma Point sampling methods for dynamic multi-stage problems where the probabilistic transition model of the MDP may be fixed, but uncertain. In the case of poorly known transition model governing a MDP, this paper shows that the total number of scenarios in a scenario-based robust optimization may be decreased by generating a small number of appropriately chosen samples of the model. The robust policy for the worst-case instance of the data can be approximated by identifying the minimum objective function obtained from these realizations. This paper compares the proposed approach to more direct sampling-based approaches in a machine repair problem. The numerical examples show reduction in the total number of simulations required to obtain robust solutions while achieving optimal results.**

I. Introduction

Markov Decision Processes (MDPs) rely on precise transition models to find the optimal control policies. However, the parameters describing the models may be subject to uncertainty, either because there is not sufficient information to fully characterize these models, or because the models may simply be incorrect. A nominal control policy can suffer a significant performance penalty if this uncertainty is not accounted for. Poor knowledge in the state transition matrix of a system, for example, can lead to significant variations of the objective [1].

Studies on the impact of uncertainty in the parameters of decision-making processes has been addressed by numerous authors. The work of Satia in Ref. [2] considered the online identification of the state transition matrix by observing the system's transitions across the states and updating the model for the transition matrix with these observations. The work of Kumar et al. [3]–[5] considered the problem of controlled Markov Chains, when the state transition matrix governing the chain was poorly known. An additional term in the objective function was added to account for an exploration stage to identify the uncertain parameters.

More recent work (e.g., [1], [6]) incorporated the uncertainty in the state transition matrix directly in MDP formulation and found policies both "optimal" in minimizing the cost and robust to the uncertainty in the optimization parameters. In particular, Ref. [1] considers both finite and infinite horizon problems, and derives a Robust Value Iteration (RVI) algorithm that shows that the classical value iteration algorithm can be used to solve a robust counterpart problem. Nilim and El Ghaoui [1] also present numerous computationally tractable uncertainty models that can be used with the RVI. One of these models is a scenario-based method that finds a robust policy using random samples from some unknown transition model. Other approaches have also proposed techniques for adaptively identifying the state transition matrix online [7]–[9], but were not primarily concerned with the robust problem.

Recent work by Jaulmes et al. [10], [11], Mannor et al. [12] and Delage and Mannor [13] has also addressed the impact of uncertainty in multi-stage decision problems. The work by Jaulmes has addressed the uncertainty present in the parameters of Partially Observable Markov Decision Processes (POMDPs). The solution method uses a direct sampling of the uncertain parameters and the solution of multiple POMDPs in the MEDUSA (Markovian Exploration with Decision based on the Use of Sampled models Algorithm) to select a control policy from the family of control policies generated by each realization of the POMDP models. Additional recent work by Mannor has investigated the issue of bias and variance in completely observable MDPs with poorly known model parameters. In particular, Ref. [12] discusses an analytical approximation to the mean and variance of the objective function of an infinite horizon MDP with uncertain parameters.

Using the terminology introduced by Mannor et al. [12], our main interest in this paper is the parametric variance of a MDP. Rather than solely being concerned with the inherent variability of the objective function generated by the probabilistic modeling of the system (internal variance), we are primarily concerned with the impact of uncertainty in the parameters of the state transition matrix on the objective function and deriving robust control policies that account for poor modeling of these parameters.

Our computational efficiency is obtained by using Sigma Point sampling [14] to generate a small set of appropriately chosen MDP model realizations (e.g., transition matrices), and finding the optimal robust control policy over these fixed, but uncertain, models.

This paper is outlined as follows. The general decision-making problem is introduced in Section II. Section III introduces our approach using Sigma Point sampling, and Section IV discusses numerical experiments performed on a machine repair problem.

II. Markov Decision Processes

The Markov Decision Process (MDP) framework that we consider in this paper consists of a set of states $x \in S$, a

L. F. Bertuccelli, PhD Candidate, Dept. of Aeronautics and Astronautics, MIT, Cambridge, MA 02139, USA, lucab@mit.edu

J. P. How, Professor, Dept. of Aeronautics and Astronautics, MIT, Cambridge, MA 02139, USA, jhow@mit.edu

set of control action $u \in \mathcal{U}$, a transition model A^u, and a reward model $g(x,u)$. The time-additive objective function is defined as

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

where $\mu = [u_1, u_2, \ldots, u_{N-1}]$ is the control policy.

