Stochastic Approximation for Optimal Observer Trajectory Planning

Sumeetpal Singh[a], Ba-Ngu Vo[a], Arnaud Doucet[b], Robin Evans[a]
[a]Department of Electrical and Electronic Engineering
The University of Melbourne, VIC 3010, Australia
{ssss,b.vo,r.evans}@ee.mu.edu.au
[b] Signal Processing Group, Engineering Department
University of Cambridge, Cambridge CB2 1PZ, UK
ad2@eng.cam.ac.uk

***Abstract*— A maneuvering target is to be tracked based on noise corrupted measurements of the target's state that are received by a moving observer. Additionally, the quality of the target state observations can be improved by the appropriate positioning of the observer relative to the target during tracking. The bearings-only tracking problem is an example of this scenario. The question of optimal observer trajectory planning naturally arises, i.e. how should the observer manoeuvre relative to the target in order to optimise the tracking performance? In this paper, we formulate this problem as a discrete-time stochastic optimal control problem and present a novel stochastic approximation algorithm for designing the observer trajectory. Numerical examples are presented to demonstrate the utility of the proposed methodology.**

I. Introduction

Consider the problem of tracking a maneuvering target for N epochs, and let X_k denote the state of the target at epoch k. At each epoch, a sensor provides a noisy (possibly nonlinear) partial observation of the target state $Y_k = g(X_k, A_k, V_k)$, where V_k denotes noise. We have denoted by A_k some parameter of the sensor that may be adjusted to improve the "quality" of the observation Y_k. In tracking, one is interested in computing the probability density π_k of the target state X_k given the sequence of observations $\{Y_1, \ldots, Y_k\}$ received and sensor parameters $\{A_1, \ldots, A_k\}$ until epoch k; π_k is known as the *filtering* density. The *adaptive optimal tracking problem* is to minimise

$$\mathbf{E}\left\{\sum_{k=1}^{N} \beta^k \left(\psi(X_k) - \langle \pi_k, \psi \rangle\right)^2\right\} \tag{1}$$

with respect to (w.r.t.) the choice of sensors $\{A_1, \ldots, A_N\}$. $\beta \in (0, 1]$ is known as the discount factor. The term adaptive implies that the sensor parameter a time k, A_k, is to be selected based on all information available prior to time k, which is $\{Y_1, A_1, \ldots, Y_{k-1}, A_{k-1}\}$. The criterion in (1) is the discrepancy between the true state X_k and the estimate π_k, measured through a suitable function ψ. Problem (1), which is very general, is also known as a Partially Observed Markov Decision Process (POMDP).

In this paper, we are concerned with the scenario where a moving platform (or observer) is to be adaptively maneuvered to optimise the tracking performance of a maneuvering target; thus A_k denotes the position of the observer at epoch k. This problem is termed the optimal *observer trajectory planning* (OTP) problem.

Literature review: When the dynamics of the target and the observation process are linear and Gaussian, then the optimal solution to problem (1) (when ψ gives a quadratic cost) can be computed off-line, i.e. open and closed loop control yield the same performance [8]. A linear Gaussian target and a bearings only observer (non-linear observation) was studied in [6] where a sub-optimal observer trajectory in [6] was computed using a linearised observation equation. The possible observer paths were discretised and a Dynamic Programming (DP) algorithm was provided to compute the optimal discretised path. In [11], the authors only assumed Markovian target motion and bearings measurements. By discretising the target and observation equation, the OTP problem was posed as a POMDP and a sub-optimal closed-loop controller was obtained.

Contributions: 1) The methodology proposed in this paper does not assume linear Gaussian dynamics for the target and observation. We only assume a Markovian target and an observation process that admits a differentiable density; see Section II for a precise statement. Algorithms are presented in full generality such that they may be applied to solve any problem of the form (1), subject to the assumptions stated above, and not just trajectory planning problem. 2) In this paper we use Stochastic approximation (SA) to construct the optimal observer trajectory. Being an iterative gradient descent method, SA requires estimates of the gradient of the performance criterion (1) w.r.t. to the observer trajectory. In our general setting, there no closed form expression for π_k, the performance criterion (1) or its gradient w.r.t. A_k. We demonstrate how low variance estimates of the gradient may be obtained by using Sequential Monte Carlo (SMC), aka Particle Filters, and the *control variate* approach. 3) Because we do not use discretisation, the SA algorithm presented in this paper solves the open-loop version of problem (1) (nearly) exactly. To obtain a closed-loop controller, we use an *open-loop feedback policy*. Note that in general, even the open-loop version of problem (1) cannot be solved exactly because a closed-form expression is not available. Only an approximate solution is possible via discretisation and DP. An iterative gradient method appears to be the only way to solve it exactly. (Detailed comments are provided in Section II-A). Although we demonstrate convergence of the SA (main) algorithm for observer trajectory design in numerical examples, its theoretical convergence has not been established and is the subject of future research.

