Sensitivity Analysis of POMDP Value Functions

Stephane Ross,
Carnegie Mellon Universiy
Pittsburgh, USA

Masoumeh Izadi, Mark Mercer, David Buckeridge
McGill University
Montreal, Canada

Abstract

In sequential decision making under uncertainty, as in many other modeling endeavors, researchers observe a dynamical system and collect data measuring its behavior over time. These data are often used to build models that explain relationships between the measured variables, and are eventually used for planning and control purposes. However, these measurements cannot always be exact, systems can change over time, and discovering these facts or fixing these problems is not always feasible. Therefore it is important to formally describe the degree to which the model can tolerate noise, in order to keep near optimal behavior. The problem of finding tolerance bounds has been the focus of many studies for Markov Decision Processes (MDPs) due to their usefulness in practical applications. In this paper, we consider Partially Observable MDPs (POMDPs), which is a more realistic extension of MDPs with a wider scope of applications. We address two types of perturbations in POMDP model parameters, namely additive and multiplicative, and provide theoretical bounds for the impact of these changes in the value function. Experimental results are provided to illustrate our POMDP perturbation analysis in practice.

1 Introduction

Partially Observable MDPs (POMDPs) [4], as a natural extension of MDPs for sequential decision making and planning under uncertainty are becoming popular in different application domains. For instance in the medical field, researchers have proposed these models to support disease treatment decisions [7, 13, 12], drug infusion [11], epidemic control [17], and biosurveillance decision making [9]. The acceptance of these models has been limited in practice, however, due to factors such as the use of qualitative research to build decision models, the uncertainty involved in fixing some modeling assumptions, and the computational inefficiency of solution methods. Increasing the acceptance of these models will require addressing several issues including: developing an improved quantitative understanding of the domains being modeled in order to decrease uncertainties(e.g. progression of a disease or effects of treatments); improving solution methodologies to efficiently overcome computational difficulties of finding good solutions; and formally characterizing properties of these models which support uncertainties involved in real settings and stability of solutions provided. Demonstrating the effects of various parameter and initial value changes on system behavior is an important issue for model builders and model users in all disciplines. Sensitivity analysis is used to determine the sensitivity of model outputs to changes in the structure of the model or to changes in the value of the model parameters. This is a critical issue as models of dynamical systems may represent quantities that are very difficult, or even impossible, to measure with complete accuracy in the real world. Moreover, it is possible that small changes in parameter values occur dynamically over time.

While there has been some consideration of sensitivity analysis on model structure and model parameters for MDPs, few research have examined sensitivity of the POMDP value function. Perturbation analysis on model parameters has been previously developed for Markov processes [3], MDPs [18], and semi-MDPs [2]. These results show that when small additive noise (with respect to some parameters) is introduced into the transition functions of a Markov model the value function remains nearly optimal.

Research on MDPs and POMDPs in the areas of aggregation and dimensionality reduction [15, 5, 6, 16] can be thought of as sensitivity analysis with respect to the structure of the model. In this type of approaches, a restricted model which contains a small set of states, a set of high level actions, or a mapping of the model to a smaller model is considered. To ensure good quality of the solutions, the reduced model should provide results that are very similar to those produced by the original model of the environment. These studies normally provide methods to compute partial or suboptimal plans in the reduced model and evaluate the effect of resulting structural changes on the value function. Sensitivity analysis of value function with respect to perturbation of its parameters has been studied implicitly in the areas of generalization and knowledge transfer [19, 14, 10].

In these studies, the goal is to use a single value function that generalizes across many similar tasks or the same task under different conditions. Most of the work in this area has focused on designing strategies or representations that are empirically shown to be robust to environmental noise or imprecise world knowledge and can be applied to new scenarios.

Perturbation analysis for a general case POMDP is very complex. In this paper, we focus on the uncertainty with respect to POMDP model parameters. We address the instability of the value function caused by perturbation of POMDP parameter values. Since a random and arbitrary types of noise are difficult to handle and perhaps not very likely to happen in real world, we narrow our focus to parameter sensitivity with respect to some structured types of noise, namely additive and multiplicative. We investigate cases in which either transition functions, observation functions, and /or reward functions are slightly different at the execution time than the planning time and try to show that near optimality can be preserved up to some bounds.

2 POMDPs

Formally, a POMDP is defined by the following components: a finite set of hidden states S; a finite set of actions A; a finite set of observations Z; transition functions $T^a : S \times S \rightarrow [0,1]$ for all $a \in A$, such that $T^a(s,s')$ is the probability that the agent will end up in state s' after taking action a in state s; an observation function $O : A \times S \times Z \rightarrow [0,1]$, such that $O(a,s',z)$ gives the probability that the agent receives observation z after taking action a and getting to state s'; a reward function $R : S \times A \rightarrow \Re$, such that $r = R(s,a)$ is the immediate reward received when the agent takes action a in hidden state s and ends up in state s'; a discount factor, $\gamma \in (0,1)$; and an initial belief state b_0, which is a probability distribution over the set of hidden states S. The belief at time $t+1$ can be computed using Bayes rule. Each component of b_{t+1} corresponding to state s_i can be determined as follows:

$$b_{t+1}(s_i|a_t,z_{t+1}) = \frac{O(a_t,s_i,z_{t+1})\sum_{s\in S} T^{a_t}(s,s_i)b_t(s)}{\sum_{s'\in S} O(a_t,s',z_{t+1})\sum_{s\in S} T^{a_t}(s,s')b_t(s)} \quad (1)$$