The goal is to find an optimal control policy, that minimizes the expected objective[1] over a finite horizon of N steps given some known transition model A^u

$$\min_\mu \ \mathbf{E} J_\mu(x_0) \tag{2}$$

The optimal control is found by solving

$$u^*(i) \in \arg\min_{u \in \mathcal{U}} \ \mathbf{E} J_\mu(x_0) \quad \forall i \in S \tag{3}$$

The optimality of the policy u^* may not be guaranteed if there is no uncertainty in these model parameters. For example, in the event of a model mismatch between the model's state transition matrix, A^u, and the true underlying state transition matrix A^u, the implemented policy may no longer be optimal, since in general

$$\min_\mu \ \mathbf{E} J_\mu(x_0, A^u) \neq \min_\mu \ \mathbf{E} J_\mu(x_0, A^{',u})$$

A. Uncertainty: Modeling

The uncertain parameters we will be primarily concerned about are the entries of the state transition matrix, A. We will assume that the probabilities of the state transition matrix, a, are described by an uncertainty set $\mathcal{A}$, that is $a \in \mathcal{A}$. A common description for modeling this uncertainty is the bounded approach, where $\mathcal{A} = \{a \mid a^- \leq a \leq a^+\}$, where the bounds a^- and a^+ are used to provide information on the effective range of the probability. In this paper we assign a prior distribution f_D to the uncertain $a \sim f_D$ as our current work is concerned with the simultaneous online identification of the uncertain parameters and control of the system with state transition observations. Results for the alternative polytopic description of the uncertainty can be found in Ref. [1].

Our choice for f_D is the Dirichlet distribution[2]. The Dirichlet distribution f_D for the *row* of the transition matrix given by $\mathbf{p} = [p_1, p_2, \ldots, p_N]^T$ and parameter (or prior counts) $\alpha = [\alpha_1, \alpha_2, \ldots, \alpha_N]^T$, is defined as

$$f_D(\mathbf{p}|\alpha) \quad = \quad K \prod_{i=1}^{N} p_i^{\alpha_i - 1}, \quad \sum_i^N p_i = 1 \tag{4}$$

where K is a normalizing factor that ensures the probability distribution integrates to unity. The primary reasons for using the Dirichlet distribution are that the uncertain p_i satisfy (by construction) both $p_i \in [0,1]$ and the explicit unit sum constraint $\sum_i p_i = 1$. Furthermore, the Dirichlet distribution is defined by parameters α_i that can be interpreted as counts, or times that a particular state transition was observed, thus easily updating the distribution based on these observations.

B. Robustness

In the presence of uncertainty in the optimization parameters, the optimal control policy u^* generated from incorrect parameters may no longer be optimal. Even if one had access to an estimator that could report the best estimates $\hat{A}$ (in some maximum likelihood sense for example), simply replacing the uncertain parameters A with their best estimates $\hat{A}$ may lead to fragile results if one of the worst-case parameters are the true underlying ones driving the dynamic system. Thus we introduce a robust counterpart to the nominal problem. The robust counterpart of Eq. (3) will be defined as

$$\min_\mu \max_{A \in \mathcal{A}} \mathbf{E} J_\mu(x_0) \tag{5}$$

Like the nominal problem, the objective function is maximized with respect to the control policy; however, for the robust counterpart, the *uncertainty set* $\mathcal{A}$ for the transition matrix is given, rather than the actual state transition matrix A for the nominal problem. The objective is then minimized with respect to the worst case realization of the transition matrix A belonging to the uncertainty set $\mathcal{A}$. The robust policy is obtained by solving

$$u_R^* = \arg\min_\mu \max_{A \in \mathcal{A}} \mathbf{E} J_\mu \tag{6}$$

C. Robustness: Computational Tractability

The solution times for the robust optimization of Eq. (5)will be heavily dependent (just like in classical Dynamic Programming) on the number of stages in the problem, the dimension of the state, the number of control actions, but also on the choice of the uncertainty model for the parameters [1]. Robust Dynamic Programming [1] can be used to solve for the robust policy.

One of the methods proposed by Nilim considered scenario approaches for the uncertainty set. Here, the decision-maker had access to, or could generate scenarios, that could form a scenario set $\mathcal{A}_s$ that could then be used in performing the robust optimization of Eq. (5). This is similar to the random sampling from the MEDUSA approach, but these scenario-based approaches may require performing tradeoff studies on the total number of simulations actually required to accurately represent the uncertain system.

