Abstract—In this paper, the problem of mapping and planning in an uncertain environment is studied. A non-Bayesian formulation of the simultaneous planning, localization and mapping (SPLAM) problem is presented wherein the environment is modeled as a stationary, spatially uncorrelated random process whose stationary probabilities are fixed but unknown, and have to be estimated as the autonomous system moves through the environment and makes observations using its sensors. Two separate cases are considered for the knowledge of the location of the autonomous system: first, it is assumed that the system knows its location perfectly and in the second part, this assumption is relaxed and the system has to localize itself with respect to the map that it is building. The environmental random process is estimated using stochastic approximation algorithms and a control-theoretic, “certainty-equivalence” based approach is taken for the planning problem. Under a certain “reliable sensor assumption”, it is shown that the estimation and planning algorithms converge with probability one, and that the convergence of the mapping algorithms is independent of the control policy, as long as it is non-anticipative, akin to the celebrated “Separation Principle” in Classical Linear Control theory. Further, at least theoretically, the computational burden of the mapping and planning algorithms in this methodology is significantly less than the current Bayesian SPLAM techniques.

I. INTRODUCTION

In this paper, the problem of mapping and planning in an uncertain environment is considered. The uncertainty of the autonomous system is modeled as a completely known Markov decision process while the environment is modeled as a stationary, spatially uncorrelated (independent) random process whose parameters are unknown and have to be estimated while the system navigates through the environment. Moreover, the autonomous system has to navigate through the environment in an “optimal fashion”. It is assumed that there is error in observing the environment. The observation model is assumed to be known, given the location of the autonomous system. Two separate cases are considered for the sensing of the location of the autonomous system: in the first case, it is assumed that the location of the system is perfectly known while in the second, this assumption is done away with and the system has to localize itself with respect to some landmarks in the environment (or with some absolute position sensor such as GPS, which is shown to be a special case of the former). A stochastic approximation based approach is presented for the estimation of the environment process while a control-theoretic, “certainty-equivalence” based approach is taken to the solution of the resultant adaptive, partially observed Markov decision problem (POMDP). It is shown that under a certain “Reliable Sensor Assumption” the estimation and planning algorithms are guaranteed to converge with probability one. Moreover, it is shown that the convergence of the estimation algorithms are completely independent of the planning algorithms and thus, satisfy a “Separation Principle” as in classical linear control theory. It is further shown that the computation burden of the algorithms are significantly less, at least theoretically, than the current simultaneous planning, localization and mapping (SPLAM) techniques.

The problem of simultaneous localization and mapping (SLAM) or additionally simultaneous planning, localization and mapping (SPLAM) has received considerable attention in the Robotics community in the past several years. The generic SLAM problem consists of an autonomous system navigating in an unknown environment, which it is trying to map while simultaneously localizing itself with respect to the map that it is building. This creates a chicken and egg problem which leads to a very high dimensional estimation problem [1]–[4]. In most SLAM techniques, the localization and mapping problem is posed as a Bayesian filtering problem wherein the environment is considered to be a fixed but unknown parameter. There are two basic approaches to solving the Bayes filtering problem. The first alternative is to use the Kalman filtering technique which is applicable to linear-Gaussian systems [4]–[7]. However, this method cannot accommodate cases where the distributions are non-Gaussian and cannot provide a solution to the so called “data association” problem [2], [8], [9]. The second method consists of solving the Bayesian filtering problem using particle filtering techniques [10], [11]. These methods can accommodate the multi-modal nature of the probability distributions and the data association problem gracefully [2], [12]. The basic drawback with the Bayesian formulation of the SLAM problem is that the estimates of the various environmental components (features) become correlated even though their measurements are mutually independent. This is a basic structural property of the Bayesian formulation of the SLAM problem [1]–[4], [12]. Thus, the filter has to estimate a very high dimensional probability distribution on the environment which becomes increasingly computationally intractable as the size of the environment increases [2], [3], [12], [13]. In some approaches such as those in [3], [10], [12], [14], [15], the correlations in the environment disappear by conditioning on the trajectory of the autonomous system instead of the current location, the so called “Rao-Blackwellized” particle filtering approach. However, in this approach, the curse of dimensionality (in the environment) is traded off for the curse of history (in the trajectory of the autonomous system), and is still computationally very hard [14], [15]. The addition of planning to the SLAM problem, resulting in the SPLAM problem, adds further to the complexity of the problem [16]–[19]. In fact, the planning problem on its own is computationally quite intractable under uncertainty (more on this is in the following paragraph).
In this paper, an alternative to the Bayesian formulation of the SPLAM problem is proposed. In this formulation, the environment is modeled as a stationary but unknown random process that has to be estimated as the autonomous system moves through it and makes observations of the environment. It is assumed that the robot localizes in the environment based on a few landmarks located throughout the environment (which is the Bayesian SLAM problem but with very few parameters compared to case when the whole environment is considered to be an unknown parameter), and then maps the rest of the environment based on this estimate. This problem formulation ensures that the estimates of the landmarks and the environmental components never get correlated. In addition, each individual environmental component can be estimated completely independently of the other components. This drastically reduces the computational burden of the SPLAM problem as the correlation of the various features in the environment need not be considered. The environmental components are estimated using stochastic approximation algorithms. Further, the interaction between the planning and estimation algorithms is considered, and it is shown that the estimation and planning algorithms converge with probability one.