The goal of a POMDP agent is to find a long term plan or *policy* for acting in such a way as to maximize the total expected reward received. The best such plan is called an *optimal policy* or an *optimal solution* for the POMDP. The belief state is known to be a sufficient statistic for computing an optimal policy in POMDPs [1]. Hence a policy is a mapping $\pi : B \rightarrow A$. The amount of total expected reward that an agent can accumulate over its lifetime as given by the horizon h and following a policy π is called the *value function* of π. The optimal policy π^*, in particular, is the one that maximizes the total expected future reward from any given belief state b:

$$\pi^*(b) = \arg\ \max_{\pi} E[\sum_{t=0}^{h-1} \gamma^t r_{t+1}|b] \quad (2)$$

Associated with each policy π is a set of conditional plans of the form π_p. Each such plan has an α-vector α_p. Each α-vector defines a hyperplane in belief space. The value of executing a conditional plan π_p from a belief state b is:

$$V_{\pi_p}(b) = \sum_{s\in S} V_{\pi_p}(s)b(s) \quad (3)$$

Therefore, in terms of the α-vectors we can write: $V_{\pi_p}(b) = \alpha_p \cdot b$. Finding good policies for POMDPs is generally difficult. POMDPs with as few as two states can have an optimal value function which needs an exponential number of α-vectors in their parsimonious set. The optimal value function for a POMDP is defined as:

$$V^*(b) = \max_{a\in A} \sum_{s\in S} R(s,a)b(s) + \gamma \sum_{z\in Z} P(z|a,b)V^*(b^{az}), \quad (4)$$

where $\sum_{s\in S} R(s,a)b(s)$ is the immediate reward at belief state b, and b^{az} is the next belief state after taking action a and making observation z form belief b.

3 Parameter Perturbations in POMDPs

In this section, we provide a general lower bound on the performance of the optimal policy computed from an approximate POMDP model $\hat{M}$, when executed in the exact POMDP model M, using $\hat{M}$ to maintain its belief. We study how this lower bound carries to different perturbations in the model, such as additive and multiplicative perturbations of the transition/observations probabilities and rewards. Such perturbations are common in practice, as often we do not know exactly the true values for T, O and R, and approximations $\hat{T}$, $\hat{O}$ and $\hat{R}$ are constructed from sampled data or imperfect domain knowledge. In such cases, we can often guarantee that the absolute error (related to additive noise) or relative error (related to multiplicative noise) on each parameter is smaller than some threshold with high probability, e.g. using the Hoeffding or Chernoff bounds. Given these thresholds, the following bounds can be useful to guarantee that the performance of the agent will not be too far from the optimal if it computes its policy and maintains its belief from the approximate model based on $\hat{T}$, $\hat{O}$ and $\hat{R}$. Such guarantees can be very important for the adoptions of POMDP solution methods to critical real-world applications. In the following sections we show how the alpha vectors representing the optimal value function might diverge with respect to T, O, R, and then deduce how the value function and the value of the corresponding policy also diverge.

3.1 General Bound

We first derive a general bound on the difference between the value functions obtained with the exact model M and approximate model $\hat{M}$. To achieve this, we first bound the difference between the state-values computed for any α-vector, and then use this bound to bound the difference in value functions. We then provide a general lower bound on the performance of the optimal policy computed from $\hat{M}$ when executed in M.

Let $M = (S,A,Z,T,O,R)$ be the exact POMDP model and $\hat{M} = (S,A,Z,\hat{T},\hat{O},\hat{R})$ be an approximate POMDP model where the transition function $\hat{T}$, observation function $\hat{O}$ and reward function $\hat{R}$ differ respectively from T, O, and R in the exact model. Let us also denote the joint transition-observation probabilities $P_{s,a}(s',z) = T^a(s,s')O(a,s',z)$ and similarly $\hat{P}_{s,a}(s',z) = \hat{T}^a(s,s')\hat{O}(a,s',z)$.

Recall that any α-vector $\alpha_t \in \Gamma_t$ represents the expected value obtained by a particular t-step conditional plan $\pi_t \in \Pi_t$, which can be specified by an immediate action a and a mapping $f_{t-1} : Z \rightarrow \Pi_{t-1}$ associating to every possible observation $z \in Z$ a $(t-1)$-step conditional plan $\pi_{t-1} \in \Pi_{t-1}$. The value of such α-vector, denoted $\alpha_t^{a,f_{t-1}}$, is then defined as:

$$\alpha_t^{a,f_{t-1}}(s) = R(s,a) + \gamma \sum_{s' \in S, z \in Z} P_{s,a}(s',z)\alpha_{f_{t-1}(z)}(s'), \qquad (5)$$

where α_π denotes the α-vector associated to conditional plan π. We now bound the difference between the values of the two α-vectors α_{π_t} and $\hat{\alpha}_{\pi_t}$ representing the same conditional plan, but where α_{π_t} is computed with the exact model M and $\hat{\alpha}_{\pi_t}$ is computed with model $\hat{M}$.

