Motion Planning in Uncertain Environments with Vision-like Sensors

Suman Chakravorty [a,*] John L. Junkins [b]

[a] *Assistant Professor, Department of Aerospace Engineering, Texas A&M University, College Station, TX 77843-3141*

[b] *Distinguished Professor, Department of Aerospace Engineering, Texas A&M University, College Station, TX 77843-3141, Member NAE*

Abstract

In this work we present a methodology for intelligent path planning in an uncertain environment using vision like sensors, i.e., sensors that allow the sensing of the environment non-locally. Examples would include a mobile robot exploring an unknown terrain or a micro-UAV navigating in a cluttered urban environment. We show that the problem of path planning in an uncertain environment, under certain assumptions, can be posed as the adaptive optimal control of an uncertain Markov decision process, characterized by a known, control dependent system, and an unknown, control independent environment. The strategy for path planning then reduces to computing the control policy based on the current estimate of the environment, also known as the "certainty equivalence" principle in the adaptive control literature. Our methodology allows the inclusion of vision like sensors into the problem formulation, which empirical evidence suggests, accelerates the convergence of the planning algorithms. Further we show that the path planning and estimation problems, as formulated in this paper, possess special structure which can be exploited to significantly reduce the computational burden of the associated algorithms. We apply this methodology to the problem of path planning of a mobile rover in a completely unknown terrain.

Key words: Intelligent control, Vision sensors, Stochastic Adaptive Optimal Control, Dynamic Programming

1 Introduction

In this paper, we present a methodology for "intelligent path planning" in an uncertain environment using vision-like sensors. The state space of any path planning problem can be expressed as the ordered pair $(s, q(s))$ where s represents the system state and $q(s)$ represents the state of the environment at the state s. For example, in the case of a robot exploring an unknown terrain, s corresponds to the (x, y) co-ordinates of the robot and $q(s)$ corresponds to the height of the terrain $z(x, y)$ at the point (x, y). We show that the planning problem can be modeled as a Markov decision process, characterized by a known, control dependent exploration system and an unknown, uncontrollable environment part. Then, we show that the "intelligent path planning" paradigm reduces to the adaptive optimal control of a Markov decision process. Our formulation allows the integration of vision-based or similar sensors, i.e., sensors that allow the sensing of an environment non-locally, into the planning methodology. We show that the planning and estimation problems, as formulated in this paper, have special structure which can be exploited to significantly reduce the dimensionality of the associated algorithms. There has been substantial research in the adaptive control of controlled Markov Chains, or Markov Decision Processes, in the past two decades and it is usual to distinguish the different methods into "direct" and "indirect" adaptive control. In indirect adaptive control, the transition probabilities of the underlying Markov chain are estimated and the control is applied based on the most recent estimate of the transition probabilities

* Corresponding author.

Email addresses: schakrav@aero.tamu.edu (Suman Chakravorty), junkins@aero.tamu.edu (John L. Junkins).

(Kumar and Varaiya, 1986). This is known as the so-called "certainty equivalence principle". The "direct" approach to stochastic adaptive control falls under the category of "reinforcement learning" methodologies wherein the optimal control is calculated directly without resorting to estimating the transition probabilities of the underlying Markov chain (Bertsekas and Tsitsiklis, 1996; Sutton and Barto, 1998). These methods can be further distinguished into "Q-learning" and "adaptive-critics", please refer to (Bertsekas and Tsitsiklis, 1996; Sutton and Barto, 1998) for more details. Underlying all these methods is Bellman's "principle of optimality" or Dynamic Programming, a methodology for sequential decision-making under uncertainty (Bertsekas and Tsitsiklis, 1996). In this work, we show that the path planning problem can be reduced to the adaptive optimal control of a Markov decision process and thus, the above methodologies can be applied to the same.

Robot motion planning has been an active area of research over the past few decades. Various different approaches have been devised for the collision-free path planning of mobile robots in known environments (Latombe, 1991). The uncertainty in the environment is treated on a deterministic worst case (Latombe et. al, 1991; Mason, 1989) or in a probabilistic average case basis (Lavalle, 2000, 2006), for the case of planning under uncertainty. In "probabilistic robotics", there has been substantial research in the localization of a mobile robot while simultaneously mapping the environment (Thrun, 2001, 2000), the so called SLAM problem. However, the primary focus of this work has been the estimation of the robot position while simultaneously mapping the operating environment and not much attention has been paid to the control aspect of this problem. In (Lavalle, 2000), a game-theoretic framework is proposed for robotic motion planning wherein the uncertainties inherent in the robotic system and the environment are modeled probabilistically. Then, the authors resort to Bellman's principle of optimality (Bertsekas and Tsitsiklis, 1996) in order to tackle the motion-planning problem. However, even a probabilistic model of a robotic system and its environment suffers from uncertainties, i.e., the transition probabilities of the underlying Markov processes are unknown or uncertain. Thus, the problem of control of such a robotic system needs to take this uncertainty into account and as such can be posed as a stochastic adaptive control problem. In (Lavalle, 2000; Hu and Brady, 1997), a Bayesian adaptive control (Kumar and Varaiya, 1986) approach is suggested for the solution of this problem. However, the Bayesian approach suffers from the issue of dimensionality and is unsuitable for high dimensional environments. Under an "incoherent" assumption on the environment, which is typically satisfied by a large component of any environment, and corresponds to an environment that is "stationary" in a probabilistic sense. This allows us to break down the dimensionality of the associated estimation problem. We adopt a non-Bayesian adaptive control framework. Specifically, we propose a non-Bayesian parameter estimation approach, based on the Monte-Carlo method and non-local sensing using vision like sensors, to estimate the probabilities governing the environment process.