The idea of our paper stems from the Sigma Point sampling approach used to approximate a Gaussian distribution with a small deterministic number of samples [14]. The key idea that we extract from this is the following: *Approximate the uncertainty in the state transition matrix with a set of deterministically sampled transition matrices.* Thus, unlike the sampling approach of [10] (which uses a small number of arbitrary samples) a Sigma Point implementation of an MDP with uncertainty in the probabilistic description will deterministically select transition matrices that capture the full spectrum of the uncertainty about these distributions. This is explained in greater detail in the next section.

[1]Maximization or minimization of the objective function depends on whether the objective is seen as a profit (therefore, maximization) or a cost (therefore, minimization).

[2]The Dirichlet distribution is the multi-dimensional extension to the Beta distribution [15].

III. Sigma Point Sampling

A. Sigma Point Sampling

Sigma Point sampling [14] is a deterministic sampling technique that selects *statistically relevant* samples to approximate a Gaussian distribution for nonlinear filtering problems; the key idea is that it is easier to approximate a Gaussian distribution rather than linearizing an arbitrary nonlinear system. The Sigma Point algorithm is defined as follows for a Gaussian random vector x. If the random vector $\mathbf{x}$ is normally distributed with mean $\bar{\mathbf{x}}$ and covariance $\mathbf{R}$, $x \sim N(\bar{\mathbf{x}}, \mathbf{R})$, then the Sigma Points $\mathcal{M}_i$ can be formed deterministically as follows

$$\begin{aligned}
&\mathcal{M}_0 = \bar{\mathbf{x}}, \qquad w_0 = \kappa/(N+\kappa) \\
&\mathcal{M}_i = \bar{\mathbf{x}} + \left(\sqrt{(N+\kappa)\mathbf{R}}\right)_i, \quad \forall i = 1, \ldots, N \\
&\mathcal{M}_i = \bar{\mathbf{x}} - \left(\sqrt{(N+\kappa)\mathbf{R}}\right)_i, \quad \forall i = N+1, \ldots, 2N
\end{aligned}$$

The notation $(\mathbf{R}^{1/2})_i$ denotes the i^{th} row of the square root matrix of $\mathbf{R}$. Each of the samples carries a weight $w_i = 1/(2(N+\kappa))$ and a tuning parameter κ is used to modify the weights appropriately. For example, in the Gaussian case, an optimal choice of κ is to ensure that $N + \kappa = 3$. After these samples are propagated through the dynamic model, the posterior distribution can be recovered as

$$\begin{aligned}
\bar{\mathbf{x}}^+ &= \sum_i w_i \mathcal{M}_i^+ \\
\mathbf{R}^+ &= \sum_i w_i (\mathcal{M}_i^+ - \bar{\mathbf{x}}^+)(\mathcal{M}_i^+ - \bar{\mathbf{x}}^+)^T
\end{aligned} \tag{7}$$

where $\mathcal{M}_i^+$ are the Sigma Points propagated through the dynamic model. While this algorithm was developed for Gaussian distributions, we can obtain a two-moment approximation for the Dirichlet distribution to deterministically select samples from an uncertain probability distribution since:

1) the Dirichlet distribution is well-approximated by a mean and a covariance.
2) the samples $\mathcal{M}_i$ satisfy the requirements of a probability mass function, namely [15]: $\sum_i \mathcal{M}_i = 1$, and $0 \leq \mathcal{M}_i \leq 1$

The first point is satisfied since the parameters α_i can be recovered from a set of Dirichlet-distributed random variables only using first and second moment information [16]. Thus, it remains to show that the Sigma Point samples in the case of a Dirichlet satisfy a probability mass function subject to an appropriate choice of the weights w_i. The following propositions (proves in the Appendix) show that the Sigma Points generated for a probability distribution in fact satisfies the assumptions of a probability mass function, subject to an appropriate choice of weights.

Proposition 1: If $\mathbf{E}[\mathbf{p}]$ and $\mathbf{\Sigma}$ are the mean and covariance of a Dirichlet distribution, then each Sigma Point satisfies a probability mass function (pmf); namely, each

$$\mathcal{Y}_i = \mathbf{E}[\mathbf{p}] \pm \beta \mathbf{\Sigma}_i^{1/2} \tag{8}$$

satisfies $\mathbf{1}^T\ \mathcal{Y}_i = 1$, where $\mathbf{\Sigma}_i^{1/2}$ is the i^{th} column of the square root of the covariance matrix $\mathbf{\Sigma}$ ∎

The following additional proposition constrains the choice of the parameter β to ensure that the Sigma Points generated completely satisfies the requirements of a probability mass function.