Lemma 1 *For any horizon t, $\sup_{\pi_t \in \Pi_t} ||\alpha_{\pi_t} - \hat{\alpha}_{\pi_t}||_\infty \leq \frac{||R-\hat{R}||_\infty}{1-\gamma} + \frac{\gamma||\hat{R}||_\infty}{(1-\gamma)^2} \sup_{s \in S, a \in A} ||P_{s,a} - \hat{P}_{s,a}||_1$.*

Proof: Consider any $\pi_t \in \Pi_t$ and its corresponding immediate action a and mapping f_{t-1}. Then for any state $s \in S$:

$$\begin{aligned}
& |\alpha_{\pi_t}(s) - \hat{\alpha}_{\pi_t}(s)| \\
= \; & |\alpha_t^{a,f_{t-1}}(s) - \hat{\alpha}_t^{a,f_{t-1}}(s)| \\
= \; & |R(s,a) + \gamma \textstyle\sum_{s' \in S, z \in Z} P_{s,a}(s',z)\alpha_{f_{t-1}(z)}(s') - \hat{R}(s,a) \\
& - \gamma \textstyle\sum_{s' \in S, z \in Z} \hat{P}_{s,a}(s',z)\hat{\alpha}_{f_{t-1}(z)}(s')| \\
\leq \; & |R(s,a) - \hat{R}(s,a)| \\
& + \gamma \textstyle\sum_{s' \in S, z \in Z} |P_{s,a}(s',z)\alpha_{f_{t-1}(z)}(s') - \hat{P}_{s,a}(s',z)\hat{\alpha}_{f_{t-1}(z)}(s')| \\
= \; & |R(s,a) - \hat{R}(s,a)| \\
& + \gamma \textstyle\sum_{s' \in S, z \in Z} |P_{s,a}(s',z)\alpha_{f_{t-1}(z)}(s') - P_{s,a}(s',z)\hat{\alpha}_{f_{t-1}(z)}(s') \\
& + P_{s,a}(s',z)\hat{\alpha}_{f_{t-1}(z)}(s') - \hat{P}_{s,a}(s',z)\hat{\alpha}_{f_{t-1}(z)}(s')| \\
\leq \; & |R(s,a) - \hat{R}(s,a)| + \gamma \textstyle\sum_{s' \in S, z \in Z} P_{s,a}(s',z)|\alpha_{f_{t-1}(z)}(s') - \hat{\alpha}_{f_{t-1}(z)}(s')| \\
& + \gamma \textstyle\sum_{s' \in S, z \in Z} |P_{s,a}(s',z) - \hat{P}_{s,a}(s',z)||\hat{\alpha}_{f_{t-1}(z)}(s')| \\
\leq \; & |R(s,a) - \hat{R}(s,a)| + \frac{\gamma||\hat{R}||_\infty}{1-\gamma}||P_{s,a} - \hat{P}_{s,a}||_1 \\
& + \gamma \sup_{\pi_{t-1} \in \Pi_{t-1}} ||\alpha_{\pi_{t-1}} - \hat{\alpha}_{\pi_{t-1}}||_\infty
\end{aligned}$$

Thus we have that:

$$\begin{aligned}
& \sup_{\pi_t \in \Pi_t} ||\alpha_{\pi_t} - \hat{\alpha}_{\pi_t}||_\infty \\
\leq \; & ||R - \hat{R}||_\infty + \frac{\gamma||\hat{R}||_\infty}{1-\gamma} \sup_{s \in S, a \in A} ||P_{s,a} - \hat{P}_{s,a}||_1 \\
& + \gamma \sup_{\pi_{t-1} \in \Pi_{t-1}} ||\alpha_{\pi_{t-1}} - \hat{\alpha}_{\pi_{t-1}}||_\infty
\end{aligned}$$

For $t = 1$, since π_1 is just an immediate action a, $\alpha_a(s) = R(s,a)$ and we have that $\sup_{\pi_1 \in \Pi_1} ||\alpha_{\pi_1} - \hat{\alpha}_{\pi_1}||_\infty = ||R - \hat{R}||_\infty$. Unfolding the previous recurrence up to $t = 1$, we find that:

$$\begin{aligned}
\sup_{\pi_t \in \Pi_t} ||\alpha_{\pi_t} - \hat{\alpha}_{\pi_t}||_\infty \; \leq \; & (\textstyle\sum_{i=0}^{t-1} \gamma^i)||R - \hat{R}||_\infty \\
& + (\textstyle\sum_{i=0}^{t-2} \gamma^i) \frac{\gamma||\hat{R}||_\infty}{1-\gamma} \sup_{s \in S, a \in A} ||P_{s,a} - \hat{P}_{s,a}||_1
\end{aligned}$$

Thus for all t:

$$\begin{aligned}
\sup_{\pi_t \in \Pi_t} ||\alpha_{\pi_t} - \hat{\alpha}_{\pi_t}||_\infty \; \leq \; & \frac{||R-\hat{R}||_\infty}{1-\gamma} \\
& + \frac{\gamma||\hat{R}||_\infty}{(1-\gamma)^2} \sup_{s \in S, a \in A} ||P_{s,a} - \hat{P}_{s,a}||_1
\end{aligned}$$

□ Note that as $\hat{R} \rightarrow R$, $\hat{T} \rightarrow T$ and $\hat{O} \rightarrow O$, the bound for $\sup_{\pi_t \in \Pi_t} ||\alpha_{\pi_t} - \hat{\alpha}_{\pi_t}||_\infty \rightarrow 0$. Hence we believe this bound is fairly tight for very small perturbations. We now use this bound to bound the difference between the value functions V_π and $\hat{V}_\pi$ of any policy π, where V_π is computed with the exact model M and $\hat{V}_\pi$ is computed with model $\hat{M}$.