Planning under uncertainty has been a topic of interest to multiple communities such as Artificial Intelligence [20], [21], Control Theory [22]–[25] and Operations Research [26]. The motion planning of an autonomous agent in an uncertain environment is an application of the the methods developed in this research area. The most popular method of planning under uncertainty is through the use of Markov Decision problems (MDP) [22], [23], [27] or when the state of the system cannot be sensed perfectly, through partially observed Markov decision problem (POMDPs) [20], [21], [26]. However, the uncertainty models of the MDPs themselves might not be known perfectly and may depend on certain unknown parameters, which have to be estimated based on observations of the process as it evolves over time. This corresponds to the so-called stochastic adaptive control problem [24], [25], or an adaptive MDP, and in the absence of perfect knowledge of the state of the system, the problem is transformed into an adaptive POMDP. It can be seen that the problem of SPLAM is in fact an adaptive POMDP problem since the environment is unknown while at the same time the location of the autonomous systems cannot be sensed perfectly. Various researchers have applied MDPs/ POMDPs to the solution of the motion planning problem [28], [29], the second reference, in particular, has extensive references to Robotics related applications. The problem of simultaneous mapping and planning is relatively new in the Robotics community even though some researchers have considered this problem before [16]–[19]. The computational burden of solving MDPs gets very unmanageable as the size of the state space becomes large while that of POMDPs is almost intractable even for relatively small state spaces. Since the environment is a part of the state space in the motion planning problem, and because they are typically of large dimension in most realistic problems, the resultant POMDP is computationally intractable though some recent research has focused on highly efficient ways of solving these problems [13], [30]–[32]. Hence, the SPLAM problems are even more computationally intractable when compared to the pure planning problem because of the added burden of estimation (mapping) that comes along with the planning problem in this case. Moreover, the interaction of the mapping and planning algorithms need to be stable. In this paper, it is shown that the problem of planning can be completely “separated” from the estimation problem in the context of the SPLAM problem, as long as the control is non-anticipative, i.e., it does not depend on the future of the algorithm. This is akin to the celebrated “separation principle” in classical linear Control theory [33]. Further, the planning policies are based on a “certainty-equivalence” approach to the problem of solving POMDPs, instead of solving an information space or belief state MDP [20]–[22], leading to a computationally more tractable approach. In addition the structure of the problem affords certain simplifications of the underlying dynamic programming (DP) algorithms used to solve the MDPs, which can further reduce the computational burden of the planning algorithms.

The rest of the paper is organized as follows. In section II, certain basic facts about MDPs and POMDPs are briefly enumerated. Section III concerns itself with the modeling and formulation of the SPLAM problem. In section IV, the problem of mapping and planning, when the location of the autonomous system is known perfectly, is considered. Mapping and control algorithms are proposed and their strong consistency is established. In part II of the paper, the full SPLAM problem is considered. In section IV, mapping, localization and planning algorithms are proposed for the problem and their strong consistency is established. In section V, a detailed discussion on the relationship of the work presented in this paper to previous research is undertaken. In the appendix, the convergence results in section III and IV are proved.