Proposition 2: If $\mathbf{E}[\mathbf{p}]$ and $\mathbf{\Sigma}$ are the mean and covariance of a Dirichlet distribution, the maximum positive value for the parameter β, $\beta_{\max}$, that guarantees that each Sigma Point $\mathcal{Y}_i = \mathbf{E}[\mathbf{p}] \pm \beta_{\max,i} \mathbf{\Sigma}_i^{1/2}$ is a pmf, is given by

$$\beta_{\max,i} = \frac{1}{|\mathbf{\Sigma}_{ij}^{1/2}|} \min\left(\mathbf{E}[\mathbf{p}]_i, 1 - \mathbf{E}[\mathbf{p}]_i\right) \tag{9}$$

where $\mathbf{\Sigma}_{ij}^{1/2}$ is the $(i,j)^{th}$ entry of the square root of the covariance matrix $\mathbf{\Sigma}$, and $\mathbf{E}[\mathbf{p}]_i$ is the i^{th} row of the mean probability vector. Then, $\beta_{max} = \min_i \beta_{\max,i}$. ∎

Based on this statistical description on the uncertainty in the distribution $\mathbf{E}[\mathbf{p}]$, the Sigma Point sampling algorithm applied to uncertain MDPs selects the following Sigma Points (note that each Sigma Point $\mathcal{Y}_i$ in fact represents a deterministic realization of a row of the uncertain state transition matrix)

$$\begin{aligned}
&\mathcal{Y}_0 = \mathbf{E}[\mathbf{p}] \\
&\mathcal{Y}_i = \mathbf{E}[\mathbf{p}] + \beta_{\max} \left(\mathbf{\Sigma}^{1/2}\right)_i \ \forall i = 1, \ldots, N \\
&\mathcal{Y}_i = \mathbf{E}[\mathbf{p}] - \beta_{\max} \left(\mathbf{\Sigma}^{1/2}\right)_i \ \forall i = N+1, \ldots, 2N
\end{aligned} \tag{10}$$

Remark 1: The Sigma Point algorithm for an N_S dimensional vector requires $2N_S + 1$ total samples. Hence, even for a 100-state system, only 201 total samples are generated. Random sampling methods like MEDUSA [11] often use a heuristic number of samples, or need large-scale Monte Carlo investigation of the total number of simulations required to achieve a desired confidence level since the sampling is done in a completely random fashion. The Sigma Point algorithm however, explores along the principal components of the probability simplex identifying samples that have a β deviation along those components, and so captures the statistically relevant regions of uncertainty. Furthermore, since the number of samples scales linearly with the number of dimensions, the uncertainty can be absorbed readily in more sophisticated problems, without necessarily adding significant computation.

Remark 2: The two-moment approximation of the Dirichlet distribution implies that there might be inaccuracies in the third and higher moments of a reconstructed Dirichlet distribution. However, the higher moments of the Dirichlet decay to zero very rapidly (see for example Ref. [12]), and experience has shown that the two-moment approximation is quite accurate.

B. Robust Counterpart Using Sigma Point Sampling

The new robustness objective of Eq. (5) can now be specified in terms of the finite number of Sigma Point samples. Rather than solving the typically harder problem

$$J_R^* = \min_{\mu} \max_{A \in \mathcal{A}} \mathbf{E}\ J_\mu(x_0) \tag{11}$$

Algorithm 1 Sigma Point Sampling for Uncertain MDP

1: Select $\beta = [0, \beta_{\max}]$ using Proposition 2
2: Select uncertainty model for i^{th} row of transition matrix by choosing appropriate parameters α for the Dirichlet distribution, $A_{i,\cdot} \sim f_D(\mathbf{p} \mid \alpha)$
3: Calculate the mean and covariance

$$\begin{aligned} \mathbf{E}[\mathbf{p}] &= \mathbf{E}[A_{i,\cdot}] = \alpha_i / \sum_i \alpha_i \\ \mathbf{\Sigma} &= \mathbf{E}[(A_{i,\cdot} - \mathbf{E}[\mathbf{p}])(A_{i,\cdot} - \mathbf{E}[\mathbf{p}])^T] \end{aligned}$$