II. Problem Formulation

Let X_k, A_k and Y_k respectively denote the state of the target, the position of the observer and the observation (target state measurement) received at time k where k denotes a discrete-time index. The target state $\{X_k\}_{k\geq 0}$ is an unobserved Markov process with initial distribution and transition law given by

$$X_0 \sim \pi_0, \quad X_{k+1} \sim p\left(\cdot | X_k\right), \tag{2}$$

respectively. (The symbol "$\sim$" implies distributed according to.) The transition density p does not depend on the observer motion. The observation process $\{Y_k\}_{k\geq 0}$ is generated according to the state

[a]This work was supported in part by CENDSS, ARC and the Defense Advanced Research Projects Agency of the US Department of Defense and was monitored by the Office of Naval Research under Contract No. N00014-02-1-0802.

and observation dependent probability density

$$Y_k \sim q\left(\cdot|X_k, A_k\right). \tag{3}$$

We assume that the observation process admits a density and the density is differentiable w.r.t the second conditioning argument, namely A_k. We place no restrictions on the target motion model accept that the target motion is Markovian. Note that we do not assume a linear Gaussian model.

In this paper, we concentrate on the *jump Markov linear model* (JMLM) for maneuvering targets and a bearings-only measurement process. For this application of the above general framework, the components of the state are

$$X_k = [r_{x,k}, v_{x,k}, r_{y,k}, v_{y,k}, \theta_k]^T \in \mathbb{R}^4 \times \Theta =: \mathbb{X} \tag{4}$$

where $(r_{x,k}, r_{y,k})$ denotes the target's (Cartesian) coordinates, $(v_{x,k}, v_{y,k})$ denotes the target velocity in the x and y direction and θ_k denotes the mode of the target, which belongs to the finite set Θ. (The subscript k indicates time k.) The state of the target is comprised of continuous and discrete valued variables and is used to model maneuvering targets – the target switches between models as indicated by θ_k, between constant velocity maneuvers. Let $\tilde{X}_k \in \mathbb{R}^4$ be the continuous valued components of X_k, i.e. $X_k = \left[\tilde{X}_k^T, \theta_k\right]^T$. In the JMLM, $\{\theta_k\}_{k\geq 0}$ is a time-homogeneous Markov chain while

$$\tilde{X}_{k+1} = F(\theta_{k+1})\tilde{X}_k + G(\theta_{k+1})W_k \tag{5}$$

where $\{W_k\}_{k\geq 0}$ is a sequence of independent and identically distributed (i.i.d.) Gaussian random vectors taking values in $\mathbb{R}^2$, $W_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, Q)$, and represents the uncertainty of the acceleration in the x and y direction.

The state of the observer is similarly defined:

$$X_k^o = \left[r_{x,k}^o, v_{x,k}^o, r_{y,k}^o, v_{y,k}^o\right]^T \tag{6}$$

with

$$A_k = \left[r_{x,k}^o, r_{y,k}^o\right]^T. \tag{7}$$

The observation process $\{Y_k\}_{k\geq 0}$ is generated according to the state and observation dependent probability density (3).In the context of bearings-only tracking,

$$Y_k = \arctan\left(\frac{r_{x,k} - A_k(1)}{r_{y,k} - A_k(2)}\right) + V_k \tag{8}$$

where $V_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_Y^2)$, in which case

$$q\left(y|X_k, A_k\right) = \frac{1}{\sigma_Y\sqrt{2\pi}} \exp\left[-\frac{\left(y - \operatorname{atan}\left(\frac{r_{x,k}-A_k(1)}{r_{y,k}-A_k(2)}\right)\right)^2}{2\sigma_Y^2}\right]. \tag{9}$$

($z(i)$ denotes the i–th component of the vector z).

In this paper, we will assume a simple linear model for the observer motion X_k^o, $k = 0, 1, \ldots$, namely

$$X_{k+1}^o = F^o X_k^o + G^o U_{k+1} \tag{10}$$

where $U_{k+1} := [u_{x,k+1}, u_{y,k+1}]^T$ and X_0^o is given. The variables $u_{x,k+1}$ and $u_{y,k+1}$ denote the acceleration in the x and y direction respectively is the subject of control.