II. PRELIMINARIES

In this section, we provide a brief overview of Markov decision processes (MDP) and partially observed Markov decision processes (POMDP). More comprehensive treatments of these can be found in [20]–[22], [27], [29]. We shall only consider discrete-time finite state MDPs here. Let \( s \) denote the state of a finite MDP, \( s \in S \), \( S \) being a finite set. Let \( u \) denote the control action which can take a discrete number of values. The MDP is characterized by a transition probability function, \( p(r/s, u) \), which is the probability that the system will transition from state \( s \) to \( r \) under control \( u \), at the end of one time step. The goal of the MDP is to solve the infinite horizon discounted optimal control problem given by

\[
\mu^*(s_0) = \arg \min_{\mu \in \{u_0, u_1, \ldots\}} \{ E \sum_{t=0}^{\infty} \beta^t c(s_t, u_t)/s_0 \},
\]

(1)
where \( \mu = \{ u_0, u_1, \ldots \} \) is an infinite horizon control policy, \( c(s, u) \) is the cost that the system incurs in taking control action \( u \) at state \( s \) and \( \beta < 1 \) is a given discount factor. It is well known that the optimal control policy corresponding to the problem posed above is stationary and is given by [22], [27]

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

(2)

where \( J^*(s) \) is the optimal cost-to-go function and is found as the solution of the Bellman fixed point equation/ Dynamic Programming equation:

\[
 J^*(s) = \arg \min_u \{ c(s, u) + \beta \sum_r p(r/s, u)J^*(r) \}.
\]

(3)

It is very well known that the (Bellman) operator underlying the above equation is a contraction operator with contraction factor \( \beta \) and thus, the solution can be found through successive approximations [22], [27], [34]. This is the basic principle behind value and policy iteration, two methods to solve the DP equation above [22], [23], [34]. This is the basic principle behind value and policy iteration, two methods to solve the DP equation above [22], [23], [34].

In some cases, the true state of the system might not be observable and only a measurement, \( m \) of the same might be available. The measurement is a known probabilistic function given the true state and is denoted by \( p(m/s) \). This problem seems non-Markovian to the observer since the observation at the current instant \( m_t \) seems correlated to the past \( F_t = \{ m_0, m_1, \ldots, m_{t-1} \} \). Thus, the problem is known as a partially observed Markov decision process (POMDP). In order to solve the above POMDP, it is transformed into an MDP though the introduction of a sufficient statistic, i.e., a quantity that is Markovian in nature. Such a quantity is the belief state or the information state which is the probability distribution over the states of the system given the observed past, \( p(s/F_t) \), and is denoted by the belief state \( b_t(s) \). It can be seen that the belief state evolves according to an MDP. The resultant MDP is called the belief state MDP or the information space MDP citebertsekas1, lavalle2, white. The various parameters of the MDP are found as follows: The incremental cost in taking control \( u \) at belief state \( b \) is given by

\[
 \bar{c}(b, u) = \sum_s c(s, u)b(s).
\]

(4)

The transition probability density function is found as follows:

\[
 p(b/b', u) = \sum_m p(b', m, b', u)p(m/b', u),
\]

(5)

\[
 p(m/b', u) = \sum_{s,s'} p(m/s)p(s/s', u)b'(s'),
\]

(6)

and the density function \( p(b', m, b', u) \) is formed as follows.

\[
 p(b/m, b', u) = 1 \text{ if } b = b', \quad 0, \text{ otherwise},
\]

(7)

and the belief \( \bar{b}(s) \) is formed using the so-called Bayes filter,

\[
 \bar{b}(s) = \eta p(m/s) \sum_{s'} p(s/s', u)b'(s'),
\]

(8)

where \( \eta \) is a suitable normalizer and is equal to \( p(m/b', u) \). Thus, the belief state MDP is completely defined and the methods for the solution of MDPs, i.e., value and policy iteration can be used to solve it. However, note that the belief state is a probability distribution on the states of the underlying MDP and thus, can take a continuum of values. Hence, the resultant belief MDP is a continuous state MDP which becomes very hard to solve for anything but a moderate number of states, though very efficient ways of solving such POMDPs have been devised [20], [21], [31].