4: Generate the samples using the Sigma Point algorithm according to Eq. (10)
5: Solve the robust problem using the Sigma Points and Robust Dynamic Programming

$$J^*_{SP} = \min_{\mu} \max_{\mathcal{Y}_i} \mathbf{E} J_\mu(x_0)$$

over the entire parameters $A \in \mathcal{A}$, the robust optimization is solved over the restricted set of Sigma Points $\mathcal{Y} \subseteq \mathcal{A}$,

$$J^*_{SP} = \min_{\mu} \max_{A \in \mathcal{Y}} \mathbf{E}\ J_\mu(x_0) \tag{12}$$

The full implementation of the Sigma Point sampling approach for an uncertain MDP is shown in Algorithm 1. The choice of β and the selection of the Dirichlet distribution $f_D(\mathbf{p} \mid \alpha)$ are made prior to running the algorithm. Using the uncertainty description given by $f_D(\mathbf{p} \mid \alpha)$, the mean and covariance are used to generate the Sigma Points $\mathcal{Y}_i$, which are the realizations for each of the models of the uncertain MDP. Robust Dynamic Programming [1] is used to find the optimal robust policy.

IV. Numerical Results: Machine Repair Problem

This section considers numerical examples using a machine repair problem adapted from Ref [17], and will investigate the case when there is uncertainty in the state transition matrix of the system.

A machine can take on one of two states x_k at time k: i) the machine is either *running* ($x_k = 1$), or ii) broken (not running, $x_k = 0$). If the machine is running, a profit of \$100 is made. The control options available to the user are the following: if the machine is running, a user can choose to either i) perform maintenance (abbreviated as $u_k = m$) on the machine (thereby presumable decreasing the likelihood the machine failing in the future), or ii) leave the maching running without maintenance ($u_k = n$). The choice of maintenance has cost, C_{maint}, e.g., the cost of a technician to maintain the machine.

If the machine is broken, two choices are available to the user: i) repair the machine ($u_k = r$), or ii) completely replace the machine ($u_k = p$). Both of these two options come at a price, however; machine repair has a cost C_{repair}, while machine replacement is $C_{replace}$, where for any sensible problem specification, the price of replacement is greater than the repair cost $C_{replace} > C_{repair}$. If the machine is replaced, it is *guaranteed* to work for at least the next stage.

For the case of the machine running at the current time step, the state transitions are governed by the model

$$\begin{aligned} \text{Pr } (x_{k+1} = \text{fails} \mid x_k = \text{running}, u_k = \text{m}) &= \gamma_1 \\ \text{Pr } (x_{k+1} = \text{fails} \mid x_k = \text{running}, u_k = \text{n}) &= \gamma_2 \end{aligned}$$

For the case of the machine not running at the current time step, the state transition are governed by the following model

$$\begin{aligned} \text{Pr } (x_{k+1} = \text{fails} \mid x_k = \text{fails}, u_k = \text{r}) &= \gamma_3 \\ \text{Pr } (x_{k+1} = \text{fails} \mid x_k = \text{fails}, u_k = \text{p}) &= 0 \end{aligned}$$

Note that, consistent with our earlier statement that machine replacement guarantees machine function at the next time step, the transition model for the replacement is deterministic. From these two models, we can completely describe the transition model if the machine is running or not running at the current time step:

$$\begin{aligned} \text{Running } (x_k = 1) : A_1 &= \begin{bmatrix} \gamma_1 & 1-\gamma_1 \\ 1-\gamma_2 & \gamma_2 \end{bmatrix} \\ \text{Not Running } (x_k = 0) : A_0 &= \begin{bmatrix} \gamma_3 & 1-\gamma_3 \\ 1 & 0 \end{bmatrix} \end{aligned}$$

The objective is to find an optimal control policy such that $u_k(x_k = 0) \in \{$ r, p $\}$ if the machine is not running, and $u_k(x_k = 1) \in \{$ m, n $\}$ if the machine is running, for each time step. The state of the machine is assumed to be perfectly observable, and this can be solved using Dynamic Programming

$$J_k(i) = \max_{u_k \in U} \left[g(x_k, u_k) + \sum_j A^u_{ij} J_{k+1}(j) \right]$$

A. Uncertain Transition Models

In this numerical example, it is assumed that the transition model A_0 is uncertain; that is, there is uncertainty in the likelihood of the machine failing after is repaired. This is a credible assumption if the person repairing it is new to the job, for example, or there is some uncertainty on the original cause of the machine failure.

The robust control $u^*_{R,k}$ maximizes the objective function over all matrices A_0 in the uncertainty set $\mathcal{A}_0$ that minimize the objective function

$$J^*_k(i) = \max_{u_k \in U} \min_{\tilde{A} \in \mathcal{A}} \left[g(x_k, u_k) + \sum_j \tilde{A}^u_{ij} J^*_{k+1}(j) \right]$$

Note that since the transition model A_1 is well-known, the robust counterpart of the nominal problem only needs to be formulated for the model A_0.