The original contributions of the current work are as follows. We identify that the problem of "intelligent path planning" can be reduced to the adaptive optimal control of an uncertain Markov decision process, characterized by a known, control dependent system and an unknown, control independent environment. Posing the problem in this fashion significantly reduces the computational burden of the estimation and the planning algorithms. We propose a Monte-Carlo based estimation scheme for the estimation of the probabilities governing the environment process using vision-like sensors and use the "certainty equivalence principle" to adaptively/intelligently change the control policies. We demonstrate the convergence of the planning and estimation algorithms. The rest of the paper is organized as follows. Section 2 contains the formulation of the exploration problem as a Markov decision problem and a heuristic analysis of the estimation and control schemes proposed to solve the decision problem. In section 3, we provide formal results regarding the convergence of the exploration policies under the planning methodology introduced in section 2. In section 4, we present a numerical example where a mobile rover navigates in an unknown terrain as an application of the proposed methodology.

2 Intelligent Exploration Systems

2.1 Preliminaries

Let the state of the exploration system be denoted by s, $s \in S$, where S denotes a finite state space. We denote the state of the environment at the system state s, i.e., the local state of the environment at system state s, by $q(s)$. From hereon we shall assume that the system state and the local environment are sensed perfectly and only the non-local environment is sensed imperfectly. For simplicity, we shall assume that the state space is finite. Let the number of system states be N and let the number of possible environment states, at any system state s, be D, and we denote this set by Q. We denote the local state of the exploration system by the ordered pair $(s, q(s))$. Let the set of control actions be denoted by U. We shall denote any particular control action by u. Let $\mathcal{F}^t = \{(s_0, q_0(s_0)), u_0, \cdots, (s_{t-1}, q_{t-1}(s_{t-1})), u_{t-1}\}$ represent the history of the process till time t, where s_t represents the system state at time t and $q_t(s_t)$ represents the state of the environment at state s_t at time t.

A **2.1** *We assume that the current system state, s_t, is dependent only on the system state and control input at the previous time instant, i.e.,*

$$p(s_t/\mathcal{F}^t) = p(s_t/s_{t-1}, u_{t-1}). \tag{1}$$

The environment is assumed to satisfy the "incoherence" assumption below.

A **2.2** *The environment process is "incoherent", i.e., the environment (random) process is spatially uncorrelated and temporally stationary. In other words, if* $\{q_t(s), s \in S\}$ *denotes the environment process,* $q_t(s)$ *is a stationary process for all* $s \in S$*, i.e.,* $q_t(s)$ *is independent of* $q_\tau(s)$ *for all* $t \neq \tau$*, for all* $s \in S$*, and the random variables are identically distributed. Moreover,* $q_t(s)$ *is independent of* $q_\tau(s')$ *whenever* $s \neq s'$*, for all* t, τ*.*

Then, the following result for the system transition probabilities is easily shown:

Proposition 1 *Under assumptions A2.1,A2.2, the following holds:*

$$p((s_t, q_t(s_t))/\mathcal{F}^t) = p(s_t/s_{t-1}, u_{t-1})p^*(q_t(s_t)). \quad (2)$$

Here, $p^*(q(s))$ represents the true probability of the local environment in state s being at the state $q(s)$.
The goal of path planning in an uncertain environment is framed as an infinite horizon discounted stochastic optimization problem, i.e., given any initial state $(s_0, q_0(s_0))$, find the optimal control sequence $\mu^*(s_0, q_0(s_0)) = \{u_0, u_1, ...\}$ corresponding to the optimal path such that

$$\mu^*(s_0, q_0(s_0)) = arg\min_{\mu} E_\mu(J/(s_0, q_0(s_0))),$$
$$J = \sum_{t=1}^{\infty} \beta^t c((s_t, q_t(s_t)), (s_{t-1}, q_{t-1}(s_{t-1})), u_{t-1}), \quad (3)$$

where $c((s_t, q_t(s_t)), (s_{t-1}, q_{t-1}(s_{t-1})), u_{t-1})$ is a positive pre-defined cost that the system incurs in making the transition from state $(s_{t-1}, q_{t-1}(s_{t-1}))$ to $(s_t, q_t(s_t))$ under the control action u_{t-1}, $E_\mu(.)$ denotes the expectation operator with respect to the policy μ, and $\beta < 1$ is a given discount factor. The optimal control policy is sationary and given by the optimality equation (Bertsekas and Tsitsiklis, 1996):

$$u^*(s, q(s)) = arg\min_{u} \sum_{(r,\bar{q}(r))} p(r/s, u)p^*(\bar{q}(r))[c((r, \bar{q}(r)), (s, q(s)), u) + \beta J^*(r, \bar{q}(r))], \quad (4)$$

where $J^*(r, \bar{q}(r))$ is the optimal cost-to-go from the state $(r, \bar{q}(r))$. Moreover, J^* satisfies the following fixed point equation:

$$J^*(s, q(s)) = \min_{u} \sum_{(r,\bar{q}(r))} p(r/s, u)p^*(\bar{q}(r))[c((r, \bar{q}(r)), (s, q(s)), u) + \beta J^*(r, \bar{q}(r))]. \quad (5)$$