Theorem 1 *For any policy π, $||V_\pi - \hat{V}_\pi||_\infty \leq \frac{||R-\hat{R}||_\infty}{1-\gamma} + \frac{\gamma||\hat{R}||_\infty}{(1-\gamma)^2} \sup_{s \in S, a \in A} ||P_{s,a} - \hat{P}_{s,a}||_1$.*

Proof: Consider any horizon t, policy $\pi_t \in \Pi_t$, and belief b, then:

$$\begin{aligned}
|V_{\pi_t}(b) - \hat{V}_{\pi_t}(b)| \; &= \; |\textstyle\sum_{s \in S} b(s)\alpha_{\pi_t}(s) - \sum_{s \in S} b(s)\hat{\alpha}_{\pi_t}(s)| \\
&\leq \; \textstyle\sum_{s \in S} b(s)|\alpha_{\pi_t}(s) - \hat{\alpha}_{\pi_t}(s)| \\
&\leq \; ||\alpha_{\pi_t} - \hat{\alpha}_{\pi_t}||_\infty \\
&\leq \; \frac{||R-\hat{R}||_\infty}{1-\gamma} + \frac{\gamma||\hat{R}||_\infty}{(1-\gamma)^2} \sup_{s \in S, a \in A} ||P_{s,a} - \hat{P}_{s,a}||_1
\end{aligned}$$

Taking the limit as $t \rightarrow \infty$ proves the theorem. □ Now, let π^* be the optimal policy for model M and $\hat{\pi}^*$ the optimal policy computed from model $\hat{M}$. The following theorem bounds the performance of the agent if it executes $\hat{\pi}^*$ in the true model M, using $\hat{M}$ to maintain its belief.

Theorem 2 *For any belief b, $V_{\hat{\pi}^*}(b) \geq V_{\pi^*}(b) - \frac{2||R-\hat{R}||_\infty}{1-\gamma} - \frac{2\gamma||\hat{R}||_\infty}{(1-\gamma)^2} \sup_{s \in S, a \in A} ||P_{s,a} - \hat{P}_{s,a}||_1$.*

Proof: Consider any horizon t and belief b. Let $\pi_t^* = \arg\max_{\pi_t \in \Pi_t} \sum_{s \in S} b(s)\alpha_{\pi_t}(s)$ and $\hat{\pi}_t^* = \arg\max_{\pi_t \in \Pi_t} \sum_{s \in S} b(s)\hat{\alpha}_{\pi_t}(s)$, be the optimal t-step conditional plans starting in belief b for model M and model $\hat{M}$ respectively.

To prove this lower bound, we consider that b is the initial belief, and that the agent executes its conditional plan

$\hat{\pi}_t^*$ using the history of action/observation to follow its plan. This is equivalent to first executing the immediate action specified $\hat{\pi}_t^*$, computing the next belief $\hat{b}'$ with $\hat{M}$ and then proceeding with the conditional plan $\hat{\pi}_{t-1}^*$ that is optimal for belief $\hat{b}'$ in $\hat{M}$. Hence looking at conditional plans allow us to ignore the fact that the belief maintained with $\hat{M}$ differs from the exact belief maintained with M after the first step if we bound $V_{\hat{\pi}_t^*}(b)$ by bounding $\sum_{s\in S} b(s)\alpha_{\hat{\pi}_t^*}(s)$, as $\alpha_{\hat{\pi}_t^*}$ implicitly takes this into account.

We know that for all $s \in S$:

$$\alpha_{\hat{\pi}_t^*}(s) \geq \hat{\alpha}_{\hat{\pi}_t^*}(s) - ||\alpha_{\hat{\pi}_t^*} - \hat{\alpha}_{\hat{\pi}_t^*}||_\infty$$

and also we have:

$$\hat{\alpha}_{\pi_t^*}(s) \geq \alpha_{\pi_t^*}(s) - ||\alpha_{\pi_t^*} - \hat{\alpha}_{\pi_t^*}||_\infty$$

It follows that

$$\sum_{s\in S} b(s)\alpha_{\hat{\pi}_t^*}(s) \geq \left(\sum_{s\in S} b(s)\hat{\alpha}_{\hat{\pi}_t^*}(s)\right) - ||\alpha_{\hat{\pi}_t^*} - \hat{\alpha}_{\hat{\pi}_t^*}||_\infty$$

and that

$$\sum_{s\in S} b(s)\hat{\alpha}_{\pi_t^*}(s) \geq \left(\sum_{s\in S} b(s)\alpha_{\pi_t^*}(s)\right) - ||\alpha_{\pi_t^*} - \hat{\alpha}_{\pi_t^*}||_\infty$$

Since $\hat{\pi}_t^*$ was considered better than π_t^* in $\hat{M}$, then

$$\sum_{s\in S} b(s)\hat{\alpha}_{\hat{\pi}_t^*}(s) \geq \sum_{s\in S} b(s)\hat{\alpha}_{\pi_t^*}(s)$$