The solution approach using Sigma Point Sampling generate realizations of the matrix A_0 based on Algorithm 1, and in particular, the Sigma Points were found by

$$\begin{aligned} \mathcal{Y}_0 &= \mathbf{E}[A_0] \\ \mathcal{Y}_i &= \mathbf{E}[A_0] + \beta_{\max} \left(\mathbf{\Sigma_A}^{1/2} \right)_i \ \forall i = 1, \ldots, N \\ \mathcal{Y}_i &= \mathbf{E}[A_0] - \beta_{\max} \left(\mathbf{\Sigma_A}^{1/2} \right)_i \ \forall i = N+1, \ldots, 2N \end{aligned} \tag{13}$$

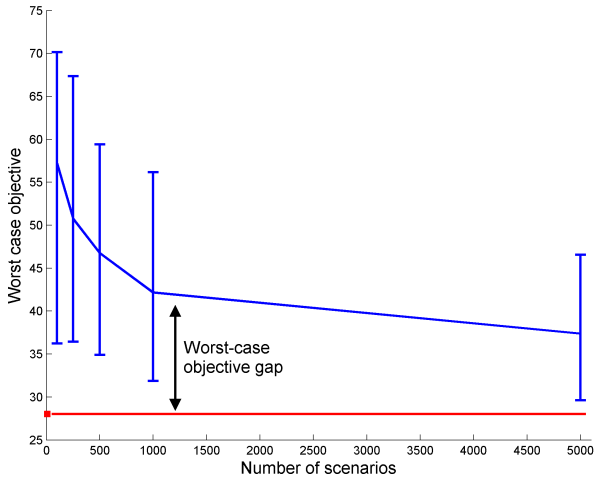

Fig. 1. The difference between the worst case objective through sampling (blue) and Sigma Point sampling (red) decreases only slightly as the number of simulations are increased significantly. The Sigma Point sampling strategy only truly requires 5 samples to find the worst case objective of $J^* = 28$, but the line has been extended for comparison.

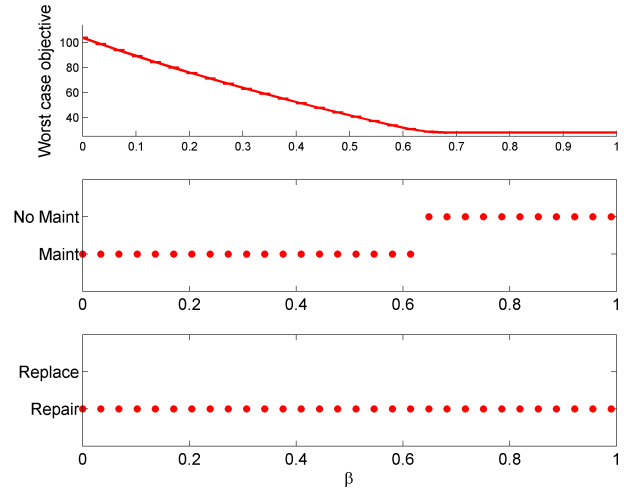

Fig. 2. Sigma Point sample tradeoff of robust performance (top subfigure) vs. normalized β shows that increasing the robustness also decreases the objective. The robust policy (bottom two figures) switches at $\beta = 0.65$.

B. Numerical Results

The machine repair problem with uncertain A_0 was evaluated multiple times with random realizations for the transition matrix A_0, and compared with the Sigma Point algorithm.

The main result comparing the Sigma Point approach to random sampling is shown in Figure 1 where the worst case objective (y-axis) is plotted as a function of the number of samples required. The blue line is the worst case found by using conventional sampling, and the red line is the Sigma Point worst-case using $\beta = 3$. This choice of β was in fact sufficient for this example to find the worst case of $J_{wc} = 28$. Note the slow convergence of the brute force sampling, with a significant gap even with 1200 samples. The Sigma Point only required 5 samples, since the uncertainty was only in one transition model of dimension $\mathcal{R}^{2\times 2}$. Hence, $N_s = 2 \times 2 + 1 = 5$. Note that the number of scenarios required to find the worst case varied significantly with the choice of hyperparameters α_i of the Dirichlet distribution. When $\alpha_i \approx 100$, for example, the Dirichlet distribution has a much smaller variance than when $\alpha_i \approx 10$ and the total number of samples required to find the worst case for $\alpha_i \approx 10$ is smaller than $\alpha_i \approx 100$.