We envisage the problem of path planning as one of adaptive control of an uncertain Markov decision process, (i.e., the transition probabilities of the Markov decision process are not known) since the environmental probabilities $p^*(q(s))$ are unknown and have to be estimated during the course of the exploration based upon the observation of the environment.
We adopt the following environment observation model: At every instant t, the system at state s_t, can observe the environment state $q_t(s)$, (i.e., the current environment state at the state s),if $s_t \in F(s) \subseteq S$, where $F(s)$ is assumed to be known beforehand. The set $F(s)$ constitutes a "footprint" of the sensor system.
Associated with every observation-vantage point pair, $(q(s), s')$, $s' \in F(s)$, there exists a known measurement error model, $p(\hat{q}(s)/q(s), s')$, $\hat{q}(s), q(s) \in Q$, and $s, s' \in S$, i.e., the probability that $\hat{q}(s)$ is observed when the environment is actually at the state $q(s)$, at system state s. The above observation model facilitates the inclusion of visual sensors, or any other sensor system that allows the observation of the environment non-locally, into the problem formulation.

2.2 Heuristic Analysis

2.2.1 Estimation

In the following, we shall present the estimation and control methodologies that we intend to use in order to solve the exploration problem. Consider the following relationship:

$$\pi(\hat{q}(s)) = \sum_{s' \in F(s), q(s)} p(\hat{q}(s)/q(s), s')p^*(q(s))\pi(s'/F(s)), \quad (6)$$

where $\pi(\hat{q}(s))$ denotes the probability of observing the noise corrupted environment state $\hat{q}(s)$ during the course of the exploration, i.e., the fraction of the time that the environment at state s is observed to be at $\hat{q}(s)$ during the course of the exploration, while in the footprint set $F(s)$, $p^*(q(s))$ denotes the true probability that the environment state is $q(s)$ at the state s, $\pi(s'/F(s))$ denotes the probability that the system is at state s', while in the footprint set $F(s)$, during the course of the exploration, i.e., the fraction of the time that the system is at state s' while in the footprint set $F(s)$.
Since the noise model is known, and the values of $\pi(\hat{q}(s))$ and $\pi(s)$ can be estimated during the course of the exploration using the Monte-Carlo method, it is possible to obtain the true environment probabilities using equation (6). Mathematically, we have:

$$\pi_t(\hat{q}(s)) := \frac{1}{T(F(s))} \sum_{k=1}^{T(F(s))} 1(\hat{q}_{t_k}(s) = \hat{q}(s)), \quad (7)$$

$$\pi_t(s) := \frac{1}{t} \sum_{n=1}^{t} 1(s_n = s), \quad (8)$$

$$\pi_t(F(s)) := \frac{1}{t}\sum_{n=1}^{t} 1(s_n \in F(s)), \quad (9)$$

where $1(A)$ denotes the indicator function of the event A, $T(F(s))$ is the total time spent by the system in the footprint set $F(s)$ until time t, and $\{t_k\}$ represents the set of time instants when the system is in the footprint set $F(s)$. Then, the true probabilities of the environment process $p(q(s))$ are obtained recursively as:

$$P_t(s) := arg\min_{P\in\mathcal{P}} ||\bar{b}_t(s) - \bar{A}_t(s)P||^2, \quad (10)$$

where

$$P_t(s) = [p_t(q_1(s)), \cdots, p_t(q_D(s))]', \quad (11)$$
$$\bar{b}_t(s) = [\pi_t(q_1(s)), \cdots \pi_t(q_D(s))]', \quad (12)$$
$$\bar{A}_t(s) = [\alpha_t^{ij}(s)], \quad (13)$$
$$\alpha_t^{ij}(s) = \frac{1}{\pi_t(F(s))}\sum_{s'\in F(s)} p(q_i(s)/q_j(s), s')\pi_t(s'), \quad (14)$$

$||.||$ denotes the euclidean norm in $\Re^D$, and $\mathcal{P}$ represents the space of all probability vectors in $\Re^D$.

