Information Space Receding Horizon Control

Suman Chakravorty
Associate Professor, Aerospace Engineering
Texas A&M University
College Station, TX
Email: schakrav@tamu.edu

R. Scott Erwin
Principal Research Scientist, Space Vehicles Directorate
Air Force Research Laboratory
Albuquerque, NM
Email: richard.erwin@kirtland.af.mil

***Abstract*—In this paper, we present a receding horizon solution to the problem of optimal sensor scheduling problem. The optimal sensor scheduling problem can be posed as a Partially Observed Markov Decision Process (POMDP) whose solution is given by an Information Space (I-space) Dynamic Programming (DP) problem. We present a simulation based stochastic optimization technique that, combined with a receding horizon approach, obviates the need to solve the computationally intractable I-space DP problem. The technique is tested on a simple sensor scheduling problem where a sensor has to choose among the measurements of N dynamical systems such that the information regarding the aggregate system is maximized over an infinite horizon.**

I. Introduction

In this paper, we consider the problem of optimal sensor scheduling such that the information gained by the sensor is maximized. It is easily shown that the scheduling problem, in general, may be posed as a Partially Observed Markov Decision Problem (POMDP) whose solution is given by an information space (I-space) Dynamic Programming (DP) problem. We propose a receding horizon control (I-space RHC: IS-RHC) approach to solve such I-space DP problems. The online stochastic optimization problems that result from the receding horizon approach are solved using a simulation based gradient ascent technique. The IS-RHC technique is tested on a simple scheduling problem where the sensor has a choice between measurements of N dynamical systems.

In recent years, the optimal sensor scheduling problem has garnered a lot of interest in the Control and Robotics community and is variously known as Information-theoretic Control/ Active Sensing and Dual Control [1]–[5]. Discrete dynamic scenarios such as target tracking [3], [4], and linear spatially distributed systems [6] have been considered, but relatively very little has been done on the optimal sensing of nonlinear dynamical phenomenon. In the linear dynamical scenario, the optimal scheduling problem results in a deterministic optimal control problem which can be solved online using Model Predictive Control (MPC: see below for a dicussion on the MPC literature). In the nonlinear case, the problem is stochastic and thus, is significantly harder to solve because of the associated computationally intractable stochastic DP problem. In this paper, we suggest a receding horizon control approach to the solution of such stochastic sequential decision making problems, in particular, I-space sequential decision making problems, that bypasses the need to solve the stochastic DP problem.

It is very well known that stochastic control problems with sensing uncertainty, of which sensor scheduling problems are a special case, can be posed as a Markov Decision Problem (MDP) on the Information state (I-state), which is usually the conditional filtered pdf of the state of the system [7]–[9]. Unfortunately, it is also equally well known that such problems are notoriously difficult to solve owing to the twin curses of dimensionality and history, so much so that such problems have only been solved for small to moderate sized discrete state space problems (i.e., wherein the underlying state space of the problem is discrete). Initially, exact solution of the POMDPs were sought [9] utilizing the convexity of the cost-to-go function in terms of the I-state. However, these techniques do not scale well. Thus, focus shifted to solving such I-space problems using point based value iteration in which a set of I-states are sampled in the I-space and an approximate MDP defined on these states is solved using standard DP techniques such as value/ policy iteration [10], [11]. Recently, there has been also a growing interest in online solution techniques for POMDPs [12]. These methods have resulted in the solution of much higher dimensional problems when compared to the ones that can be solved using exact techniques, however, they still do not scale to continuous state, observation and control space problems.

Theoretically, the DP problem can be solved online using feedback policy gradient algorithms from the Reinforcement Learning (RL) literature [13]–[15]. These techniques parametrize the feedback policy in terms of a parameter θ which used in conjunction with an approximation architecture gives the parametrized feedback policy $\pi(\chi, \theta)$ where χ is the information state of the system. However, getting this parametrization is exceedingly difficult especially in the case of the continuous I-space problems considered in this paper. Further, there is the issue of getting a good initial feedback policy in terms of a good initial parameter estimate θ_0 which is again exceedingly difficult for continuous I-space problems. In contrast, by considering the open loop policy as advocated here and using the policy gradient technique on such a policy, we get rid of the practical difficulties mentioned above. This is also reflected in the fact that, to the

best of our knowledge, feedback policy gradient techniques have not been successfully applied to continuous I-space problems as the ones considered here (in fact, to the best of our knowledge, existing techniques for I-space problems are for small to moderate sized discrete problems only).

Model Predictive or Receding Horizon Control (MPC/ RHC) has been one of the most successful applications of control theoretic techniques in the industry [16]–[18]. The MPC technique and the Dynamic Programming technique essentially give the same answer in that they provide the optimal feedback control solution to deterministic optimal control problems. The MPC techniques solve a sequence of finite horizon open loop control problems in a receding horizon fashion instead of solving the infinite dimensional DP equation offline. In this fashion, constraints on the systems can be taken into account, which is very difficult in DP, provided the open loop optimal control problems can be solved online. This has led to many successful applications in the process control industry [16], [17]. We propose a similar approach to solve I-space sequential decision making problems, wherein a sequence of open loop stochastic optimization problems are solved online in a receding horizon fashion. However, in the stochastic case, the answers of the RHC and the DP techniques do not coincide because in the DP formulation, the optimization is over feedback policies and not open loop control sequences. However, such DP problems, in particular, I-space problems, are virtually computationally intractable in continuous state spaces and thus, the IS-RHC techniques provides a computationally attractive solution to the I-space problems. At the same time, the empirical results show that the IS-RHC techqniue does lead to better payoffs in terms of information gains when compared to a shortsighted policy. We mention here that the idea of IS-RHC as proposed here has antecedents in the Control Systems literature of the 60s/ 70s when researchers were trying to obtain computationally tractable solutions to the Dual Control problem as posed by Feldbaum in his seminal work [19]. These references [20], [21] advocated solving the I-space problems in an open loop receding horizon fashion as advocated here. These techniques, in general, solved deterministic open loop optimization problems that were approximations of the true stochastic optimization problem in a receding horizon fashion. Our technique is different from this early work in that 1) we actually perform the stochastic optimization using a simulation based technique, without having to consider a deterministic approximation based on the nominal path, and 2) it is developed for discrete control spaces.