We now state the optimal OTP problem. Assume a suitable the function $\psi : \mathbb{R}^4 \times \Theta \to \mathbb{R}$ is given (e.g., ψ could pick out a component of interest of the state vector). Consider a fixed initial observer state X_0^o and initial target distribution π_0, as well as a fixed sequence of acceleration $U_{1:N} = \{U_1, U_2, \ldots, U_N\}$. The observer trajectory is completely determined, using (7, 10), once X_0^o and $U_{1:N}$ are given and

$$A_k = \begin{bmatrix} 1\,0\,0\,0 \\ 0\,0\,1\,0 \end{bmatrix}\left[(F^o)^k X_0^o + \sum_{i=1}^{k} (F^o)^{k-i} G^o U_i\right]. \tag{11}$$

The optimal OTP problem is to solve

$$\begin{aligned} &\min_{A_{1:N}} \mathbf{E}_{(\pi_0, A_{1:N})}\left\{\sum_{k=1}^{N} \beta^k \left(\psi(X_k) - \langle \pi_k, \psi\rangle\right)^2\right\} \\ &\text{s.t. } (11), \quad U_k \in \mathcal{U}_k, \quad 1 \leq k \leq N, \end{aligned} \tag{12}$$

where N is the *horizon* of interest, $\beta \in (0, 1)$ is a discount factor, π_k is the filtering density at time k (see (13) below), $\langle \pi_k, \psi\rangle$ denotes $\int \psi(x)\pi_k(x)dx$, and $\mathcal{U}_k \subset \mathbb{R}^2$. Note that we are solving for $A_{1:N}$ in terms of the accelerations $U_k = [u_{x,k}, u_{y,k}]^T$. $\mathcal{U}_k$ bounds the x and y components of the acceleration as determined by the physical limitations of the observer. Note that we are assuming a known (fixed) initial observer state X_0^o and target distribution π_0; hence the subscript $(\pi_0, A_{1:N})$ in the expectation operator. (See (16) below for details of the probability density w.r.t which the expectation $\mathbf{E}_{(\pi_0, A_{1:N})}\{\cdot\}$ is taken.)

The aim is to optimise tracking performance with respect to the observer trajectory described by variables $A_{1:N}$. The observer locations $A_{1:N}$ are completely determined once X_0^o and $U_{1:N}$ are specified, (11). The independent variables of the optimisation problem are $U_{1:N}$, which themselves are subject to bounds.

Feedback control: The OTP problem stated in (12) is an open-loop stochastic control problem. In order to utilise feedback, we will use the *open-loop feedback control* (OLFC) approach [1]. Let $U_{1:N}^*$ be the solution to (12). The control applied at epoch 1 is then U_1^*, for which an observation Y_1 is received and the filtering density is updated to π_1. At epoch k, $1 < k \leq N$, let π_{k-1} be the filtering density (density of X_{k-1}) corresponding to all controls taken and observations received up to (and including) epoch $k-1$, $\{U_1, Y_1, \ldots, U_{k-1}, Y_{k-1}\}$, and let the state of the observer be X_{k-1}^o. The filtering density satisfies the *Bayes* recursion

$$\pi_k(x) = \frac{q\left(Y_k|x, A_k\right)\int p(x|x')\pi_{k-1}(x')dx'}{\int\int q\left(Y_k|x, A_k\right)p(x|x')\pi_{k-1}(x')dx'dx}. \tag{13}$$

Now solve problem (12) for initial target distribution π_{k-1}, observer state X_{k-1}^o and horizon $N-k$, and let the solution (with an abuse of notation) be denoted by $U_{k:N}^*$. The control for epoch k is U_k^*. This procedure is repeated until epoch N. The propagation of the filtering density π_{k-1} can be done using SMC as detailed in [2].

Certain key issues concerning problem (12) are now discussed. Firstly, we place no restrictions on the target motion model accept as stated in (2), i.e. the target motion is Markovian. We do not assume a linear Gaussian model. The specific model in (5) is used only to provide a concrete scenario for the numerical example. Similarly, we do not assume a linear Gaussian observation processes. The only requirement on the observation process is that it has a known density (3) which is differentiable w.r.t the second conditioning argument (A_k). The bearings only case in (9) is one example of an observation density that satisfies these assumptions. Under these general conditions, the filtering density π_k cannot be represented in closed form. Hence, we will resort to a SMC method [2] in order to evaluate π_k and the criterion in (12). The algorithms in this paper are presented in enough generality such that they may be applied to solve any problem of the form (12), subject to the assumptions stated above. Finally, we remark that the gradient method proposed can be used to solve general state space POMDPs where the choice

of control also influences the dynamics of the state process, as was done in [3].