2.2.2 Control

In this subsection, we shall study the control problem associated with the exploration problem. Consider the stochastic optimal problem posed in equation (3). Let $T = \{t_1, t_2, \cdots, t_k, \cdots\}$ denote the set of all times at which the control policy is updated during the path planning. Let the updated control policy at time instant t_k be denoted by $\mu_k(s, q(s))$. Let $p_t(q(s))$ denote the estimated environmental uncertainty at the time t, obtained from the estimation equation (10). Then, the control update at time $t_k \in T$, $\mu_k(.)$, using the principle of optimality and the "certainty equivalence principle", is given by

$$\mu_k(s, q(s)) = argmin_u \sum_{(r,\bar{q}(r))} p(r/s, u)p_{t_k}(\bar{q}(r))[c((r, \bar{q}(r)), (s, q(s)), u) + \beta J_k(r, \bar{q}(r))], \quad (15)$$

where

$$J_k(s, q(s)) = \min_u \sum_{(r,\bar{q}(r))} p(r/s, u)p_{t_k}(\bar{q}(r))[c((r, \bar{q}(r)), (s, q(s)), u) + \beta J_k(r, \bar{q}(r))]. \quad (16)$$

3 Adaptive/ Intelligent Path Planning

In this section, first we shall assume that the control policy for path planning is a given fixed stationary policy $\mu(s, q(s))$, and study the convergence of the estimation scheme specified by eq. (10). Let us fix some $s \in S$. Let the exploration system be in the state $(s_t, q_t(s_t))$ at time t and let the observed environment at s, at time t, be denoted as $q_t(s)$.

A **3.1** *We assume that* $\sum_{j\neq i} p(q_j(s)/q_i(s), s') \leq 0.5 - \epsilon$, $\epsilon > 0$.

The above assumption implies that we have a "good sensor", i.e., a sensor that is, on an average, right more times than its wrong.

Proposition 2 *If every environment state $q(s)$ is observed infinitely often, and under assumption* **A** *3.1, the estimates of the environmental probabilities, $p_t(q(s)) \rightarrow p(q(s))$, for all $q(s) \in Q, s \in S, w.p.1$.*

Proof: See Appendix.
The quantities $\pi_t(s)$ and $\pi_t(\hat{q}(s)$ need not converge for the convergence result above to hold. If we further assume that the Markov Chain underlying the evolution of the system is strongly ergodic, then given any initial distribution, the probability distribution of the states of the Markov chain approaches the unique steady state distribution of the Markov chain asymptotically. Please refer to (Isaacsen and Madsen, 1978) for more details. In fact, in that situation the quantities $\pi_t(s)$ and $\pi_t(\hat{q}(s)$ would then converge to these invariant probabilities.
Next, we list a few of the salient properties of the path planning/ control problem (i.e., the adaptive control policy) as presented in section 2. In the following, propositions 3 and 4 are presented without proof for lack of space. Please refer to [Chakravorty et. al. (2005)] for the proofs.

Proposition 3 *Under assumptions A2.1- A2.2, the cost-to-go functions $J_t(s, q(s)) \rightarrow J^*(s, q(s))$ as $t \rightarrow \infty$, if $p_t(q(s)) \rightarrow p^*(q(s))$, for all $s \in S, q(s) \in Q$.*

Next, we simplify the optimality equations based on the special structure of the path planning problem due to the incoherent environment assumption, which allows us to reduce the dimensionality of the problem.
Let

$$\sum_{(r,\bar{q}(r))} p((r, \bar{q}(r))/(s, q(s)), u)c((r, \bar{q}(r)), u, (s, q(s))) = \bar{c}(s, u, q(s)). \quad (17)$$

It follows that

$$J(s, q(s)) = \min_u[\bar{c}(s, u, q(s)) + \beta\sum_r p(r/s, u)\bar{J}(r)], \quad (18)$$

where

$$\bar{J}(s) = \sum_{q(s)} p^*(q(s))J(s, q(s)). \quad (19)$$

Noting that

$$u^*(s,q(s)) = arg\min_u[\bar{c}(s,u,q(s)) + \beta\sum_r p(r/s,u)\bar{J}(r)], \quad (20)$$

we can conclude that we need only have an average value of the cost-to-go function $J(s,q(s))$, at the system state s, in oredr to evaluate the optimal control. Note that $\bar{J}$ is the fixed point of the "average" dynamic programming operator $\bar{T}: \Re^N \rightarrow \Re^N$, defined by the following equation:

$$\bar{T}\bar{J}(s) = \sum_{q(s)} p^*(q(s)) \min_u\{\bar{c}(s,u,q(s)) + \beta\sum_r p(r/s,u)\bar{J}(r)\} \quad (21)$$

Proposition 4 *The average DP operator, $\bar{T}$, is a contraction mapping in $\Re^N$ under the ∞ norm.*

Next, we have the following result regarding the convergence of the path planning process:

Proposition 5 *If every environmental state $q(s)$ is observed infinitely often, and assumption* **A** *3.1 is satisfied, the estimates $p_t(q(s)) \rightarrow p^*(q(s))$, for all $s \in S, q(s) \in Q$, w.p. 1. Furthermore, $\bar{J}_t(s) \rightarrow \bar{J}^*(s)$, for all $s \in S$, w.p. 1.*

The above result holds not only for the control policy outlined in this paper, but any policy that is non-anticipative, i.e., the control at the current instant only depends on the past of the algorithm and not the future. The above result implies that the estimates of any environment state that is observed infinitely often converges to the true values. It is entirely possible that some of the environment states will not be observed infinitely often, and thus nothing can then be said about the convergence of the estimates of such states, which is a manifestation of the exploration-exploitiation tradeoff.