The original contributions of this paper are as follows: 1) we propose an online receding horizon approach to the solution of the POMDP problem that results from the sensor scheduling problem, and 2) we propose a simulation based gradient ascent approach to the solution of the stochastic optimization problems resulting from the receding horizon approach at every time step. Our technique is valid for continuous state and observation space POMDP problems with a discrete control space.

The rest of the paper is organized as follows. In Section II, we formulate the sensor scheduling problem. In Section III, we present the IS-RHC technique. In Section IV, we present a simple numerical example as application of the IS-RHC technique.

II. Motivation and Problem Formulation

In this section, we introduce the sensor scheduling problem that we wish to solve in this paper. We show that the problem may be posed as an Information space (I-space) Markov Decision Problem (MDP). However, the high dimensionality of the resulting I-space Dynamic Programming (DP) problem precludes a computational solution to the problem thereby motivating the need for a computationally tractable approach to solving the I-space MDP that is different from the DP based approach.

Consider a dynamical system with state denoted by X where $X = [X^{(1)}, X^{(2)}, ...X^{(N)}]'$ and $X^{(i)}$ is a vector that represents the state of a dynamical subsystem whose dynamics may (or may not) be coupled with the dynamics of the other dynamical subsystems. Let the dynamics of the entire system be represented by the following nonlinear difference equation:

$$X_k = F(X_{k-1}) + G(X_{k-1})W_{k-1}, \tag{1}$$

where $F(.)$ and $G(.)$ are nonlinear functions, and $\{W_k\}$ is an uncorrelated white noise sequence. If the sub dynamical systems were decoupled the above equation would decompose into N independent difference equations, one for each of the sub-states $X^{(i)}$. The measurement equation for the state of the system is denoted by the following (possibly) nonlinear equation:

$$z_k = H_{u_k}(X_k) + V_k, \tag{2}$$

where $\{V_k\}$ is a zero mean uncorrelated white noise sequence, and $H_{u_k}(.)$ is a nonlinear measurement function where the variable u_k is a control variable that can take values from 1 to N, and denotes that we make a measurement of sub-state $X^{(i)}$ at time k, if $u_k = i$. This implies that we can only measure one sub-component of X at any time step, and the control u_k denotes which sub-state is measured. Of course, we might have the choice of making $P > 1$ measurements as well, however, for notational simplicity we shall concentrate only on $P = 1$ in the following. The generalization to $P > 1$ is quite straightforward.

Let χ_k denote the pdf of the state X at time k. We shall call χ_k the information state of the system since it encodes our knowledge (or lack thereof) about the system state X. In the Gaussian case, i.e., under the approximation that the pdf remains Gaussian, χ_k is encoded by the mean and covariance of the state vector X. Given the information state (I-state) at time $k-1$, χ_{k-1}, the I-state at time k, χ_k, will depend on the particular component that is chosen for measurement at time

k and hence, on the control variable u_k. It is also clear that the I-state χ_k is dependent on the noisy observation at time k, z_k. However, z_k is a random variable and thus, the I-state χ_k is also random given the previous I-state χ_{k-1} and the control u_k. In fact, the evolution of the I-state is governed by a Markov chain (MC) whose transition density function may be found as follows (for notational convenience, we drop the explicit reference to time k):

$$p(\chi'/\chi,u) = \int p(\chi'/z,\chi,u)p(z/\chi,u)dz, \quad (3)$$

$$= \int \underbrace{p(\chi'/z,\chi)}_{\delta(\chi'-T(z,\chi))} p(z/\chi,u)dz, \quad (4)$$

$$T(z,\chi)(X) = \eta p_u(z/X)\int p(X/X')\chi(X')dX', \quad (5)$$

$$p(z/\chi,u) = \eta = \int p_u(z/X)(\int p(X/X')\chi(X')dX')dX. \quad (6)$$

In the above equations, $p_u(z/X)$ represents the likelihood of the measurement z given that the underlying state is X and sub-state denoted by control u is measured, and can be found from Eq. 2, $p(X/X')$ is the transition density of the underlying Markov chain governing the evolution of the state X (not to be confused with the MC governing the evolution of the I-state χ) and is found from Eq. 1, $\chi(X)$ represents the probability that the underlying true state of the system is X and $\delta(.)$ is the indicator function for the event $\chi' = T(z,\chi)$. The filtering Eq. 5 is the optimal Bayes recursion governing the evolution of the conditional density of the state in a nonlinear system [22]. This recursion is replaced by the Kalman recursion given the approximation that the state pdf remains Gaussian (the Kalman filter, the Extended Kalman Filter (EKF), the Unscented Kalman Filter (UKF) etc.). We note here that the evolution of the I-Space MDP in the field of Partially Observed Markov Decision Processes (POMDP) [7] is similar except that in that case the transition density of the MC governing the evolution of the hidden state is control dependent , i.e., $p(X'/X,u)$, as opposed to the measurement equation being control dependent as is in this case through the controlled likelihood function $p_u(z/X)$. In any case both result in an I-state MDP whose control dependent transition density is given by $p(\chi'/\chi,u)$. The methods of this paper apply equally well to any POMDP with a finite set of control variables.