Figure 2 shows the performance of the worst case as a function of the parameter $\beta \in [0, 1]$. The objective of this figure is to show the tradeoff between protecting against the worst-case and choice of the parameter β. Since the Sigma Points only require a small number of samples to find the worst case in this smaller machine repair example, this tradeoff can be performed very quick.y

The worst case objective was found for each value of β and is shown in the top figure. The bottom two subfigures show the policy as a function of β. For $\beta < 0.65$, the optimal (robust) policy is to perform maintenance, while if $\beta \geq 0.65$, the outcome of the maintenance is too uncertain, and it will be more cost effective (in a worst-case sense) to not perform maintenance at all. Hence, there is a very discrete policy switch at $\beta = 0.65$ that says that a different, decision should be made in response to the high uncertainty in the transition model.

V. Conclusions and Future Work

This paper has presented a computationally tractable technique for efficiently simulating an uncertain Markov Decision Process. We have shown computational savings over otherwise brute force numerical simulation, while at the same time maintaining consistent performance with more numerically intensive approaches. Further investigations on the optimal choice for $\beta_{\max}$ for a Dirichlet distribution are warranted.

Current work is addressing the case of uncertain rewards. When the rewards are not well known, they can be approximated by a Gaussian distribution with known mean and variance [13]. Since the original Sigma Point algorithm was designed for a Gaussian framework, it is rather natural to extend this to sampling from both the Dirichlet distribution (for the state transition matrix), and the Gaussian distribution (for the rewards) using the Sigma Point algorithm. Another immediate area of future work is extensions to the case of time-varying uncertainty models and incorporating adaptation of the uncertainty models. As additional observations become available to update the uncertainty of the transition models, the goal is to use this new information to generate new policies, that should help reduce the potential conservatism of otherwise worst-case policies.

Acknowledgements

Research supported by AFOSR grant FA9550-04-1-0458. The authors are grateful to Han-Lim Choi for critical reading and comments.

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] J. Satia and R. E. Lave., "Markovian decision processes with uncertain transition probabilities," *Operations Research*, 1973.

[3] P. Kumar and A. Becker., "A new family of optimal adaptive controllers for markov chains.," *IEEE Trans. on Automatic Control*, vol. AC-27, no. 1, 1982.

[4] P. Kumar and W. Lin., "Optimal adaptive controllers for unknown markov chains.," *IEEE Trans. on Automatic Control*, vol. AC-27, no. 4, 1982.

[5] P. 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.

[6] C. White and H. K. Eldeib., "Markov decision processes with imprecise transition probabilities," *Operations Research*, 1994.

[7] J. Buckley and E. Eslami, "Fuzzy markov chains: Uncertain probabilities.," *Mathware and Soft Computing*, vol. 9, pp. 33–41, 2002.

[8] R. Israel, J. S. Rosenthal, and J. Z. Wei., "Finding generators for markov chains via empirical transition matrices with applications to credit ratings," *Mathematical Finance, Vol. 11, No. 2*, vol. 11, no. 2, pp. 245–265, 2001.

[9] M. Sato, K. Abe, and H. Takeda., "Learning control of finite markov chains with unknown transition probabilities," *IEEE Trans. on Automatic Control*, 1982.

[10] R. Jaulmes, J. Pineau, and D. Precup., "Active learning in partially observable markov decision processes," *European Conference on Machine Learning (ECML)*, 2005.

[11] 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.

[12] 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.

[13] E. Delage and S. Mannor, "Percentile optimization for markov decision processes with parameter uncertainty," *subm to Operations Research*, 2007.

[14] S. Julier and J. Uhlmann, "Unscented filtering and nonlinear estimation," *Proc. of IEEE*, vol. 92, no. 3, 2004.

[15] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*. McGraw-Hill, 1991.

[16] N. Wicker, J. Mullera, R. Kiran, R. Kalathura, and O. Pocha, "A maximum likelihood approximation method for dirichlet's parameter estimation," *Computational Statistics and Data Analysis*, vol. 52, no. 3, pp. 1315–1322, 2008.

[17] D. Bertsekas, *Dynamic Programming and Optimal Control.* Athena Scientific, 2005.