A. Motivation for Gradient Methods and OLFC

Problem (12) is a Partially Observed Markov Decision Process (POMDP) with a continuous state space. Because of the absence of a closed-form expression, standard state-space discretisation appears to be the only way towards numerical implementation. In discretisation, the target, observation and control state-space is discretised to obtain a *finite* POMDP. This was the approach adopted in [11]. As an example, consider a target state comprised of a continuous and discrete component, $(x, \theta) \in \mathbb{R}^d \times \Theta$. If we discretise each component of x to L values then, the discretised target state-space has $L^d|\Theta|$ elements! The observation and control spaces also need to be discretised. The problem is compounded even further by observer path constraints. The previous location of the observer has to be augmented to the state descriptor to yield states (x_k, θ_k, A_{k-1}) where A_{k-1} is the location of the observer at the previous epoch $k-1$. However, observer trajectory constraints are handled nicely by a gradient method as one only needs to do a projection of the computed gradient to ensure satisfaction of the path constraints; see Section III-A.

Although the solution for a finite POMDP can be obtain exactly, it is often too computationally intensive. It is well known that the controller (or policy) of a finite horizon finite POMDP is characterised by a finite set of vectors [10]. However, the number of vectors needed to characterised the optimal policy grows exponentially with the horizon and computing these vectors may be computationally infeasible for even values of $N = 5$ or greater. In practise, various *pruning* methods are employed [7], which effectively amounts to another level of discretisation.

The open-loop problem (12) cannot be solved exactly because a closed-form expression for the criterion is not available. Only an approximate solution is possible via discretisation (as described above) and DP. An iterative gradient method appears to be the only way to solve it exactly.

III. Gradient of the OTP Criterion

Stochastic approximation is a stochastic gradient descent method. Let $J(A_{1:N})$ denote the cost function to be minimised in (12). We will use the following iterative procedure

$$A_{1:N}^{(k+1)} = A_{1:N}^{(k)} - \alpha_k \left(\nabla J(A_{1:N})|_{A_{1:N}=A_{1:N}^{(k)}} + \text{noise} \right) \tag{14}$$

where $\nabla J(A_{1:N})$ denotes the gradient of J w.r.t $A_{1:N}$. (Actually, we have to perform a projection step in (14) to ensure the constraints are satisfied.) Once again, we do not have a closed-form expression for ∇J for the same reasons as in J – the filtering density π_k and integration with respect to it cannot be evaluated in close-form in our general setting. In this section we will show how one may obtain an asymptotically unbiased estimates of ∇J, denoted by $\widehat{\nabla J}$, via the the so called *score-function* method [9]. The score-function methods gives an unbiased estimate of ∇J but unfortunately with high variance. We propose and an *adaptive control variate method* to reduce the variance of $\widehat{\nabla J}$ by coupling with (14) a second stochastic iteration that estimates the optimal control variate for $\widehat{\nabla J}$. We then demonstrate in numerical examples that the variance of $\widehat{\nabla J}$ is reduced by several orders of magnitude and the convergence of (14) to the minimising solution is rapid.

A. Derivation of the OTP Criterion Gradient

In this section, we derive the gradient of the cost function (12) with respect to $A_{1:N}$. Henceforth, without loss of generality, we assume a fixed initial observer state X_0^o and initial target distribution π_0. Also, to simplify exposition, *we consider the problem of optimising the tracking performance at time N only*. Define the mapping $J : \left(\mathbb{R}^2\right)^N \to \mathbb{R}$ as

$$J\left(A_{1:N}\right) := \mathbf{E}_{(\pi_0, A_{1:N})}\left\{ \left(\psi(X_N) - \langle \pi_N, \psi \rangle\right)^2 \right\}. \tag{15}$$

We call J the *OTP criterion* and the omission of (π_0, X_0^o) in the definition of J follows from the opening sentences of this paragraph. For any integrable function $h : \mathbb{X}^N \times \left(\mathbb{R}^2\right)^N \times \mathbb{R}^N \to \mathbb{R}$,

$$\mathbf{E}_{(\pi_0, A_{1:N})}\left\{h(X_{1:N}, A_{1:N}, Y_{1:N})\right\} = \int h(x_{1:N}, A_{1:N}, y_{1:N}) \times \Pi_{i=1}^N q\left(y_i|x_i, A_i\right) p\left(x_i|x_{i-1}\right) \pi_0(x_0) dx_{1:N} dy_{1:N} dx_0 \tag{16}$$

The integral in (16) follows by definition of the dynamics in (2) and (3). Note that the integrand in (15) is indeed some function with the same domain as h above; this follows since π_k (13) is a function of $(A_{1:k}, Y_{1:k})$ only. In the unconstrained setting, we solve

$$\min_{A_{1:N} \in \left(\mathbb{R}^2\right)^N} J(A_{1:N}) \tag{17}$$

instead of (12) since the observer can move between any 2 points on the xy-plane in one epoch.