Our objective in this work is to maximize the total information about the dynamical system over the infinite horizon. To this end, let us define the information gain metric $\Delta I(\chi,u)$ denoting the expected information gain in choosing control u at I-state χ. For instance, this could be the trace of the reduction that a measurement would make in the estimated covariance of the state vector X; the determinant of the inverse of this covariance reduction (the information gain); or divergence measures taken between forcasted values of χ and expected values of χ if a measurement is taken. An excellent discussion of metrics for sensor tasking problems can be found in [23]. Let $0 < \beta < 1$ denote a discount factor that quantifies the fact that the information gains in the immediate future are more important to us than the information gains further out in the future. Thsi allows us to effectively reduce the infinite horizon optimization into a finite horizon optimization. We wish to solve the following discounted sequential decision making problem over all possible feedback policies $u(.)$:

$$V^*(\chi) = \max_{u(.)} V(\chi,u(.)), \text{where} \quad (7)$$

$$V(\chi,u(.)) \equiv E[\sum_{t=1}^{\infty}\Delta I(\chi_t,u(\chi_t))\beta^t/\chi_0 = \chi], \quad (8)$$

for all possible information states χ. Since the I-state χ is governed by a controlled Markov chain, the answer to the above problem is provided by solving the following Dynamic Programming problem:

$$V^*(\chi) = \max_u[\Delta I(\chi,u) + \beta\int p(\chi'/\chi,u)V^*(\chi')d\chi']. \quad (9)$$

Thus, at least theoretically, the I-space sequential decision problem is solved if the above DP problem can be solved. However, in the general nonlinear case, χ belongs to a function space (the space of pdfs). Even in the case when the Gaussian approximation holds, the I-state χ is encoded by the mean and joint covariance of the state X, thereby making the I-state dimension $Nd+N^2d^2$, where d is the dimension of each sub-component $X^{(i)}$ of X. Thus, it is clear that the above DP problem resides in a continuous and very high dimensional state space thereby making the problem computationally intractable.

III. Information Space Receding Horizon Control (IS-RHC)

In this section, we shall propose a simulation based Receding Horizon Control approach (IS-RHC) to solve the I-space MDP problem that was posed in the previous section. However, the feedback solution that is obtained using the technique is, in general, different from the feedback solution that would be obtained if the I-space DP problem was computationally tractable. In the special case when the I-space sequential decision problem is deterministic, for instance when the underlying system is linear and the pdfs are Gaussian, the DP solution and the I-space RHC solution proposed here are one and the same. Please see the remark at the end of this section for a more detailed discussion of this issue.

A. Stochastic Relaxation of Optimization Problem

Consider again the statement of the I-space MDP given in Eq. 7. Given that the expected information gain is uniformly bounded above, i.e., $|\Delta I(\chi,u)| < M < \infty$ for all (χ,u) and given the discount factor $\beta < 1$, and given any arbitrarily small error tolerance $\epsilon > 0$, there always exists a finite time T such that the finite horizon T-step discounted information gain $J_T(\chi,u_0,u_1...,u_T)$ for any infinite horizon control sequence

$\{u_t\}_{t=1}^{\infty}$ is arbitrarily close to the infinite horizon discounted cost-to-go for the same control sequence, i.e.,

$$J(\chi, \{u_t\}_{t=1}^{\infty}) = E[\sum_{t=1}^{\infty} \Delta I(\chi_t, u_t)\beta^t / \chi_0 = \chi]$$
$$\approx E[\sum_{t=1}^{T} \Delta I(\chi_t, u_t)\beta^t / \chi_0 = \chi] = J_T(\chi, \{u_t\}_{t=1}^{T}),$$

in the sense that $|J_T(\chi, \{u_t\}_{t=1}^{T}) - J(\chi, \{u_t\}_{t=1}^{\infty})| < \epsilon$ for all χ. Thus, in the following we shall concentrate on solving the discounted finite horizon I-space MDP assuming that a finite horizon T and discount factor β is given such that the above approximation holds thereby leaving us with a finite dimensional optimization problem as opposed to the infinite dimensional problem resulting from the original infinite horizon case.

Define the T-step information gain in following the T-step control sequence $U = \{u_1, \cdots, u_T\}$ from I-state χ as follows:

$$J(\chi, u_1, \cdots, u_t) = E_{\chi}[\sum_{t=1}^{T} \Delta I(\chi_t, u_t)\beta^t / \chi_0 = \chi], \quad (10)$$

where the notation $E_{\chi}[.]$ denotes that the expectation is over the sample paths $\{\chi_0 = \chi, \chi_1, \cdots, \chi_T\}$ that are particular $T-$step realizations of the I-space process. We have dropped the subscript T for notational convenience. The above equation is different from Eq. 8 because the expectation above is with respect to an open loop policy while the one in Eq. 8 is with respect to a feedback policy (see Remark 4 at the end of this Section for more details). Further, we define the optimal T-step information gain as follows:

$$J^*(\chi) = \max_{u_1 \cdots u_T} J(\chi, u_1, \cdots u_T).$$

Define a randomized policy $\Pi = \{\pi_1, \cdots \pi_T\}$ where π_t is a probability vector such that $\pi_{t,j} = \Pr(u_t = j)$ where $\pi_{t,j}$ denotes the j^{th} component of π_t. Thus, in the randomized policy, we do not take a particular control action at time t, instead we take the control action $u_t = j$, $j = 1 \cdots N$, with a probability $\pi_{t,j}$ and $\sum_j \pi_{t,j} = 1$ for all t. Further, define the T-step information gain in following stochastic policy Π from I-state χ as follows:

$$J_s(\chi, \Pi) = E_{\chi,u}[\sum_{t=1}^{T} \Delta I(\chi_t, u_t)\beta^t / \chi_0 = \chi],$$

where the notation $E_{\chi,u}[.]$ denotes the fact that the expectation in the above equation is now with respect to both the sample paths $\{\chi_1, \cdots \chi_T\}$ and control sequences $\{u_1, \cdots u_T\}$. Then, it can be seen that the following relationship holds:

$$J_s(\chi, \Pi) = \sum_{u_1, \cdots u_T} J(\chi, u_1, \cdots u_T)\pi_{1,u_1} .. \pi_{T,u_T}, \quad (11)$$

the summation above is a T-dimensional sum where each u_t can take one of N values. In the following, for notational conveneience, we shall abuse notation and denote the expected information gain due to a stochastic policy $J_s(.)$ by $J(.)$ (the symbol for information gain due to a deterministic policy). We wish to solve the optimization problem:

$$J^*(\chi) = \max_{\Pi} J(\chi, \Pi), \quad (12)$$

given some I-state χ. We want to solve the randomized problem since this allows us to use continuous optimization techniques such as gradient descent where the gradients can be found from simulations of the I-space process as shall be shown in the following. Contrast this to the optimization problem for the deterministic policies which results in a combinatorial optimization problem with N^T choices that is intractable for even moderate number of choices N and lookahead horizon T.

B. Simulation based Stochastic Gradient Method

Let us write $\pi_{t,N} = 1 - \pi_{t,1} - ... - \pi_{t,N-1}$. Then, Eq. 11 reduces to the following equation:

$$J(\chi, \Pi) = \sum_{u_1 \ldots u_T} [\sum_{j=1}^{N} J(\chi, u_1, ..., u_t = j, .., u_T)\pi_{t,j}]$$
$$\pi_{1,u_1} \ldots \pi_{T,u_T},$$
$$= \sum_{u_1 \ldots u_T} \pi_{1,u_1} \ldots \pi_{T,u_T} [\sum_{j=1}^{N-1} J(\chi, u_1, ..., u_t = j, .., u_T)\pi_{t,j}$$
$$+ J(\chi, u_1, ..., u_t = N, ..., u_T)(1 - \pi_{t,1} - ... - \pi_{t,N-1})]. \quad (13)$$

Note that $J(\chi, \Pi)$ is a multi-linear function of the probabilities $\pi_{t,j}$. Then, from Eq. 13, it follows that:

$$\frac{\partial J(\chi, \Pi)}{\partial \pi_{t,j}} = \sum_{u_1 .. u_T} \pi_{1,u_1} .. \pi_{T,u_T}$$
$$\{J(\chi, u_1, .., u_t = j, .., u_T) - J(\chi, u_1, .., u_t = N, .., u_T)\}. \quad (14)$$

Consider the term $\sum_{u_1 .. u_T} \pi_{1,u_1} .. \pi_{T,u_T} J(\chi, u_1, .., u_t = j, .., u_T)$. This is nothing but the expected T-step information gain in following stochastic policy Π whenever $u_t = j$. Define

$$J_{(t,j)}(\chi, \Pi) = \sum_{u_1 .. u_T} \pi_{1,u_1} .. \pi_{T,u_T} J(\chi, u_1, .., u_t = j, .., u_T), \quad (15)$$

where subscript (t, j) denotes the gradient of $J(\chi, \Pi)$ with respect to $\pi_{t,j}$. Then, using the above definition and Eq. 14, it follows that

$$\frac{\partial J(\chi, \Pi)}{\partial \pi_{t,j}} = J_{(t,j)}(\chi, \Pi) - J_{(t,N)}(\chi, \Pi), \quad (16)$$

for all t and all j. Thus, by simulating sample I-space trajectories under the stochastic policy Π, we can estimate the gradient of the T-step information gain function $J(\chi, \Pi)$ with respect to each of the control probabilities $\pi_{t,j}$. Then, the policy Π can be improved by ascending along the gradient $\frac{\partial J(\chi, \Pi)}{\partial \Pi}$. Note that the gradient $\frac{\partial J(\chi, \Pi)}{\partial \Pi}$ is a $T \times N$ matrix whose (t, j) element is $\frac{\partial J(\chi, \Pi)}{\partial \pi_{t,j}}$. Mathematically, this means we adjust the stochastic policy at iteration n (not to be confused with time t) as follows:

$$\Pi_{n+1} = \mathcal{P}_P\{\Pi_n + \epsilon_n \frac{\partial J(\chi, \Pi)}{\partial \Pi}|_{\Pi = \Pi_n}\}, \quad (17)$$

where ϵ_n is a small step size and $\mathcal{P}_P(.)$ denotes a projection onto the space of stochastic policies P. The projection is necessary since the new policy update need not satisfy the constraints required to be satisified by a stochastic policy. This projection results in a quadratic programming problem whenever the constraints are violated. However, estimating $\frac{\partial J(\chi,\Pi)}{\partial \Pi}$ exactly is intractable owing to the large number of simulations required to do the estimation. Instead, we can form a noisy estimate of $\frac{\partial J(\chi,\Pi)}{\partial \Pi}$ from a single sample path (simulation) of the I-space process as follows. Recall Eq. 16 and suppose that ω is a sample realization of the I-space process, where $\{\chi_1(\omega), u_1(\omega), \cdots \chi_T(\omega), u_T(\omega)\}$ denotes the sample I-space/ control space path, with associated information gain $J(\omega)$. Then, the information gradient equation 16 can be approximated by using the noisy information gradient estimate:

$$\begin{aligned}\frac{\widehat{\partial J(\chi,\Pi)}}{\partial \pi_{t,j}} &= \frac{J(\omega)}{\pi_{t,j}} \text{ if } u_t(\omega) = j, \\ &= \frac{-J(\omega)}{\pi_{t,N}} \text{ if } u_t(\omega) = N, \\ &= 0, \text{ o.w..} \end{aligned} \tag{18}$$