Thus it follows that:

$$\sum_{s\in S} b(s)\alpha_{\hat{\pi}_t^*}(s) \geq \left(\sum_{s\in S} b(s)\alpha_{\pi_t^*}(s)\right) - 2\sup_{\pi_t\in\Pi_t} ||\alpha_{\pi_t} - \hat{\alpha}_{\pi_t}||_\infty$$

By Lemma 1 we have:

$$V_{\hat{\pi}_t^*}(b) \geq V_{\pi_t^*}(b) - \frac{2||R-\hat{R}||_\infty}{1-\gamma} - \frac{2\gamma||\hat{R}||_\infty}{(1-\gamma)^2}\sup_{s\in S,a\in A} ||P_{s,a} - \hat{P}_{s,a}||_1$$

Taking the limit as $t \to \infty$ proves the theorem. □ This result provides a general bound for any perturbations that we may have in the approximate model $\hat{M}$. The following sections consider particular noise models, such as additive and multiplicative noise, and show how this general bound can be applied to such perturbations. It should be noted that this bound is much tighter than the known worst case bound of $\frac{\hat{R}||_\infty}{1-\gamma}$.

3.2 Additive Bound

Definition: Let M and $\hat{M}$ be POMDPs with same sets of states, actions and observations. We say $\hat{M}$ is an (α,β,δ)-additive approximation of M if:

1. $\forall s,a,s' : T^a(s,s') - \alpha \leq \hat{T}^a(s,s') \leq T^a(s,s') + \alpha$.
2. $\forall a,s',z : O(a,s',z) - \beta \leq \hat{O}(a,s',z) \leq O(a,s',z) + \beta$.
3. $\forall s,a : R(s,a) - \delta \leq \hat{R}(s,a) \leq R(s,a) + \delta$.

We show using the previous general bound that if two POMDPs with the same structure for states, actions, and observations are sufficiently close in the (α,β,δ)-additive sense, then for any policy, the value function is similar in both POMDPs.

Theorem 3 *Let $\hat{M}$ be an (α,β,δ)-additive approximation of POMDP M. Then for any policy π, $||V_\pi - \hat{V}_\pi||_\infty \leq \frac{\delta}{1-\gamma} + \frac{\gamma||\hat{R}||_\infty}{(1-\gamma)^2}\min\{2, \alpha|S| + \beta|Z|\}$.*

Proof: By definition of $\hat{M}$, $||R-\hat{R}||_\infty \leq \delta$. Furthermore, we have that $\sup_{s\in S,a\in A} ||P_{s,a} - \hat{P}_{s,a}||_1 \leq |S|\alpha + |Z|\beta$:

$$\begin{aligned}
& ||P_{s,a} - \hat{P}_{s,a}||_1 \\
= \;& \textstyle\sum_{s'\in S,z\in Z} |P_{s,a}(s',z) - \hat{P}_{s,a}(s',z)| \\
= \;& \textstyle\sum_{s'\in S,z\in Z} |T^a(s,s')O(a,s',z) - \hat{T}^a(s,s')\hat{O}(a,s',z)| \\
= \;& \textstyle\sum_{s'\in S,z\in Z} |T^a(s,s')O(a,s',z) - T^a(s,s')\hat{O}(a,s',z) \\
& \quad + T^a(s,s')\hat{O}(a,s',z) - \hat{T}^a(s,s')\hat{O}(a,s',z)| \\
\leq \;& \textstyle\sum_{s'\in S,z\in Z} T^a(s,s')|O(a,s',z) - \hat{O}(a,s',z)| \\
& \quad + \textstyle\sum_{s'\in S,z\in Z} O(a,s',z)|T^a(s,s') - \hat{T}^a(s,s')| \\
\leq \;& \beta|Z| + \alpha|S|
\end{aligned}$$

Note, however, that in all cases $||P_{s,a} - \hat{P}_{s,a}||_1 \leq 2$, so we know that $\sup_{s\in S,a\in A} ||P_{s,a} - \hat{P}_{s,a}||_1 \leq \min\{2, \alpha|S| + \beta|Z|\}$. Thus by Theorem 1, we get $||V_\pi - \hat{V}_\pi||_\infty \leq \frac{\delta}{1-\gamma} + \frac{\gamma||\hat{R}||_\infty}{(1-\gamma)^2}\min\{2, \alpha|S| + \beta|Z|\}$. □

The following result lower bounds the performance of the optimal policy $\hat{\pi}^*$ computed from an (α,β,δ)-additive approximation of POMDP M, when executed in model M.

Theorem 4 *Let $\hat{M}$ be an (α,β,δ)-additive approximation of POMDP M. Then for any belief b, $V_{\hat{\pi}^*}(b) \geq V_{\pi^*}(b) - \frac{2\delta}{1-\gamma} - \frac{2\gamma||\hat{R}||_\infty}{(1-\gamma)^2}\min\{2, \alpha|S| + \beta|Z|\}$.*

Proof: The proof proceeds similarly to Theorem 2, but uses the bounds on $||R-\hat{R}||_\infty$ and $||P_{s,a} - \hat{P}_{s,a}||_1$ derived in Theorem 3 to bound the bound given by Lemma 1 at the last step. □

3.3 Multiplicative Bound

Definition: Let M and $\hat{M}$ be POMDPs with same sets of states, actions and observations. We say $\hat{M}$ is an (α,β,δ)-multiplicative approximation of M if:

1. $\forall s,a,s' : (1-\alpha)T^a(s,s') \leq \hat{T}^a(s,s') \leq (1+\alpha)T^a(s,s')$.