For a function $f : \mathbb{R}^n \to \mathbb{R}$ with arguments $z \in \mathbb{R}^n$, we denote $\left(\partial f / \partial z(i)\right)(z)$ by $\nabla_{z(i)} f(z)$. For a fix $(y, x) \in \mathbb{R} \times \mathbb{X}$, consider the mapping $a \left(\in \mathbb{R}^2\right) \to q(y|x, a) \in \mathbb{R}$ Define the *score* [9] as follows:

$$S_i(y, x, a) := \frac{\nabla_{a(i)} q(y|x, a)}{q(y|x, a)}, \quad i \in \{1, 2\}. \tag{18}$$

For any $\widetilde{A}_{1:N} \in \left(\mathbb{R}^2\right)^N$, it may be shown by taking the derivative inside the integral in (15) and applying the product rule, that

$$\nabla_{i,j} J(\widetilde{A}_{1:N}) := \nabla_{A_i(j)} J(\widetilde{A}_{1:N}) = \mathbf{E}_{(\pi_0, \tilde{A}_{1:N})}\left\{ \left(\psi(X_N) - \langle \pi_N, \psi \rangle\right)^2 S_j(Y_i, X_i, \tilde{A}_i) \right\}, \tag{19}$$

$j \in \{1, 2\}$, $i = 1, 2, \ldots, N$. We define

$$\nabla J(A_{1:N}) := \left[\begin{bmatrix} \nabla_{1,1} J(A_{1:N}) \\ \nabla_{1,2} J(A_{1:N}) \end{bmatrix}^T, \ldots, \begin{bmatrix} \nabla_{N,1} J(A_{1:N}) \\ \nabla_{N,2} J(A_{1:N}) \end{bmatrix}^T \right]^T. \tag{20}$$

Let the mapping from $U_{1:N} \to A_{1:N}$ in (11) be denoted H, i.e. $A_{1:N} = H(U_{1:N})$. Problem (12) can be written as

$$\min_{U_{1:N} \in \left(\mathbb{R}^2\right)^N} J(H(U_{1:N})), \quad \text{s.t. } U_k \in \mathcal{U}_k, 1 \le k \le N. \tag{21}$$

If $\nabla J(A_{1:N})$ is known then, the gradient of J w.r.t. $U_{1:N}$ may be obtained by an application of the chain rule since ∇H is known precisely. Therefore, we concentrate on the problem of deriving a low variance estimator for $\nabla J(A_{1:N})$ only.

B. Monte Carlo (MC) Approximation to ∇J

The gradient in (19) cannot be computed analytically and we will resort to MC.

By Bayes' rule, the density of $X_{0:N}$ given $Y_{1:N}$ and $A_{1:N}$ is

$$\pi_{0:N}(x_{0:N}) = \frac{\left(\prod_{i=1}^N q\left(Y_i|x_i, A_i\right) p\left(x_i|x_{i-1}\right)\right) \pi_0(x_0)}{\int \left(\prod_{i=1}^N q\left(Y_i|x_i, A_i\right) p\left(x_i|x_{i-1}\right)\right) \pi_0(x_0) dx_{0:N}}. \tag{22}$$

Assume that we have L i.i.d. target trajectory samples, $\{X_{0:N}^{(j)}\}_{j=1}^{L}$, simulated from $\left(\Pi_{i=1}^{N} p\left(x_i|x_{i-1}\right)\right)\pi_0(x_0)$. Then, for any integrable function of interest $h : \mathbb{X}^{N+1} \to \mathbb{R}$,

$$\int h(x_{0:N})\pi_{0:N}(x_{0:N})dx_{0:N} \approx \int h(x_{0:N})\hat{\pi}_{0:N}(x_{0:N})dx_{0:N}. \tag{23}$$

where

$$\hat{\pi}_{0:N}(x_{0:N}) := \sum_{j=1}^{L} w_N^{(j)} \delta_{X_{0:N}^{(j)}}(x_{0:N}), \tag{24}$$

$\delta_{X_{0:N}^{(j)}}$ denotes the Dirac delta-mass located at $X_{0:N}^{(j)}$ and $w_N^{(j)}$ are the *importance weights* defined as

$$w_N^{(j)} := \prod_{i=1}^{N} q(Y_i|X_i^{(j)}, A_i)[\sum_{j=1}^{L}\prod_{i=1}^{N} q(Y_i|X_i^{(j)}, A_i)]^{-1}. \tag{25}$$

The MC approximation in (23) is asymptotically unbiased [2]. Henceforth, $\hat{\pi}_{0:N}$ is to be understood as the MC approximation to posterior distribution $\pi_{0:N}$.