Thus, using the noisy information gradient from the Eq. 18, the policy update equation may be written as follows:

$$\Pi_{n+1} = \mathcal{P}_P\{\Pi_n + \epsilon_n \frac{\widehat{\partial J(\chi,\Pi)}}{\partial \Pi}|_{\Pi=\Pi_n}\}, \tag{19}$$

where $\frac{\widehat{\partial J(\chi,\Pi)}}{\partial \Pi}$ is a $T \times N$ matrix whose (t,j) element is $\frac{\widehat{\partial J(\chi,\Pi)}}{\partial \pi_{t,j}}$. Using the noisy policy update Eq. 19 above, we improve the policy by ascending the gradient and in the limit, we would hope to reach an optimum point for the information gain function $J(\chi,\Pi)$. A convergence analysis for the above procedure is provided in Section IV. A few remarks regarding the noisy information gradient equation 18 are in order here.

C. IS-RHC Algorithm

Thus far in this section, we have outlined a simulation based stochastic gradient technique that allows us to find an optimum of the T-step information gain function $J(\chi,\Pi)$ with respect to the stochastic policy Π given some initial I-state χ and some initial guess for the stochastic policy Π_0. In the following, we outline a receding horizon approach which in combination with the stochastic gradient technique allows us to find an online solution to the I-space MDP problem without having to solve the corresponding DP equation.

Suppose at time $t = 0$, the I-state of the system is χ_0. Also suppose that we are given some initial guess for the optimal T-step stochastic policy, say Π_0. Then using the simulation based noisy gradient estimate from Eq. 18 and the policy improvement step from Eq. 19, we can ascend the gradient of the function $J(\chi,\Pi)$ and find an optimum w.r.t Π. This gives us a T-step policy $\Pi_0^* = \{\pi_1^*...\pi_T^*\}$. As in the standard Receding Horizon control approach, we apply the control u_1 according to π_1^*. Next we observe the noisy measurement at time 1, z_1 and update our I-state according to the Bayes filtering Eq. 5 to get the I-state at time 1, χ_1. Assuming that the underlying system is autonomous (note that Eq. 1 is time independent and hence, autonomous), then we can reset time to 0, make χ_1 our new initial I-state χ_0 and repeat the procedure outlined above. In this fashion, at every time step, given the current I-state, the T-step stochastic optimization can be done online and applied in a receding horizon fashion. Mathematically, the RHC-based feedback control for I-state χ can be written as:

$$u_{RHC}(\chi) = e_1^{\mathsf{t}} arg\min_{\Pi} J(\chi,\Pi), \tag{20}$$

where e_1 is the first unit vector in R^T (e_1 isolates the control at the first time instant of the T-step open loop control policy). It should be clear from the above procedure that this amounts to solving the I-space MDP problem given in Eq. 7 without having to solve the associated DP equation, however, the solution obtained using the IS-RHC procedure is, in genral, not the same as that which would be obtained from solving the DP equation (please see remark below for a more detailed discussion). The above recursive procedure is summarized in the pseudo-code IS-RHC.

Algorithm 1 Algorithm IS-RHC

- Given initial information state χ_0, lookahead horizon T, initial policy Π_1 and error tolerance δ
  1) $n = 1$, define $||\Pi_1 - \Pi_0|| = \delta + 1$
  2) WHILE $||\Pi_n - \Pi_{n-1}|| > \delta$ DO
     a) Generate sample I-space path $\{\chi_t(\omega)\}_{t=1}^T$ starting with initial I-state χ_0.
     b) Use Eq. 18 to form the noisy estimate of the information gradient using the sample path.
     c) Use Eq. 19 to update the policy.
  3) Output converged T-step policy $\Pi^* = [\pi_1^*..\pi_T^*]$ and choose control u_1 according to π_1^*.
  4) Observe noisy measurement z_1 and update using Eq. 5 to obtain the new I-state χ_1.
  5) Set $\chi_0 = \chi_1$ and go to Step 1.
- End

Remark 1. *The receding horizon approach is the same for both the IS-RHC and the deterministic MPC techniques since they amount to resolving a T-step optimization problem at every time step using the current (I-)state as the initial condition. However, in the deterministic MPC procedure the optimization problem is deterministic and is usually posed as a deterministic open loop control problem which is then transcribed into a nonlinear programming problem (NLP) that is solved using some standard optimization software [24]. In our case, the optimization is stochastic and no such analytic NLP can be posed due to the complicated nature of the I-state process. However, sample I-space paths can be simulated using suitable (nonlinear) filtering techniques such*

as EKF/ UKF/ particle filters etc. The stochastic gradient based technique proposed here uses these simulations to form a noisy estimate of the information gradient and utilizes this in a gradient ascent algorithm to converge to an optimum of the information gain function. This stochastic optimization is done at every time step given the current I-state, i.e., we do multiple policy update steps for every time step. Due to the receding horizon nature of the problems, this results in an online solution to the I-state MDP without having to solve the associated high dimensional I-space DP equation.