2. $\forall a,s',z : (1-\beta)O(a,s',z) \leq \hat{O}(a,s',z) \leq (1+\beta)O(a,s',z)$.

3. $\forall s,a$ such that $R(s,a) \geq 0 : (1-\delta)R(s,a) \leq \hat{R}(s,a) \leq (1+\delta)R(s,a)$.

4. $\forall s,a$ such that $R(s,a) < 0 : (1+\delta)R(s,a) \leq \hat{R}(s,a) \leq (1-\delta)R(s,a)$.

We show using the previous general bound that if two POMDPs with the same structure for states, actions, and observations are sufficiently close in the (α,β,δ)-multiplicative sense, then for any policy, the value function is similar in both POMDPs.

Theorem 5 *Let $\hat{M}$ be an (α,β,δ)-multiplicative approximation of POMDP M. Then for any policy π, $||V_\pi - \hat{V}_\pi||_\infty \leq \frac{||\hat{R}||_\infty}{(1-\gamma)}\left(\frac{\delta}{1-\delta} + \frac{\gamma(\alpha+\beta)}{1-\gamma}\right)$.*

Proof: By definition of $\hat{M}$, we have that $|T^a(s,s') - \hat{T}^a(s,s')| \leq \alpha T^a(s,s')$, $|O(a,s',z) - \hat{O}(a,s',z)| \leq \beta O(a,s',z)$ and $|R(s,a) - \hat{R}(s,a)| \leq \delta|R(s,a)|$. Furthermore, $|R(s,a)| \leq \frac{|\hat{R}(s,a)|}{1-\delta}$, so $|R(s,a) - \hat{R}(s,a)| \leq \frac{\delta|\hat{R}(s,a)|}{1-\delta}$. Hence $||R - \hat{R}||_\infty \leq \frac{\delta||\hat{R}||_\infty}{1-\delta}$. To bound $||P_{s,a} - \hat{P}_{s,a}||_1$, we consider a similar proof as in the previous bound for the additive case:

$$\begin{aligned}
& ||P_{s,a} - \hat{P}_{s,a}||_1 \\
= \;& \textstyle\sum_{s'\in S,z\in Z} |P_{s,a}(s',z) - \hat{P}_{s,a}(s',z)| \\
\leq \;& \textstyle\sum_{s'\in S,z\in Z} T^a(s,s')|O(a,s',z) - \hat{O}(a,s',z)| \\
& + \textstyle\sum_{s'\in S,z\in Z} O(a,s',z)|T^a(s,s') - \hat{T}^a(s,s')| \\
\leq \;& \beta \textstyle\sum_{s'\in S,z\in Z} T^a(s,s')O(a,s',z) \\
& + \alpha \textstyle\sum_{s'\in S,z\in Z} O(a,s',z)T^a(s,s') \\
= \;& \alpha+\beta
\end{aligned}$$

Thus by Theorem 1, we get

$$\begin{aligned}
||V_\pi - \hat{V}_\pi||_\infty &\leq \tfrac{\delta||\hat{R}||_\infty}{(1-\delta)(1-\gamma)} + \tfrac{\gamma||\hat{R}||_\infty}{(1-\gamma)^2}(\alpha+\beta) \\
&= \tfrac{||\hat{R}||_\infty}{(1-\gamma)}\left(\tfrac{\delta}{1-\delta} + \tfrac{\gamma(\alpha+\beta)}{1-\gamma}\right)
\end{aligned}$$

□ As in the additive case, we can bound the performance of the optimal policy $\hat{\pi}^*$ computed from an (α,β,δ)-multiplicative approximation of POMDP M, when executed in model M.

Theorem 6 *Let $\hat{M}$ be an (α,β,δ)-multiplicative approximation of POMDP M. Then for any belief b, $V_{\hat{\pi}^*}(b) \geq V_{\pi^*}(b) - \frac{2||\hat{R}||_\infty}{(1-\gamma)}\left(\frac{\delta}{1-\delta} + \frac{\gamma(\alpha+\beta)}{1-\gamma}\right)$.*

Proof: The proof proceeds similarly to Theorem 2, but uses the bound on $||R - \hat{R}||_\infty$ and $||P_{s,a} - \hat{P}_{s,a}||_1$ derived in Theorem 5 to bound the bound given by Lemma 1 at the last step. □ It is natural to think that a POMDP model can tolerate a relative noise (in the dynamics) better than an absolute noise of the same amount, since the transition and observation functions are all probability functions. However, this is not always true for the rewards. In the special case of $\delta = 0$, it is noted that the bound for multiplicative error in theorem 6 is tighter than the bound for an additive error in theorem 4 (with the same magnitude of α and β), even if the amount of noise is the same. More precisely, when $\delta = 0$:

$$additiveerror \leq \frac{2\gamma||\hat{R}||_\infty}{(1-\gamma)^2}\min\{2, \alpha|S| + \beta|Z|\} \qquad (6)$$

and

$$multiplicativeerror \leq \frac{2\gamma||\hat{R}||_\infty(\alpha+\beta)}{(1-\gamma)^2} \qquad (7)$$

We can also observe that the additive bound can be tighter than the general bound from theorem 2 depending on the value of α and β.