### III. Modeling and Problem Formulation

In this section, the problem of Mapping and Planning under Uncertainty is modeled and formulated. The system consists of the robotic (autonomous) system and the environment. Let \( s \) denote the state of the robotic system, in the sequel it is referred to simply as the system state, and let \( Q = \{ q_1, q_2, q_3, \ldots, q_M \} \) represent the state of the environment, where \( q_k \) corresponds to the state of the \( k \)th environmental component. Also, it is assumed that \( s \) can take one of \( N \) values while each environmental component can take one of \( D \) values, i.e., \( s \in S = \{1, 2, \ldots, N\} \) and \( q_k \in E = \{1, 2, \ldots, D\} \). Thus, the state of the entire system (autonomous system state + environment) can be represented by the pair \( (s, Q) \).

#### A. The Dynamical Model

The dynamics of the system is modeled as an MDP. The transition probabilities of the system-environment pair is obtained as follows:

\[
 p((\bar{s}, \bar{Q})/(s, Q), u) = p(\bar{s}/s, u)p^*(Q),
\]

(9)

where

\[
 p^*(Q) = \prod_{i=1}^{M} p^*(q_i),
\]

(10)

where \( u \) represents the control acting on the system, and it is assumed that there are finitely many control actions. The above dynamical model assumes that the environment is unaffected by the motion of the system and that the (vector valued) environmental random process is stationary, and its components are independent of each other (in other words, the environment is temporally stationary and spatially uncorrelated/ independent). It can be seen that an overwhelmingly large component of any environment can be modeled in this fashion and deterministic environments are a special case of such an environment. Thus, such an environment is stationary in a probabilistic sense. This paper deals with the problem of mapping and planning in such a stationary environment.

#### B. The Planning Problem

The planning problem is framed as a discounted, infinite horizon, stochastic optimal control problem. Let the incremental cost function for moving from state \( (s, Q) \) to state
\(\ddot{s}(Q)\) under control \(u\) be denoted as \(c((s, Q), (\ddot{s}, Q), u)\). Let the discount factor be denoted by \(\beta < 1\). Then the optimal control problem is to find the optimal control policy, given some initial state \((s_0, Q_0)\), that minimizes the expected value of discounted infinite horizon cost of executing such a policy on the system starting at state \((s_0, Q_0)\), i.e.,

\[
\mu^*(s_0, Q_0) = \arg\min_{\mu} E_{\mu}\left[\sum_{t=0}^{\infty} \beta^t c((s_t, Q_t), (s_{t+1}, Q_{t+1}), u_t)\right] / (\ddot{s}(Q), u).
\]

Noting that the above is an MDP, it can be seen that the optimal policies for the above problem are stationary and deterministic and can be found as follows:

\[
u^*(s, Q) = \arg\min_u \sum_{(r, Q)} p(r/s, u)\beta^t c((s, Q), (\ddot{s}, Q), u) + \beta J^*(r, Q).
\]

where

\[
J^*(s, Q) = \min_u \sum_{(r, Q)} p(r/s, u)\beta^t c((s, Q), (\ddot{s}, Q), u) + \beta J^*(r, Q).
\]

Let

\[
\Gamma^*((s, Q), u) = \sum_{(r, Q)} p(r/s, u)\beta^t c((s, Q), (\ddot{s}, Q), u) + \beta J^*(r, Q).
\]

then note that

\[
u^*(s, Q) = \arg\min_u \Gamma^*((s, Q), u).
\]

Further simplifications to the above control problem may be obtained by exploiting the stationarity and uncontrollability of the environment which reduces the computational burden of the DP algorithms that are use to solve the above fixed point equations. It can be easily shown that

\[
\hat{J}^*(s) = \sum_{Q} \hat{p}(Q) J^*(s, Q),
\]

\[
\hat{c}^*((s, Q), u) = \sum_{(r, Q)} p(r/s, u)\beta^t c((s, Q), (r, Q), u).
\]

Further, note that the average cost-to-go function \(\hat{J}^*(s)\) is the solution of the fixed point equation

\[
\hat{J}^*(s) = \sum_{Q} p^*(Q)\left\{\min_u \hat{c}^*((s, Q), u) + \beta \sum_{r} p(r/s, u)\hat{J}^*(r)\right\}.
\]

It can be shown that the operator underlying the fixed point equation is a contraction operator with contraction factor equal to the discount factor \(\beta\), and thus, usual policy and value iteration methods can be applied to the above equation in order to solve for the average cost-to-go function in the fixed point equation above.

\[\text{C. Observation of the Environment}\]