Remark 2. *The feedback policy that results from the IS-RHC (ref. Eq. 20) is different from that which would result from solving the DP equation (ref. Eq. 8). In the DP case the expectation is with respect to the sample paths that are generated as a result of a feedback policy $u(.)$ while in the case of IS-RHC the expectation is with respect to the sample paths generated by an open loop (not feedback) sequence of control actions $\{u_1, \cdots u_T\}$. Thus, the optimization problems are different for the two cases. The IS-RHC backs out a feedback policy from the subsequent open loop optimizations by recognizing that at the next time step, the particular I-state that is observed changes the online optimization problem which is then resolved for the current I-state and thereby constitutes a feedback procedure. However, this feedback policy is necessarily sub-optimal since the DP solution is the optimal feedback policy. In fact, in the deterministic case, i.e., when the I-space paths are deterministic, the two procedures are one and the same because in that case, the open loop control sequence obtained by solving the optimization problem online is exactly the same as the feedback solution that would be obtained by solving the DP equation offline (this is a well known fact in the RHC literature [18]).*

Due to paucity of space, we do not present the convergence analysis in this paper. However, under standard stochastic approximation assumptions, it may be shown that the simulation based information gradient technique presented here converges to a stationary point of the underlying information reward function.

IV. Illustrative Example

In this section, we shall present an application of the IS-RHC technique to an illustrative example containing N decoupled 1-dimensional oscillators. The decoupled nature of the oscillators only affects the filtering algorithm used to generate the sample I-space paths, otherwise, the method is independent of such coupling. The equation of motion of the i oscillator is given by the following difference equation:

$$x_k^{(i)} = (1+h)x_{(k-1)}^{(i)} - h\epsilon(x_{(k-1)}^{(i)})^3 + \sigma_w w_{k-1},$$

while the measurement equation is given by:

$$z_k = e_{u_k} x_k + \sigma_v v_k,$$

where h is the sampling time of the system, e_{u_k} is the u_k^{th} unit vector in $\Re^N$ and denotes measurement of the u_k oscillator, w_k and v_k are zero-mean, unit intensity scalar Gaussian noise processes, and σ_w and σ_v are their respective intensities.

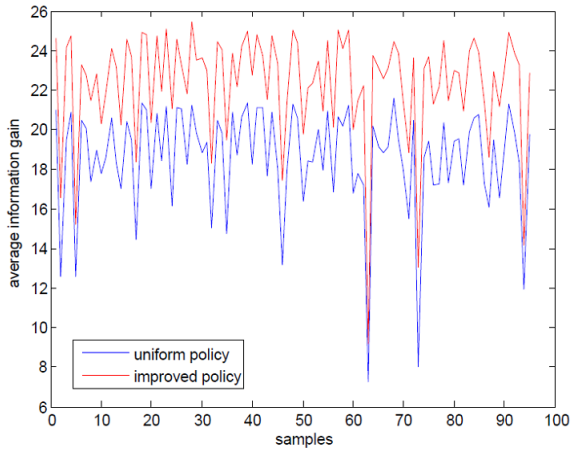

Fig. 1. Comparison of information gained by converged policy and uniform initial policy for $N = 2$ oscillators

An EKF is used to generate the I-space sample paths based on a random generation of the noisy observations based on the measurement noise model and the true "simulated" underlying process. We use an EKF because of is simplicity, however, other advanced Gaussian nonlinear filtering techniques (UKF, iterated extended Kalman filter (IEKF)) or non-Gaussian nonlinear filters (mixture of Gaussian filters, particle filters) can also be used: the sole purpose of the filtering technique is to generate the sample I-space paths and allow the evaluation of the "information gain" along these paths, which is the quantity of interest to the stochastic gradient technique. In our case, the I-state of the process is the 2-tuple (μ, K) where μ contains the means of each of the component states and K is a diagonal matrix containing the variances of the sub-components. The information gain metric that is used is the following:

$$\Delta I(\chi, u) = E[\det(K_{\chi,u}^{-1}) - \det(K_{\chi}^{-1})],$$

where $\det(A)$ represents the determinant of the matrix A, K_χ is the covariance of the I-state χ and $K_{\chi,u}$ is the covariance of the I-state resulting from taking control u at I-state χ, note that the future covariance $K_{\chi,u}$ is random and hence, the expectation is required in the expression above. The above metric results in the so called "D-optimal" design in the Experimental Design literature [25], [26].

The numerical values of the different parameters used in the system simulations are as follows: h = 0.1, ϵ = 0.01, σ_v = 0.4, and $\sigma_w = 0.5$. The discount factor chosen was $\beta = 0.9$. We also did experiments for different values of the noise intensity parameters σ_v and σ_w, and the results that are presented here are typical. We did experiments for a lookahead horizon of 20 time steps and for $N = 2, 4 and 8$ oscillators. The results of our numerical simulations are shown in Figs. 1, 2 and 3. In Figs 1, 2 and 3, we encapsulate the performance of the

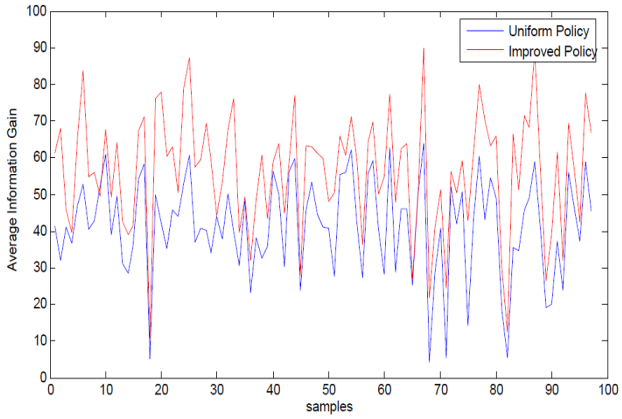

Fig. 2. Comparison of information gained by converged policy and uniform initial policy for $N = 4$ oscillators