1) Variance Reduction by Conditioning : In order to obtain an unbiased estimate of the gradient $\nabla_{A_i(j)}J(\widetilde{A}_{1:N})$, we should sample a target trajectory $X_{0:N}$ and the corresponding observations $Y_{1:N}$ according to the law $\left(\Pi_{i=1}^{N} q\left(Y_i|x_i, \tilde{A}_i\right) p\left(x_i|x_{i-1}\right)\right)\pi_0(x_0)$. Then

$$\left(\psi(X_N) - \langle \pi_N, \psi\rangle\right)^2 S_j(Y_i, X_i, \tilde{A}_i)$$

is an unbiased estimate of $\nabla_{A_i(j)}J(\widetilde{A}_{1:N})$. However, since we cannot compute π_N in the general state space setting, we are compelled to use

$$\left(\psi(X_N) - \langle \hat{\pi}_N, \psi\rangle\right)^2 S_j(Y_i, X_i, \tilde{A}_i) \tag{26}$$

instead, where $\hat{\pi}_N$ is the MC approximation to the filtering density, which is the marginal of the MC approximation to the posterior density given in (24). To reduce the variance of (26), we may exploit the fact that $var\{\mathbf{E}(X|Y)\} \leq var\{X\}$ where Y is some other random variable correlated with X. Now, since we have the posterior $\hat{\pi}_{0:N}$, we may use

$$\begin{aligned}&\mathbf{E}_{\left(\pi_0,\tilde{A}_{1:N}\right)}\left\{\left(\psi(X_N) - \langle \pi_N, \psi\rangle\right)^2 S_j(Y_i, X_i, \tilde{A}_i)\middle| Y_{1:N}\right\}\\ &= \int \left(\psi(x_N) - \langle \pi_N, \psi\rangle\right)^2 S_j(Y_i, x_i, \tilde{A}_i)\pi_{0:N}(x_{0:N})dx_{0:N}\end{aligned} \tag{27}$$

with $\hat{\pi}_N$ and $\hat{\pi}_{0:N}$ replacing π_N and $\pi_{0:N}$ respectively as the reduced variance estimate of $\nabla_{A_i(j)}J(\widetilde{A}_{1:N})$, where $Y_{1:N}$ was the sampled observation trajectory. (Note that $\langle \pi_N, \psi\rangle$ is a function $Y_{1:N}$, i.e. $\sigma(Y_{1:N})$ measurable.)

2) Variance reduction by Control Variate: Consider 2 correlated random variables W and Z where $\mathbf{E}(Z) = 0$. We would like to estimate $\mathbf{E}(W)$ for which W is an unbiased estimate. For some constant b, the estimator $W - bZ$ is also unbiased. The function

$$\begin{aligned}f(b) :=& var(W - bZ)\\ &= var(W) - 2bcov(W, Z) + b^2 var(Z)\end{aligned} \tag{28}$$

is convex and is minimised at $b^* = cov(W, Z)/var(Z)$, which implies $var(W - b^*Z) = f(b^*) < f(0) = var(W)$. Thus, the variance of the estimate $W - b^*Z$ of $\mathbf{E}(W)$ is less than the variance of W. The random variable Z is referred to as the *control variate* (CV) and we call b the CV constant.

We now detail the random variables W and Z in the context to of the gradient of the OTP criterion (19). Let $b_i(j)$, $j \in \{1, 2\}$, $i = 1, 2, \ldots, N$, be real valued constants. It is straightforward to verify that $\mathbf{E}_{\left(\pi_0,\tilde{A}_{1:N}\right)}\{S_j(Y_i, X_i, \tilde{A}_i)\} = 0$. Let $W_i(j) = \mathbf{E}_{\left(\pi_0,\tilde{A}_{1:N}\right)}\{\left(\psi(X_N) - \langle \pi_N, \psi\rangle\right)^2 S_j(Y_i, X_i, \tilde{A}_i)\Big| Y_{1:N}\}$ and $Z_i(j) = \mathbf{E}_{\left(\pi_0,\tilde{A}_{1:N}\right)}\{S_j(Y_i, X_i, \tilde{A}_i)\Big| Y_{1:N}\}$. Thus $\nabla_{A_i(j)}J(\widetilde{A}_{1:N}) = \mathbf{E}_{\left(\pi_0,\tilde{A}_{1:N}\right)}[W_i(j) - b_i(j)Z_i(j)]$. The optimal CV constant is

$$b_i^*(j) := \arg\min_b -2bcov(W_i(j), Z_i(j)) + b^2 var(Z_i(j)) \tag{29}$$

We may solve for $b_i^* = [b_i^*(1), b_i^*(2)]^T \in \mathbb{R}^2$, $i = 1, 2, \ldots, N$, using stochastic approximation, as presented in the algorithm below.

To summarise the ideas presented in sections III-B.1 and III-B.2, we now present an algorithm for estimating ∇J and the optimal CV constants $b_i^*(j)$ (29).

Algorithm : Estimating $\nabla J(\tilde{A}_{1:N})$ and optimal CV constants b_i^*, $i = 1, 2, \ldots, N$.