4 Experimental Results

Here we examine the noise in the transition functions,in the observation functions, and in rewards on one of the POMDP benchmarks, as a proof of concept. We study the precision of our bounds on a 4*x*4 grid described in the POMDP repository. In this 16-state problem, there is a single goal state which has perfect observability and there are 15 other states which look exactly the same. A decision making agent receives a reward value of 1 when it reaches the goal by choosing any of the available actions: move up, down, left, or right. Transition functions for all actions in all states are deterministic, except for the goal state, in which taking any action leaves the agent randomly in any other state. Therefore, for the case of this example, if we introduce noise in the dynamics of the model (and only in non-goal states) then the additive and multiplicative noises are the same amount. For the purpose of this paper, we can also assume that rewards can only have a real value between zero and one, so that additive and multiplicative noises in the reward have also the same amount.

We try to summarize the results of our evaluations for POMDP transition function in three scenarios: (1) goal becomes slower or harder to reach, in which the agent can gave a slight chance to jump back to states farther from the goal; (2)goal becomes easier or faster to reach, in which under some actions from some states the agent can have a slight chance to jump into states closer to the goal; (3) a fixed but random noise is introduced for all actions.

Figure 1 compares the performance of a perturbed POMDP model in the two scenarios (1) and (2) with a reference value obtained from the original 4x4 model for noise values between 10% to 40% of the original probabilities (i.e $\alpha \in [0.1, 0.4]$). In each scenario, we started the evaluation from the same given initial belief state and computed the total discounted reward after 1000 time steps. The results in

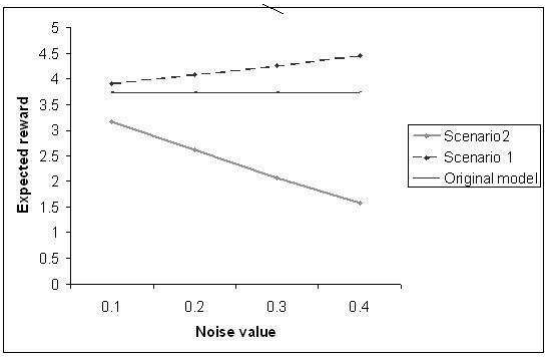

Figure 1. Comparison of the performance of POMDP policies in original and perturbed versions of 4x4 grid in scenarios 1 and 2.

this figure are averages over 250 such runs. Under the same

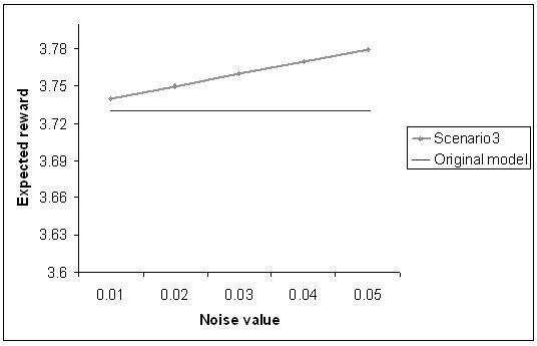

Figure 2. Comparison of the performance of POMDP policies in original and perturbed versions of 4x4 grid in scenario 3.

experimental setup, we also investigated the changes in the value function for the third scenario. This time we tried a small range of α in $[0.01, 0.05]$. Our results in Figures 1 and 2 confirm the linearity of value function error in terms of perturbation in transition functions in all three scenarios.

To investigate and compare the precision of different bounds, we used theorem 2 for the especial case of $R = \hat{R}$, and equations 6 and 7 to compute the general, additive, and multiplicative bounds respectively. For a particular value of $\alpha = 0.2$, table 1 shows these values. This table shows that multiplicative noise is the closest to the actual error achieved through our experiments. Different values of α will linearly scale the values presented in this table, according to theorems 2, 4, and 6 and our empirical results in figures 1 and 2.

Table 2 presents our experimental results with different perturbation in the observation functions, and the results of experiments with several values of noise in the reward

Table 1. The noise bounds comparison in transition function of 4x4 POMDP problem.

bound	**value for $\alpha = 0.2$**
general	24.96
additive	15.6
multiplicative	1.56
empirical error	1.12

Table 2. The noise bounds comparison in transition function of 4x4 POMDP problem.

parameter	**noise value**	**expected reward**
observation		
	0	3.73
	0.01	3.74
	0.02	3.74
	0.05	3.75
	0.1	3.83
reward		
	0	3.73
	0.05	3.55
	0.1	3.36
	1 (goal reward in state 14)	15.13
	1 (goal reward in state 5)	-

function. We tried a wide range of noise values and special cases, however, we can not present all the results here due to space constraint. In summary these results show that the reward functions of 4x4 grid were the most sensitive parameters of the model. A large value of noise can dramatically change the problem in a huge range. Here we tried to introduce new goal states in the 4x4 grid and observed a large variance in the solution both in terms of the time to find the solution and the value function.

5 Discussion

Instability in the POMDP models can occur in different situations when used in real settings. In many application the actual value of parameters for some action transitions and/or observation probabilities would not be known for certain and can be determined for instance by the accuracy of lab tests, recognition of true associations in noisy data, degree of reliability of detection devices,or sensitivity of sensors. The cost associated with different actions can also change depending on environmental factors such as population or temperature. In practice, although the exact level of accuracy is not known, the instability is not usually random either. Therefore, the idea of structured noise, investigated in this paper, can be crucial in using POMDPs in practical applications.