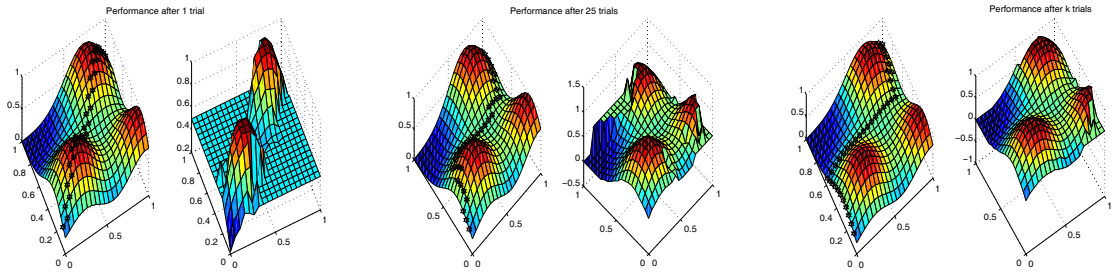
4 Illustrative Example: A Mobile Rover Navigating an Unknown Unstructured Terrain

In this section, we apply the methodology developed so far in this work to the problem of path planning of a mobile robotic rover exploring an unknown terrain. In this example the system state, s, is the (x,y) grid point and the environment variable q is the height of the terrain, z, at the grid point. We discretized the (x,y) plane into a 20x20 grid. The height of the terrain was discretized into 10 equispaced intervals. There were four allowable control actions, $N/E/S/W$, which corresponded to the rover going North/ East/ South/ West to the adjacent grid point. We assumed that there was no uncertainty in control or localization. We assumed that we had no prior knowledge of the terrain that was to be negotiated. Thus, the initial environment uncertainty, $p_0(q(s))$, was modeled as a uniform distribution. For the noise model $p(\hat{q}(s')/q(s'),s)$, we assumed that the sensors observed the environment at state s' only if the system was at states s lying in the region $||s-s'|| \leq R$, i.e., $F(s) = \{s' : ||s-s'|| \leq R\}$, where $R = 0.1$. The variance of the measurement noise grows monotonically, in a linear fashion, with distance of the observed point from the vantage point/ sensor, and also, the height of the terrain at the observed point. The noise was assumed to be Gaussian. The incremental cost function $c((r,\bar{q}(r)),(s,q(s)),u)$ penalized path length, i.e., $||(r,\bar{q}(r)) - (s,q(s))||$, and also, the distance to the goal point $(1,1)$,i.e., $||(r,\bar{q}(r)) - (1,1)||$. It was found through trial and error that penalties in the ratio 10:1 led to satisfactory performance of the rover. The control policies were updated after every trip to the goal point. The initial state was randomly varied to ensure that the terrain was explored sufficiently.

The simulation results are shown in Fig.1- Fig.2. In each subfigure of figure 1, the left subplot in each subfigure shows the actual terrain and the path of the rover along the terrain according to the current estimate of the terrain (which is shown in the right plot). Fig.1(a)-Fig.1(c) represent the progress of the algorithm when vision-like sensors are used. As can be seen from the plots, the rover is able to find a near optimal route after around 25 trials . We also tested the algorithms on two other topologies in order to study the convergence of the environmental probabilities and the cost-to-go vector. It was seen that the performance of the algorithm was quite similar in all three cases. In Fig. 2(a), a plot of the convergence of the environmental probabilities for the three different cases is shown. The environmental probabilities can be considered to be a matrix of dimension N x D, and the Frobenius norm was used as a measure of the convergence in this case. In Fig. 2(b), a similar plot for the cost-to-go vectors (Euclidean norm) is shown. It can be seen from the plots that both the environmental probabilities and the cost-to-go vector converge to their optimal values in about 100 iterations though the convergence of the cost-to-go vector is considerably smoother than that of the environmental probabilities. Further evidence of algorithmic success of this methodology can be found in [Davis and Chakravorty (2006)].

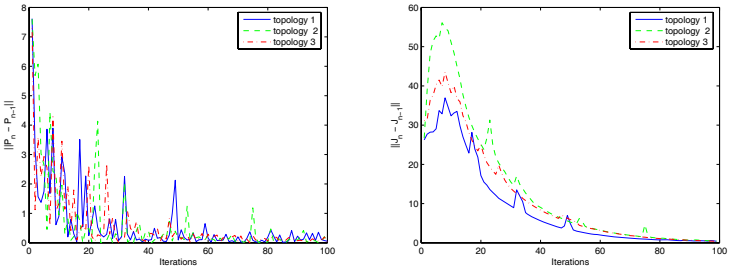
Thus, in this work, we have presented a methodology for intelligent exploration of an unknown environment with vision like sensors. We have used the "certainty-equivalence" principle in order to adaptively/intelligently update the control policies for path planning using a Monte-Carlo based estimation scheme which recursively estimates the probabilities of the environment process, based on non-local observations of the environment by vision-like sensors and proved the convergence of the scheme under certain assumptions.

The algorithms have to be extended to the case of imperfect state observations. This is an avenue of research that we are currently pursuing. The use of approximate DP methods, i.e., DP with functional approximation



(a) Performance after 1 trial (b) Performance after 25 trials (c) Performance after 50 trials

Fig. 1. Progressive learning of terrain by mobile rover



(a) Convergence of environmental probabilities (b) Convergence of cost-to-go vector

Fig. 2. Convergence of terrain and cost-to-go estimates

can be pursued to make the methodology applicable to real systems, where it can act as a high level motion planner. Local robust controllers can be designed for tracking the high level path plans and the integration of the two can result in truly intelligent autonomous systems.