- Initialisation: Fix π_0 , $\tilde{A}_{1:N}$. Set $b_i^{(0)} = [0, 0]^T \in R^2$, $i = 1, 2, \ldots, N$. Select a non-increasing positive sequence $\{\alpha_k\}_{k\geq 1}$ satisfying $\sum \alpha_k = \infty$, $\sum \alpha_k^2 < \infty$.
- Iteration k ($k \geq 1$):
  Sample a target trajectory $X_{0:N}$ and observations $Y_{1:N}$ according to the law
  $\left(\Pi_{i=1}^{N} q\left(Y_i|x_i, \tilde{A}_i\right) p\left(x_i|x_{i-1}\right)\right)\pi_0(x_0)$ and compute $\hat{\pi}_{0:N}$.
  Update : for $j \in \{1, 2\}$, $i = 1, 2, \ldots, N$

$$\begin{aligned}\nabla J_i^{(k)}(j) =& \int \left(\left(\psi(x_N) - \langle \hat{\pi}_N, \psi\rangle\right)^2 - b_i^{(k-1)}(j)\right)\\ &\times S_j(Y_i, x_i, \tilde{A}_i)\hat{\pi}_{0:N}(x_{0:N})dx_{0:N},\end{aligned} \tag{30}$$

$$\begin{aligned}b_i^{(k)}(j) =& b_i^{(k-1)}(j) - \alpha_k\Bigg[b_i^{(k-1)}(j)\\ &\times \left(\int S_j(Y_i, x_i, \tilde{A}_i)\hat{\pi}_{0:N}(x_{0:N})dx_{0:N}\right)^2\\ &- \Bigg(\int \left[\left(\psi(x_N) - \langle \hat{\pi}_N, \psi\rangle\right)^2 S_j(Y_i, x_i, \tilde{A}_i)\right]\\ &\times \hat{\pi}_{0:N}(x_{0:N})dx_{0:N}\Bigg)\\ &\times \left(\int S_j(Y_i, x_i, \tilde{A}_i)\hat{\pi}_{0:N}(x_{0:N})dx_{0:N}\right)\Bigg]\end{aligned} \tag{31}$$

$b_i^{(k)}(j)$ and $\nabla J_i^{(k)}(j)$ are, respectively, the estimates of $b_i^*(j)$ and $\nabla_{A_i(j)}J(\widetilde{A}_{1:N})$ at iteration k. A typical choice for $\{\alpha_k\}_{k\geq 1}$ is $\alpha_k = k^{-\alpha}$, where $\alpha \in (0.5, 1]$ is a constant. The right-hand side (integral) of $\nabla J_i^{(k)}(j)$ is the unbiased estimate of the gradient given in (27), which is computed using MC. The recursion for $b_i^{(k)}(j)$ is a SA algorithm that estimates the optimal CV constants in (29). The term multiplying α_k in the right-hand side of (31) is the gradient of the criterion in (28), taken w.r.t. $b_i^{(k-1)}(j)$, computed using the MC. Finally, note that the update for $b_i^{(k)}(j)$ and $\nabla J_i^{(k)}(j)$ for each i and j can be carried out in parallel as there no dependence between terms.

The rationale for the above algorithm is as follows. As $b_i^{(k)}(j)$ converges to $b_i^*(j)$, the variance of the gradient estimates $\nabla J_i^{(k)}(j)$ decreases. We expect $1/k\sum_{l=1}^{k}\nabla J_i^{(l)}(j)$ to converge (with probability one) more quickly to its limiting value than the empirical mean of (30) with $b_i^{(k-1)}(j)$ set to 0, as we are averaging terms that have smaller variance than the former case. Typical performance of the above algorithm is shown in Section V in the numerical examples.

IV. Main Algorithm: Stochastic Approximation for Optimal OTP

We present a SA algorithm for solving problem (12) or equivalently, the reformulated problem (21). Note that the OTP criterion

optimised is not the weighted sum of the filtering error in (12) but rather the filtering error at time N only (15). The same approach applies to the more general criteria in (12).

For a sequence of accelerations $U_{1:N} \in \left(\mathbb{R}^2\right)^N$, let $P(U_{1:N})$ denote the projection of $U_{1:N}$ onto the feasible set $\prod_{k=1}^N \mathcal{U}_k$.

Algorithm: Stochastic Approximation for Problem (21).