VI. APPENDIX: SIGMA POINT SELECTION PROOFS

This appendix shows that the Sigma Point selection algorithm generates samples that are probability mass functions. The first proposition shows that the row sum of the variance of an uncertain probability mass function is identically zero. This is then used to show that .

Proposition 3: (Row/column sum constraint of a probability mass function's covariance matrix $\mathbf{\Sigma}$) The row and column sums of the entries of the covariance matrix of a probability distribution $\mathbf{\Sigma}$ are equal to 0.

Proof: Given a probability mass function $\mathbf{p} = [p_0, p_1, \ldots, p_N]^T$, then the covariance matrix of this pmf is given by $\mathbf{\Sigma} = \mathbf{E}[(\mathbf{p} - \mathbf{E}[\mathbf{p}])(\mathbf{p} - \mathbf{E}[\mathbf{p}])^T]$. However, since $\mathbf{p}$ is a pmf, then $p_N = 1 - \sum_i p_i$, and thus the covariance matrix $\mathbf{\Sigma}$ will *not* be full rank, implying that $\exists v$ (a left eigenvector) such that

$$\mathbf{v}^T \mathbf{\Sigma} = \lambda \mathbf{v}^T = \mathbf{0} \tag{14}$$

where λ is the eigenvalue, equal to 0 since the matrix $\mathbf{\Sigma}$ is not full rank. One such eigenvector is the vector of ones, $\mathbf{1} = [1,\ 1,\ 1,\ \ldots,\ 1]^T$

$$\begin{aligned}
\mathbf{1}^T \mathbf{\Sigma}_i &= \mathbf{E}\left[(p_0 - \mathbf{E}[p_0]) \sum_{i=0}^{N} (p_i - \mathbf{E}[p_i])\right] \\
&= \mathbf{E}\left[(p_0 - \mathbf{E}[p_0]) \left(\underbrace{\sum_i p_i}_{=1} - \underbrace{\sum_i \mathbf{E}[p_i]}_{=1}\right)\right] \\
&= 0 \quad \blacksquare
\end{aligned}$$

We can also show that the square root of this variance (found for example by using Singular Value Decomposition) also satisfies the property that $\mathbf{1}^T\ (\Sigma^{1/2})_i = 0$. Hence, these two results show that

$$\mathbf{1^T}(\bar{\mathbf{p}} + \beta \mathbf{\Sigma^{1/2}}) = \mathbf{1^T}\bar{\mathbf{p}} = \mathbf{1} \tag{15}$$

and Proposition 1 holds.

An important point, nonetheless, is that an appropriate selection for β is still required; while the pmf constraint is implicitly satisfied by the above results, each entry is not enforced to satisfy a valid probability: i.e., there is no constraint on each probability to be non-negative or greater (in magnitude) to 1, only the sum constraint is satisfied with this approach.

Proposition 4: (Selection of β) If $\mathbf{E}[\mathbf{p}]$ and $\mathbf{\Sigma}$ are the mean and covariance of a Dirichlet distribution, the maximum positive value for the parameter β, β_{max}, that guarantees that each Sigma Point satisfies $0 \leq \mathcal{Y}_i \leq 1$ is given by

$$\beta_{max} = \frac{1}{|\mathbf{\Sigma}_{ij}^{1/2}|} \min\left(\mathbf{E}[\mathbf{p}]_i, 1 - \mathbf{E}[\mathbf{p}]_i\right) \tag{16}$$

where $\mathbf{\Sigma}_{ij}^{1/2}$ is the $(i,j)^{th}$ entry of the square root of the covariance matrix $\mathbf{\Sigma}$, and $\mathbf{E}[\mathbf{p}]_i$ is the i^{th} row of the mean probability vector

Proof: To ensure that

$$0 \leq \mathbf{E}[\mathbf{p}]_i \pm \beta \mathbf{\Sigma}_{ij}^{1/2} \leq 1 \quad \forall i \tag{17}$$

we can address each side of the inequality, and it follows that the maximal β that guarantees that each entry of the pmf is a valid probability is given by the minimum of these two quantities,

$$\beta_{max} = \frac{1}{|\mathbf{\Sigma}_{ij}^{1/2}|} \min\left(\mathbf{E}[\mathbf{p}]_i, 1 - \mathbf{E}[\mathbf{p}]_i\right) \tag{18}$$

Note that since $\mathbf{E}[\mathbf{p}]_i < 1$ and typically $\mathbf{\Sigma}_{ij}^{1/2} < \mathbf{E}[\mathbf{p}]_i$, the value of β_{max} will generally be greater than 1. $\blacksquare$