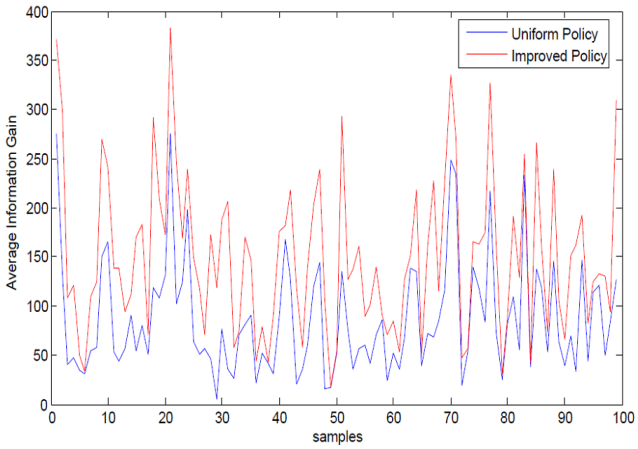

Fig. 3. Comparison of information gained by converged policy and uniform initial policy for $N = 8$ oscillators

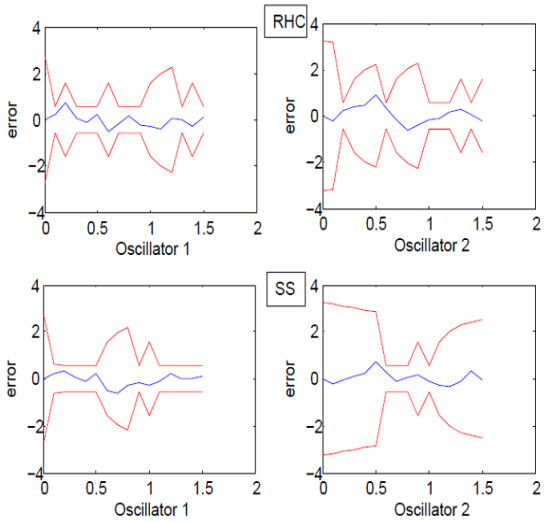

Fig. 4. Comparison of the RHC and shortsighted policies for a sample initial I-state for the case of $N = 2$ oscillators. Note the difference in the 3 σ envelopes which shows the difference between the RHC and the one step ahead policies.

stochastic gradient technique for different randomly generated initial I-states. Given the randomly generated initial I-states, and an initial guess at the T-step policy that is uniform, i.e., the probabilities $\pi_{t,j} = \frac{1}{N}$, the stochastic gradient technique is used to improve the initial policy such that the information gain is maximized. The cases for N= 2, 4 and 8 choices are shown in Figs 1, 2 and 3 respectively. As can be seen from the plots, the stochastic gradient technique does result in improved average information gain when compared to the initial uniform initial policy. Also, it can be seen that the magnitude of the information gained rises with the number of choices. It can also be seen from the plots that the average normalized information gain, i.e., the normalized information gain, (where the normalized information gain for a sample I-state is $\Delta I_f(\chi) - \Delta I_i(\chi)/\Delta I_i(\chi)$ where $\Delta I_i(.)$, $\Delta I_f(.)$ represent the information gain for the initial and final policies respectively), averaged over the different sample initial I-states increases from N = 2 through 8. The averaged normalized information gain for the case of $N = 2$ oscillators is 20.78%, it is 47.7% for the case of N=4 oscillators and is 127.1% for the case of N = 8 oscillators. Thus, the technique results in better payoffs as the number of choices N is increased. A similar behaviour is observed as the lookahead horizon T is increased. However, it has to be noted that as N, T increase, the variance of the stochastic gradient technique increases and can lead to failure more fraction of times, i.e., cases where the information gain of the converged policy is lower than the initial policy. However, in such cases, the variance of the technique can be used to our benefit as repeating the gradient descent process several times (in our case, less than 5 times) usually results in an improved policy almost always. This can be evidenced form the fact that none of the converged policies in Figs. 1, 2 and 3 have information gain lower than the intial policy.

We applied the stochastic gradient technique to solve the I-space MDP in a receding horizon fashion. We compared the solution of the I-space RHC to a greedy approach that looks only one step ahead. The RHC has a lookahead horizon of 15 steps. We performed the RHC vs. Greedy policy simulations for the $N = 2$ case. We evaluated the information gained by the two techniques over a 20 step horizon for a discount factor of β = 0.8. It is seen that the RHC technique does outperform the greedy approach in the $N = 2$ case on an average (over different initial I-states) by a factor of about 20%. In these experiments, at the initial time step, the guess initial policy is chosen to be a uniform policy. In the subsuquent receding horizon time windows, the solution of the optimization in

the previous time window is used as the initial guess for the subsuquent window along with a uniform distribution for the final time step (this is standard practice in the RHC literature [18]). This process of subsuquent initializations speeds up the convergence of the gradient ascent algorithm and also helps it converge to good solutions online. Using a uniform initial policy for the subsuquent receding horizon time windows usually results in very sub-par performance as the gradient technique fails to converge. In Fig. 4, we have shown the errors in the oscillator state estimates along with their 3σ bounds, for both the RHC and greedy policies for a sample initial I-state in the case of $N = 2$ oscillators. The figure shows that the sample behavior of the I-states for the two policies is quite different, as is evident from the distinct 3σ envelopes for the two cases and confirms the benefits of using a longer lookahed horizon when compared to a shortsighted one step ahead policy.