A Appendix

Let $P_t(s)$ denote the vector of estimated environmental probabilities at state s at time t, i.e., $P_t(s) = [p_t(q_1(s)), \cdots, p_t(q_D(s))]^T$. In the following, we fix $s \in S$ and drop the explicit reference to s in the environmental probabilities. We show that the convergence results hold for any $s \in S$ and thus, because S is a finite set, the results follow for the entire set. In the following, for notational convenience, we shall assume that the whole environment is visible from every vantage point. The proofs hold for the more general case with a few minor adjustments.
It can be shown that eq. (10) may be written as (denoting $\bar{A}_t(s), \bar{b}_t(s)$ as $\bar{A}(t), \bar{b}(t)$ respectively by dropping the explicit reference to s)

$$P_t = arg \min_{p \in \mathcal{P}} ||\bar{A}(t)P - \bar{b}(t)||, \tag{A.1}$$

where $\bar{A}(t)$ and $\bar{b}(t)$ are found recursively as

$$\bar{A}(t) = (1 - \gamma_t)\bar{A}(t-1) + \gamma_t A(X_t), \tag{A.2}$$

$$\bar{b}(t) = (1 - \gamma_t)\bar{b}(t-1) + \gamma_t b(X_t), \tag{A.3}$$

where $X_t = (s_t, \hat{q}_t(s)) = (s_t, \hat{q}_t)$ (dropping the explicit reference to s), $A(X_t) = [p_{ij}(s_t)]$ where $p_{ij}(s_t) = p(\hat{q} = q_i / q = q_j, s_t)$, $b(X_t) = [\beta_i(X_t)], \beta_i(X_t) = 1(\hat{q}_t = q_i)$, and $\gamma_t = \frac{1}{t}$. Note that $\sum \gamma_t = \infty$ and $\sum \gamma_t^2 < \infty$. In fact, the sequence $\{\gamma_t\}$ can be any deterministic sequence that satisfies these properties. The above recursions have zero initial conditions.
Let $\mathcal{F}^t = \{X_0, \hat{q}_0(s_0), u_0 \cdots, X_t, \hat{q}_t(s_t)\}$ denote the history of the algorithm till the time t, with all the random variables defined on the common probability space $(\Omega, \mathcal{F}, P)$. Note that $\mathcal{F}_t \subset \mathcal{F}_{t+1}$ for all t. Let u_t denote the control at time instant t.

Lemma A.1 *The conditional expectation, $E[A(X_{t+1})P^* - b(X_{t+1})/\mathcal{F}^t] = 0$, $\forall t$, where P^* represents the vector of true environmental probabilities.*

Proof:

$$E[A(X_{t+1})P^* - b(X_{t+1})/\mathcal{F}^t] = E[A(X_{t+1})/\mathcal{F}^t]P^* - E[b(X_{t+1})/\mathcal{F}^t]. \tag{A.4}$$

Then, $E[A(X_{t+1})/\mathcal{F}^t] = \sum_{X,u} A(X) p_{XX_t}(u) p(u/\mathcal{F}^t), X = (s, \hat{q})$, where $p_{XX_t}(u) = p(s/s_t, u)p^*(\hat{q})$ and let $p^*(\hat{q}) \equiv p^*_{\hat{q}}$, $p(s/s_t, u) \equiv p_{ss_t}(u)$. Note that $p(u/\mathcal{F}^t)$ is well-defined since the control is non-anticpative. As

mentioned previously, this is the only property that is required of the control policy in the following. Thus,

$$E[A(X_{t+1})/\mathcal{F}^t] = \sum_{s,j,u} A(s)p_{ss_t}(u)p^*_{q_j}p(u/\mathcal{F}^t)$$
$$= \sum_{s,u} A(s)p_{ss_t}(u)p(u/\mathcal{F}^t) = [\sum_{s,u} p_{ij}(s)p_{ss_t}(u)p(u/\mathcal{F}^t)]. \tag{A.5}$$

Hence,

$$E[A(X_{t+1})/\mathcal{F}^t]P^* = [\sum_{s,j,u} p_{ij}(s)p^*_{q_j}p_{ss_t}(u)p(u/\mathcal{F}^t)]. \tag{A.6}$$

We have $E[b(X_{t+1})/\mathcal{F}^t] = [E[\beta_i(X_t)/\mathcal{F}^t]]$. Thus,

$$E[\beta_i(X_t)/\mathcal{F}^t] = E[1(\hat{q}_t = q_i)/\mathcal{F}^t)$$
$$= p(\hat{q}_t = q_i/\mathcal{F}^t) = \sum_{j,s} p(q_i/q_j, s)p^*_{q_j}p(s/\mathcal{F}^t)$$
$$= \sum_{s,j,u} p_{ij}(s)p^*_{q_j}p_{ss_t}(u)p(u/\mathcal{F}^t). \tag{A.7}$$