We anticipate that the discount factor γ plays an important role in the computation of all the bounds developed in this paper. However, we did not considered changing this parameter in relation to other parameters and its effect on the value function stability. In this paper we did not discuss the time complexity that a noise value can introduce, however, we would like to mention that in some cases (in particular for observations and rewards) even a slight change in parameters can dramatically increase the time to solve a perturbed version of the original POMDP. The last row of table 2 shows this effect. This is an important area of future investigation as well.

Our results show that the bounds can have a good precision for an arbitrary POMDP problem. It might be possible that for some problems the bounds are reached in special cases. We plan to expand our empirical evaluation to other POMDP benchmarks and real applications.

6 Conclusion

In this paper we addressed sensitivity analysis in POMDP value functions. We considered two types of perturbation, additive and multiplicative, in the parameters of a POMDP model. The error of the value function in the resulting perturbed models in both cases are bounded by a linear factor of the noise. We examined our theoretical analysis on an arbitrary POMDP problem and illustrated that the linearity presented in the bounds with respect to the value of the noise parameter, is achieved in our empirical evaluations. Future research in this area will be very important in the practicality of POMDP models in real applications.

References

[1] K. J. Astrom. Optimal control of markov decision processes with incomplete state estimation. *Mathematical Analysis and Applications*,volume 10, pp.174–205.

[2] Xi-Ren Cao (2003), From Perturbation Analysis to Markov Decision Processes and Reinforcement Learning. In *Journal of Discrete Event Dynamic Systems*.

[3] Xi-Ren Cao, and Han-Fu Chen (1997) Perturbation Realization, Potentials, and Sensitivity Analysis of Markov Processes. In *IEEE TRANSACTIONS ON AUTOMATIC CONTROL*, VOL. 42, pp.1382–1392.

[4] A. R. Cassandra, M. L. Littman, and L. P. Kaelbling (1994) Acting optimally in partially observable stochastic domains. In *Proceedings of the Twelfth National Conference on Artificial Intelligence (AAAI-94)*, volume 2, pages 1023-1028,

[5] R. Givan, T. Dean, and M. Greig (2003) Equivalence notions and model minimization in markov decision processes. In *Artificial Intelligence*, 147(1-2), pp.163-223.

[6] Z. Feng and E. Hansen (2004) An approach to state aggregation for POMDPs. In *Proceedings of the 2004 AAAI workshop on learning and planning in Markov processes - Advances and challenges*.

[7] M. Hauskrecht, and H. Fraser (1998) Modeling Treatment of Ischemic Heart Disease with Partially Observable Markov Decision Processes. In *Proceedings of American Medical Informatics Association annual symposium on Computer Applications in Health Care*.

[8] K.J. Hurt-Mullen , and J. Coberly (2005) Syndromic surveillance on the epidemiologist's desktop: making sense of much data. In *MMWR Morb Mortal Wkly Rep.*, Aug 26,54, pp.141–146.

[9] M.T. Izadi and D. Buckeridge (2007). Optimizing Anthrax Outbreak Detection Methods Using Reinforcement Learning. In *International Conference on Innovative Applications of Artificial Intelligence (IAAI)*.

[10] G.D. Konidaris, and A.G. Barto (2007) Building Portable Options: Skill Transfer in Reinforcement Learning. In *Proceedings of the Twentieth International Joint Conference on Artificial Intelligence (IJCAI-07)*.

[11] Chuanpu Hu, William S. Lovejoy, Steven L. Shafer (1996) Comparison of Some Suboptimal Control Policies in Medical Drug In *Operations Research*, Vol. 44, No. 5, pp. 696–709.

[12] A. Marcel, J. van Gerven, Francisco Babs, G. Taal, and Peter J.F. Lucas (2007) Selecting treatment strategies with dynamic limited-memory influence diagrams. In *Artif. Intell. Med.*,40(3)pp.171–186.

[13] N. B. Peek (1999). A specialized POMDP form and algorithm for clinical patient management Department of Information and Computing Sciences, Utrecht University, technical report.

[14] T. J. Perkins and D. Precup (1999) Using Options for Knowledge Transfer in Reinforcement Learning. Department of Information and Computing Sciences, University of Massachusetts,technical report.

[15] P. Poupart and C. Boutilier (2003). Value-directed Compression of POMDPs. In *Proceedings of NIPS'02*, pp. 1547–1554.

[16] N. Roy (2003) Finding approximate POMDP solutions through belief compression. PhD thesis, Robotics Institute, Carnegie Mellon University, Pittsburgh.

[17] A. J. Schaefer, M. D. Bailey, S. M. Shechter, and M. S. Roberts (2004) Modeling Medical Treatment using Markov Decision Processes. In *Handbook of Operations Research/Management Science Applications in Health Care, Kluwer Academic Publishers, M. Brandeau, F. Sainfort, and W. Pierskalla, eds* pp. 597–616.

[18] Michael Kearns and Satinder Singh (2002) Near-Optimal Reinforcement Learning in Polynomial Time by . In *Machine Learning journal*, Volume 49, Issue 2, pp. 209–232.

[19] Alexander A. Sherstov and Peter Stone (2005) Improving Action Selection in MDP's via Knowledge Transfer. In *Proceedings of the Twentieth National Conference on Artificial Intelligence*.