Hence, the above illustrative example shows that the stochastic gradient based technique almost always increases the information gained regarding the system when compared to the uniform initial policy. Further, we see that the fractional amount of information gained increases as the number of choices N increases. This simple example also provides empirical evidence that the IS-RHC technique does indeed result in control policies that maximize the information gained about the system over the long run when compared to shortsighted policies.

V. Conclusion

In this paper, we have proposed a receding horizon control based approach to solve I-space MDP, termed IS-RHC, instead of solving the associated computationally intractable DP equation. We proposed a simulation based stochastic gradient technique for solving the open loop stochastic optimization problem that results at every time step due to the IS-RHC technique. We have tested the IS-RHC on a simple example involving N decoupled 1-dimensional oscillators and the results show that the IS-RHC technique does result in significant improvement in the information gained regarding the system when compared to a shortsighted policy. Further research will focus on testing the IS-RHC on more realistic examples as well as extending the formulation such that constraints on the information process and the problem of decentralized control for multiple sensors can be taken into account.

References

[1] F. Bourgault *et al.*, "Information based adaptive robot exploration," in *Proc. of 2002 IEEE/ RSJ Int. conf. on Intell. rob. sys.*, 2002.

[2] A. Makarenko *et al.*, "A decentralized architecture for active sensor networks," in *Proceedings of the 2004 IEEE INT. Conf. Rob. Automat.*, 2004.

[3] T. H. Chung *et al.*, "On a decentralized active sensing strategy using mobile sensor platforms in a network," in *Proc. IEEE Int. Conf on Dec. Cont.*, 2004.

[4] S. Aranda *et al.*, "On optimal sensor placement and motion coordination for target tracking," in *Proc. IEEE Int. conf. Dec. Cont.*, 2005.

[5] H.-L. Choi, *Adaptive Sampling and Forecasting with Mobile Sensor Networks, PhD thesis*. Cambridge, MA: Department of Aeronautics and Astronautics, MIT, 2008.

[6] A. Y. Khapalov, "Optimal measurement trajectories for distributed parameter systems," *Systems and Control Letters*, vol. 18, pp. 467–477, 1992.

[7] D. P. Bertsekas, *Dynamic Programming and Optimal Control, vols I and II*. Cambridge: Athena Scientific, 2000.

[8] P. R. Kumar and P. P. Varaiya, *Stochastic Systems: Estimation, Identification and Adaptive Control*. Prentic Hall, NJ: Prentice Hall, 1986.

[9] L. P. Kaelbling and M. L. Littman, "Planning and acting in partially observable stochastic domains," *Artificial Intelligence*, vol. 101, pp. 99–134, 1998.

[10] M. Spaan and N. Vlassis, "Perseus: Randomized point-based vallue iteration for pomdps," *Journal of Artificial Intelligence Research*, vol. 24, pp. 195–220, 2005.

[11] J. Pineau *et al.*, "Anytime point based approximations for large pomdps," *Journal of Artificial Intelligence Research*, vol. 27, pp. 335–380, 2006.

[12] S. Ross *et al.*, "Online planning algorithms for pomdps," *Journal of Artificial Intelligence Research*, vol. 32, pp. 663–704, 2008.

[13] P. Marbach, *Simulation based Optimization of Markov Reward Processes, PhD Thesis*. Boston, MA: Massachusetts Institute of Technology, 1999.

[14] J. Baxter and P. Bartlett, "Infinite horizon policy-gradient estimation," *Journal of Artificial Intelligence Research*, vol. 15, pp. 319–350, 2001.

[15] R. S. Sutton *et al.*, "Policy gradient methods for reinforcement learning with function approximation," in *Proc. 1999 Neural Information Proc. Sys.*, 1999.

[16] C. E. Garcia *et al.*, "Model predictive control: Theory and practice," *Automatica*, vol. 25, pp. 335–348, 1989.

[17] S. J. Qin and T. A. Badgwell, "An overview of industrial model predictive control technology," in *Fifth International Conference in Chemical Process Control*, 1997.

[18] D. Q. Mayne *et al.*, "Constrained model predictive control: Stability and optimality," *Automatica*, vol. 36, pp. 789–814, 2000.

[19] A. A. Feldbaum, *Optimal Control Systems*. New York, NY: Academic Press, 1965.

[20] E. Tse and M. Athans, "Adaptive stochastic control for a class of linear systems," *IEEE Transactions on Automatic Control*, vol. 17, pp. 38–52, 1972.

[21] E. Tse *et al.*, "Wide sense adaptive dual control for nonlinear stochastic systems," *IEEE Transactions on Automatic Control*, vol. 18, pp. 98–108, 1973.

[22] B. D. O. Anderson and J. B. Moore, *Optimal Filtering*. London, UK: Dover Publications, 2001.

[23] A. O. Hero III, D. A. Castan, D. Cochran, and K. Kastella, Eds., *Foundations and Applications of Sensor Management*. Springer US, 1989.

[24] I. M. Ross and F. Fahroo, "Pseudospectral knotting methods for solving optimal control methods," *J. Guidance, Control and Dynamics*, vol. 27, pp. 397–404, 2004.

[25] D. J. C. Mckay, "Information-based objective functions for active data selections," *Neural Computation*, vol. 4, pp. 590–604, 1992.

[26] E. Rafajlowicz, "Optimum choice of moving sensor trajectories for distributed parameter system identification," *International Journal of Control*, vol. 43, pp. 1441–1451, 1986.