Comparing the above equation with eq. (A.6), the result follows.
□

Lemma A.2 *Under assumption A 3.1, $||\bar{A}^{-1}(t)||_1 \leq K < \infty$, for all t, where $||.||_1$ represents the matrix norm induced by the l_1 norm in Euclidean space.*

Proof: Note that $\bar{A}(t) = [\alpha^{ij}_t(s)]$ (recall eq. (14)). Dropping the explicit reference to s, and defining $\alpha^{ij}_t(s) \equiv \alpha_{ij}(t)$, we have $\alpha_{ij}(t) = \sum_s p_{ij}(s)\pi_t(s)$, and since $[\pi_t(s)]$ is a probability vector, it follows that $\bar{A}(t)$ is a stochastic matrix, i.e., $\sum_i \alpha_{ij}(t) = 1$. Under assumption A 3.1, $\alpha_{jj}(t) = \sum_s p_{jj}(s)\pi_t(s) \geq 0.5 + \epsilon, \forall\, j, t,$, and $\sum_{i\neq j} \alpha_{ij}(t) \leq 0.5 - \epsilon, \forall\, j, t$. Then, it follows that $(1 - \alpha_{jj}(t)) + \sum_{i\neq j} \alpha_{ij}(t) \leq 1 - 2\epsilon, \forall\, j, t$. Hence, we have $||I - \bar{A}(t)||_1 \leq 1 - 2\epsilon < 1, \forall\, t$. Thus, due to Corollary 5.6.16 [Horn and Johnson, ch. 4, p 344 (1993)], $\bar{A}^{-1}(t) = \sum_{k=0}^{\infty}(I - \bar{A}(t))^k \,\forall t$, and thus, $||\bar{A}^{-1}(t)||_1 \leq \sum_{k=0}^{\infty} ||I - \bar{A}(t)||_1^k \leq \frac{1}{2\epsilon} < \infty$. Hence the result follows. □

Lemma A.3 *The random vector $(\bar{A}(t)P^* - \bar{b}(t)) \to 0$ w. p. 1.*

Proof: Consider the stochastic Lyapunov function $V_t = ||\bar{b}(t) - \bar{A}(t)P^*||^2$ where $||.||$ denotes the Euclidean norm. Then

$$V_{t+1} = ||(1 - \gamma_{t+1})(\bar{b}(t) - \bar{A}(t)P^*)$$
$$+\gamma_{t+1}(b(X_{t+1}) - A(X_{t+1})P^*)||^2$$
$$= V_t + 2\gamma_{t+1} *$$
$$(\bar{b}(t) - \bar{A}(t)P^*)^T(b(X_{t+1}) - A(X_{t+1})P^* - \bar{b}(t) + \bar{A}(t)P^*)$$
$$+\gamma^2_{t+1}||b(X_{t+1}) - A(X_{t+1})P^* - \bar{b}(t) + \bar{A}(t)P^*)||^2. \tag{A.8}$$

Noting that $||A(X_t)|| \leq K < \infty$, $||b(X_t)|| \leq K < \infty$, it follows that $||\bar{A}(t)|| \leq K < \infty$, $||\bar{b}(t)|| \leq K < \infty$, uniformly for all t, and since $||P^*|| < 1$, it follows then $||b(X_{t+1}) - A(X_{t+1})P^* - \bar{b}(t) + \bar{A}(t)P^*)||^2 \leq C < \infty$. Using eq. (A.8) and the above observation, we have

$$E[V_{t+1}/\mathcal{F}^t] \leq V_t + C\gamma^2_{t+1} + 2\gamma_{t+1}(\bar{b}(t) - \bar{A}(t)P^*)^T$$
$$E[b(X_{t+1}) - A(X_{t+1})P^* - \bar{b}(t) + \bar{A}(t)P^*)/\mathcal{F}^t]. \tag{A.9}$$

Note that due to Lemma A.1, we have $E[b(X_{t+1}) - A(X_{t+1})P^*/\mathcal{F}^t] = 0$, and hence, it follows from eq. (A.9) that

$$E[V_{t+1}/\mathcal{F}^t] \leq V_t(1 - 2\gamma_{t+1}) + C\gamma^2_{t+1}. \tag{A.10}$$

Noting that $\sum_{t=0}^{\infty} \gamma^2_t < \infty$, $V_t \geq 0$, $C\gamma^2_t \geq 0$, it follows from the Supermartingale convergence theorem [Bertsekas and Tsitsiklis (1996); Neveu (1975)] that V_t converges w.p. 1 and that $\sum_{t=0}^{\infty} \gamma_{t+1}V_t < \infty, w.p.1$. Noting that $\sum_{t=0}^{\infty} \gamma_t = \infty$, it follows that $V_t \to 0$ w.p. 1. Hence $\bar{A}(t)P^* - \bar{b}(t) \to 0$ w.p. 1.
□

Both a fixed control policy and the control policy in Proposition 5 are examples of non-anticipative policies and thus, the following proof holds for both cases.