The environment observation model is as follows. Let \(p(\bar{Q}/Q, s)\) denote the sensor observation model, i.e., the model represents the probability that an observation, \(\bar{Q}\), is made when the environment is actually at the state \(Q\) given that the observation is made form state \(s\). Further, it is assumed that the environmental observation model can be factored as follows:

\[
p(\bar{Q}/Q, s) = \prod_{k=1}^{M} p(\bar{q}_k/q_k, s),
\]

where \(\bar{q}_k\) is the noise corrupted observation of the \(k\)th environmental component. The above assumption implies that the measurement of the individual environmental components are independent of each other. To simplify notational issues, it is assumed throughout this paper that the whole environment can be observed from every state \(s\). The results of this paper hold nevertheless with some minor and relatively straightforward modifications.

Given perfect knowledge of the environment, i.e., knowledge of the probabilities \(p^*(q_i)\), the observation model above can be used to infer the current state of the environment based on the true model \(p^*(q_i)\), given the current observation of the environment \(\bar{q}_i\). However, note that in the mapping problem the environment is unknown and thus, the probabilities \(p^*(q_i)\) also need to be estimated. In this paper it is assumed that the true environmental random process, characterized by its stationary probabilities, \(p^*(q_i)\), is unknown, and thus, it is reuire to both:

- Form an estimate of the true underlying process stationary probabilities \(p^*(q_i)\) based on observations of the environment, and
- Estimate the current state of the environment based on the current observation, which in turn depends on the problem stated above.

\[\text{D. Problem Formulation}\]

Thus, from the developments so far in this section it is clear that the knowledge of the environmental probabilities \(p^*(q_i)\) is critical to both the planning (see the planning equation 19) and the mapping/estimation problems outlined above. Thus, the general problem consists of two parts:

- **Mapping:** This is the problem of estimating the probabilities of the underlying environmental random process, i.e., the quantities \(p^*(q_i)\) and also, the estimation of the current environment state given the current observation.
- **Planning:** This is the problem of taking a control action based on the current estimate of the environment and the state of the system.

Note that in the planning problem above, the state of the system might be unknown too, i.e., there is imperfect state information. Further, the autonomous system might need to estimate its own state (localization) relative to the map that it is building in the absence of an absolute state sensor such as GPS. This problem is sometimes known as the Simultaneous,
Planning, Localization and Mapping problem (SPLAM) in the Robotics community. In this paper, an answer is provided to this problem. In part I, the case with perfect state information, (i.e, the state of the autonomous system is known perfectly), is considered while in part II, the methodology is extended to the case of imperfect state observation, which in turn provides a solution to the SPLAM problem.

IV. MAPPING AND PLANNING: PERFECT STATE SENSING

In this section, we present the mapping and planning algorithms under the assumption that the state of the autonomous system is perfectly known and there is noise only in sensing the environment. In part II of the paper, the approach is generalized to the case when the state cannot be sensed perfectly.

A. Mapping/Estimation Problem

As mentioned above, in this part of the paper, the case with perfect state information is considered and thus, there is no need to estimate the state of the system. In the following, we outline the algorithms that are used to estimate the probabilities $p^*(q_k)$ characterizing the environmental random process as well as the estimate of the current state of the environment, $Q$, based on the current observation $\hat{Q}$.

Let

$$p(q_k = j | q_k = i, s) \equiv p^j_i(s),$$

and let the environmental probabilities for the $k^{th}$ environmental component, $p^k(q_k)$, be represented by the vector $P_k^k$. Note that the vector is $D$-dimensional, and its $i^{th}$ element is the probability that the $k^{th}$ environmental component is in state $i$ among the $D$ possible states. Let the state of the system at time $t$ be denoted by $s_t$ and let the observed value of the $k^{th}$ environment component at the same time be $\hat{q}_{k,t}$. Then define

$$A^{(k)}(s_t) = [p^j_i(s_t)],$$

$$b^{(k)}(\hat{q}_{k,t}) = [\beta_j(q_{k,t})], \beta_j(q_{k,t}) = 1(\hat{q}_{k,t} = 1).$$

Let $\{\gamma_t\}$ be any deterministic sequence with the following property:

$$\sum_t \gamma_t = \infty,$$

$$\sum_t \gamma_t^2 < \infty.$$

Then the probability vectors $P_k^*$ can be approximated using one of the following recursive schemes:

**Estimator E1:**

$$P_k(t) = \arg \min_{p \in P} ||A_k(t)p - C_k(t)||,$$

$$A_k(t) = (1 - \gamma_t)A_k(t-1) + \gamma_t A^{(k)}(s_t),$$

$$b_k(t) = (1 - \gamma_t)b_k(t-1) + \gamma_t b^{(k)}(\hat{q}_{k,t}).$$

**Estimator E2:**

$$P_k(t) = \Pi_P \{(1 - \gamma_t)P_k(t-1) + \gamma_t (b^{(k)}(\hat{q}_{k,t} - A^{(k)}(s_t)P_{t-1})\}.$$

where $\Pi_P$ represents the projection into the subspace of probability vectors.

The above algorithms are instances of stochastic approximation algorithms. Note that there is a separate recursive estimator for every environmental component. This exploits the fact that the environmental components are independent of each other and that their observations are independent of each other. An estimate of the current environmental state is then formed based on the current estimate of the probabilities of the environment process, $P_k(t)$, and the current observation of the environment, $\hat{Q}_t$, using simple Bayesian inference as follows:

$$b_t(Q) = \frac{p(\hat{Q}_t/Q, s)p^*(Q)}{\sum_Q p(\hat{Q}_t/Q, s)p^*(Q)}.$$

The above equation can be further simplified to show that the estimate/belief on the environment is the product of the beliefs on the individual components, i.e,

$$b_t(Q) = \prod_{k=1}^{M} b_t(q_k) = \prod_{k=1}^{M} \eta_k p(\hat{q}_{k,t}/q_k, s_t)p_t(q_k),$$

where $\eta_k$ is the normalization factor for the $k^{th}$ component of the environment.

B. Planning/Adaptive Control

Given the mapping/estimation schemes outlined in the previous section, the control policies can then be defined as follows. Recall that $P_k(t)$ represents the estimate of the stationary probabilities of the environmental process, and $b_t(Q)$ represents the belief on the environment at time $t$. Let

$$\Gamma_t(s, Q, u) = \{\bar{c}_t(s, Q, u) + \beta \sum_r p(r/s, u)J_t(r)\},$$

where $\bar{c}_t(s, Q, u) = \sum_{(r/Q)} p(r/s, U)p_t(\bar{Q})c((s, Q), (r, Q), u)$, $J_t(r) = \sum_Q p_t(\bar{Q})J_t(r, \bar{Q})$, and $\bar{J}_t(r)$ is the fixed point of the equation

$$\bar{J}_t(s) = \sum_Q p_t(Q)\min_u [\bar{c}^*(s, Q, u) + \beta \sum_r p(r/s, u)\bar{J}_t(r)],$$

which can be evaluated using value or policy iteration schemes. Note that

$$p_t(Q) = \prod_{k=1}^{M} p_t(q_k),$$

where note that $p_t(q_k)$ is the $q_k^{th}$ element of $P_k(t)$.

A continuity result for the above equations with respect to the estimates of the environment probabilities now stated.

**Proposition 1.** Let the estimates $p_t(Q) \rightarrow p^*(Q)$ for all $Q$. Then, $J_t(s) \rightarrow J^*(s)$ for all $s \in S$ and $\Gamma_t(s, u) \rightarrow \Gamma^*(s, u)$ for every state, control pair $(s, u)$. 


The above result is not proved in this but can be done using straightforward continuity arguments for the nonlinear operator underlying the fixed point equations above.

Then, the control policy is based upon one of the following control schemes:

**Control Policy C1:** Let

\[
\hat{Q}(t) = \arg \max_Q b_t(Q), \quad (37)
\]

\[
u_t = \arg \max_u \Gamma_t((s_t, \hat{Q}_t), u). \quad (38)
\]

**Control Policy C2:**

\[
u_t = \arg \min_u \sum_Q \Gamma_t((s_t, Q), u)b_t(Q). \quad (39)
\]

Note that the full belief state MDP may also be solved for the control policy, however, that problem is very computationally intensive especially as the environment becomes large. Hence, the above control algorithms. The first policy above is based on traditional control practice wherein a maximum likelihood estimate of the current state of the system is formed and then the control is based on the estimate somewhat called the “certainty-equivalence” principle. The second policy weights the various different environment values differently according to the belief on Q and thus, might be better in a probabilistic sense [?].

**C. Convergence Analysis**

In this section, the convergence of the mapping and planning algorithms presented in the previous section is established. The main steps in the proofs are presented while the actual proofs are relegated to the appendix and can be left for later without any loss of continuity.

Let \( F^t = \{ (s_0, Q_0), \ldots, (s_t, \hat{Q}_t) \} \) denote the history of the process time \( t \). Also, note that \( Q_t = (\hat{q}_{1,t}, \hat{q}_{2,t}, \ldots) \).

Let \( X_{k,t} = (s_t, q_{k,t}) \), i.e., the state-kth environmental component observation pair. Then, the equation (22) and (23) can be written as functions of \( X_{k,t} \) instead of \( s_t \) and \( \hat{q}_{k,t} \) respectively. Thus, the estimation algorithms (E1) may be written as:

\[
P_k(t) = \arg \min_{p \in P} || A_k(t) P - b_k(t) ||, \quad (40)
\]

\[
A_k(t) = (1 - \gamma_t) A_k(t - 1) + \gamma_t A^{(k)}(X_{k,t}), \quad (41)
\]

\[
c_k(t) = (1 - \gamma_t) b_k(t - 1) + \gamma_t b^{(k)}(X_{k,t}). \quad (42)
\]

The estimation algorithm (E2) may be reframed accordingly as a function of \( X_{k,t} \).

The following result is a key structural property of the mapping/estimation problem as posed in this paper and is at the basis of proving the convergence of the estimation and planning algorithms.

**Lemma 1.** Let \( A^{(k)}(X_{k,t}) \) and \( b^{(k)}(X_{k,t}) \) be as defined above, and let \( P^*_k \) denote the true stationary probabilities corresponding to the the \( k^{th} \) environmental component. Then, the following holds for all environmental components \( k \):

\[
E[A^{(k)}(X_{k,t+1})P^*_k - b^{(k)}(X_{k,t+1})/F^t] = 0. \quad (43)
\]

**Proof.** See Appendix.

We make the following assumption about the sensors, which is essentially a condition for being a reliable sensor. Recall that \( \hat{q}_k = j/q_k = i, s \) \( \equiv p^k_{ij}(s) \), the observation model for sensor observation of the \( k^{th} \) environmental component from the state \( s \).

**A 1.** It is assumed that for the observation model corresponding to every environmental component \( k \), from any possible state \( s \), there exist \( \epsilon > 0 \), such that

\[
p^k_{ij}(s) \geq 0.5 + \epsilon. \quad (44)
\]

The above assumption is the other key to the convergence of the mapping and planning algorithms. The above assumption, in layman’s terms, implies that the correct observation is made by the sensor at least half the time, and thus, corresponds to a reliable sensor. Note that otherwise the sensor is just as reliable as tossing a fair coin, and making the inference based on the outcome of the toss. Of course, requiring that the above assumption is valid for every \( s \) is very restrictive, however, it is done purely for notational convenience in the theoretical development. In practice, in forming the estimate of any particular environmental component, only readings from those states should be trusted where the above assumption is satisfied, and the estimation algorithms can then be suitable modified to reflect the same.

Then, with the above assumption, the following result can be shown:

**Proposition 2.** Let assumption A 1 hold. Then, under control policy (C1) or (C2), and estimation algorithms (E1) or (E2), (i.e., any combination of the planning and mapping algorithms), the environmental probability estimates, \( \hat{P}_k(t) \to P^*_k \), for all \( k \), w.p. 1. Furthermore, the optimal cost-to-go estimates \( \Gamma_t(s) \to J^*(s) \), w.p. 1, as well as the optimal Q-values \( \Gamma_t(s, u) \to \Gamma^*(s, u) \), w.p. 1.

**Proof.** See Appendix.

The above result holds not only for control policies (C1) and (C2) but in fact for any non-anticipative control policy, i.e., the control at the current time instant is independent of the future of the algorithm. Mathematically, this means that the control variable \( u_t \) should be \( F^t \) measurable, which is indeed true for control policies (C1) and (C2).

Also, there is a Separation Principle, as in classical linear control theory [33], at work here. In other words, the estimators/mapping algorithms can be designed independently of the control algorithms, without affecting the convergence of the composite algorithms. This is indeed another key structural property of the problem as formulated in this paper.