Proof of Proposition 5: From Lemma A.3, we have that $\bar{A}(t)P^* - \bar{b}(t) \to 0$, $w.p.1$. Thus for every $\omega \notin \mathcal{N}$, where $\omega \in \Omega$, $(\Omega, \mathcal{F}, P)$ being the probability space underlying the algorithm, $\mathcal{N}$ being a set of probability zero in $\mathcal{F}$, $||\bar{A}_t(\omega)P^* - \bar{b}_t(\omega)|| \to 0$, where $\bar{A}_t(\omega) \equiv \bar{A}(t, \omega)$ and $\bar{b}_t(\omega) \equiv \bar{b}(t, \omega)$. Since

$$P_t(\omega) = arg \min_{P' \in \mathcal{P}} ||\bar{A}_t(\omega)P' - \bar{b}_t(\omega)||,$$
$$||\bar{A}_t(\omega)P_t(\omega) - \bar{b}_t(\omega)|| \leq ||\bar{A}_t(\omega)P^* - \bar{b}_t(\omega)||, \tag{A.11}$$

which implies that $\lim_{t\to\infty} ||\bar{A}_t(\omega)P_t - \bar{b}_t(\omega)|| = 0$. Note that due to Lemma A.2, $\bar{A}_t^{-1}$ exists for all t, and thus,

$$||P_t(\omega) - P^*(\omega)||_1 \leq ||\bar{A}_t^{-1}(\omega)||_1(||\bar{A}_t(\omega)P_t - \bar{b}_t(\omega)||_1 +$$
$$||\bar{A}_t(\omega)P^* - \bar{b}_t(\omega)||_1).$$

Since $\bar{A}_t(\omega)P_t(\omega) - \bar{b}_t(\omega) \to 0$ and $\bar{A}_t(\omega)P^* - \bar{b}_t(\omega) \to 0$ there exists $N(\omega) < \infty$ such that for all $t \geq N(\omega)$, $||\bar{A}_t(\omega)P_t(\omega) - \bar{b}_t(\omega)||_1 \leq \frac{\epsilon}{2K}$ and $||\bar{A}_t(\omega)P^*(\omega) - \bar{b}_t(\omega)||_1 \leq \frac{\epsilon}{2K}$, where $||\bar{A}_t^{-1}(\omega)||_1 \leq K < \infty$ (from

Lemma A.2), and thus, $||P_t(\omega) - P^*(\omega)||_1 \leq K(\frac{\epsilon}{2K} + \frac{\epsilon}{2K}) = \epsilon$. Noting that the above holds for every sample $\omega \notin \mathcal{N}$ (a set of probability zero), it follows that the sequence $P_t \to P^*$ w.p. 1. That $J_t \to J^*$ w.p. 1 follows from Proposition 3.
□

References

P. R. Kumar & P. Varaiya, *Stochastic Systems: Estimation, Identification and AdaptiveControl,* Englewood Cliffs, NJ: Prentice Hall, 1986.

D. P. Bertsekas & J. N. Tsitsiklis, *Neuro-Dynamic Programming,* Belmont, MA: Athena, 1996.

R. S. Sutton & A. G. Barto, *Reinforcement Learning: An Introduction*, Cambridge, MA: MIT Press, 1998.

J. C. Latombe, "Robot Motion Planning," Kluwer, Boston, MA, 1991.

J. C. Latombe, A. Lazanas & S. Shekhar, "Robot Motion Planning with Uncertainty in Control and Sensing", *Artificial Intelligence*, 52:1-47, 1991.

M. T. Mason, "Automatic Planning of Fine Motions: correctness and Completeness", *Proc. of IEEE Conf. Robotics. Automat.*, pp 484-489, 1989.

S. Thrun, "A Probabilistic On-Line Mapping Algorithm for Teams of Mobile Robots," *The International Journal of Robotic Research*, vol. 20, no. 5, May 2001, pp. 335-363.

S. Thrun, "Probabilistic Algorithms in Robotics," *AI Magazine*, 21(4): 93-109

S. M. Lavalle, "Robot Motion Planning: A Game-Theoretic Foundation", *Algorithmica*, vol. 26, pp. 430-465, 2000.

S. M. Lavalle, *Planning Algorithms*, Cambridge University Press, Cambridge, UK, 2006.

H. Hu & M. Brady, "Dynamic Global Path Planning with Uncertainty for Mobile Robots in Manufacturing," *IEEE Transactions on Robotics and Automation,* vol. 13, no. 5, pp. 760-767, October 1997.

D. L. Isaacsen & R. W. Madsen, *Markov Chains Theory and Applications*, Wiley Series in probability and Mathematical Statistics, 1978

L. P. Kaelbling, M. L. Littman & A. R. Cassandra, "Planning and Acting in Partially Observable Stochastic Domains", *Artificial Intelligence*, vol. 101, 1998, pp. 99-134

S. Chakravorty and J. L. Junkins, "A Methodology for Intelligent Path Planning", *Proceedings of the 2005 IEEE Int. Symp. Intell. Cont.*, pp. 592-597, 2005.

J. Davis and S. Chakravorty, "Motion Planning in Uncertain Environemnts: Application to an Unmanned Helicopter", to appear in *Proceedings of the 2006 IEEE Conference on Decision and Control (CDC 06)*, San Diego, CA, 15-18 Dec., 2006.

R. A. Horn & Charles R. Johnson, *Matrix Analysis*, Cambridge University Press, New york, NY, 1993

J. Neveu, *Discrete Parameter Martingales*, North-Holland, Amsterdam, 1975.