- Initialisation: Fix π_0, X_0^o. Pick initial feasible $U_{1:N}^{(0)}$. Set $b_i^{(0)} = \nabla J_i^{(0)} = [0,0]^T \in \mathbb{R}^2$, $i = 1, 2, \ldots, N$. Select a non-increasing positive sequences $\{\alpha_k\}_{k\geq 1}$, $\{\beta_k\}_{k\geq 1}$, satisfying $\sum \alpha_k = \infty$, $\sum \alpha_k^2 < \infty$, $\sum \beta_k = \infty$, $\sum \beta_k^2 < \infty$, $\limsup \beta_k/\alpha_k < \infty$.
- Iteration k ($k \geq 1$):
  Set $A_{1:N}^{(k-1)} = H(U_{1:N}^{(k-1)})$. Sample a target trajectory $X_{0:N}$ and observations $Y_{1:N}$ according to the law $\left(\Pi_{i=1}^N q\left(Y_i|x_i, A_i^{(k-1)}\right) p\left(x_i|x_{i-1}\right)\right) \pi_0(x_0)$ and compute $\hat{\pi}_{0:N}$.
  Update

$$\begin{aligned}
\nabla J_i^{(k)}(j) = &\int \left(\left(\psi(x_N) - \langle \hat{\pi}_N, \psi\rangle\right)^2 - b_i^{(k-1)}(j)\right) \\
&\times S_j(Y_i, x_i, A_i^{(k-1)}) \hat{\pi}_{0:N}(x_{0:N}) dx_{0:N}, \qquad (32)\\
b_i^{(k)}(j) = &\, b_i^{(k-1)}(j) - \alpha_k \Bigg[b_i^{(k-1)}(j) \\
&\times \left(\int S_j(Y_i, x_i, A_i^{(k-1)}) \hat{\pi}_{0:N}(x_{0:N}) dx_{0:N}\right)^2 \\
&- \left(\int \left[\left(\psi(x_N) - \langle \hat{\pi}_N, \psi\rangle\right)^2 S_j(Y_i, x_i, A_i^{(k-1)})\right]\right. \\
&\left.\times \hat{\pi}_{0:N}(x_{0:N}) dx_{0:N}\right) \\
&\times \left(\int S_j(Y_i, x_i, A_i^{(k-1)}) \hat{\pi}_{0:N}(x_{0:N}) dx_{0:N}\right)\Bigg] \qquad (33)
\end{aligned}$$

  $j \in \{1,2\}$, $i = 1, 2, \ldots, N$.

$$\tilde{U}_{1:N}^{(k)} = U_{1:N}^{(k-1)} - \beta_k \nabla H(U_{1:N}^{(k-1)}) \nabla J^{(k)}, \qquad (34)$$

$$U_{1:N}^{(k)} = P(\tilde{U}_{1:N}^{(k)}). \qquad (35)$$

In (34), $\nabla J^{(k)} = [(\nabla J_1^{(k)})^T, \ldots, (\nabla J_N^{(k)})^T]^T$, which is the estimate of the gradient of OTP criterion at $A_{1:N}^{(k-1)}$. Note that the exact expression for $\nabla H(U_{1:N}^{(k-1)})$ is known. A typical choice for the step-sizes are $\alpha_k = k^{-\alpha}$, $\beta_k = k^{-\beta}$, $\alpha, \beta \in (0.5, 1]$ and $\beta > \alpha$. In which case, $\lim \beta_k/\alpha_k = 0$. Because β_k tends to 0 more quickly than α_k, the recursion (34) is said to evolve on a slower time-scale than (33). By having $U_{1:N}^{(k)}$ evolve more slowly than $b_i^{(k)}(j)$, we allow the $b_i^{(k)}(j)$ recursions to "track" the optimal CV constants in (29) which depend on the point at which the gradient ∇J of OTP criterion is evaluated. In the numerical example presented in Section V, we use constant step-sizes for $\alpha_k = \alpha$ and $\beta_k = \beta$ and demonstrate "convergence". For SA in general, decreasing step-sizes are essential for w.p.1 convergence. If fixed step-sizes are used, then we may still have convergence but now the iterates "oscillate" about their limiting values with variance proportional to the step-size. In our case, theoretical convergence is more difficult to establish since we have 2 coupled SA algorithms using biased estimates of the gradients of interest. See [9] for theoretical convergence issues of basic SA.

V. Numerical Example

In the numerical examples below, we demonstrate the variance reduction achieved in the gradient estimator as well as the convergence of the SA algorithm to the solution of the unconstrained OTP problem.

Consider the following problem of tracking a target with a one dimensional state with bearings only measurements:

$$\begin{aligned}
X_{k+1} &= X_k + 1.5 + \sigma_X W_k, \quad X_0 = 0, \\
Y_k &= \arctan\left(X_k - A_k\right) + \sigma_Y V_k, \qquad (36)
\end{aligned}$$

where $W_k, V_k (\in R) \sim N(0,1)$ are independent and $A_k \in R$ is the position of the observer at epoch k. At each epoch, the target advances its position by an amount equal to the sum of 1.5 and a zero-mean Gaussian random variable. The target evolves on the x-axis and the observation equation corresponds to an observer whose y-position is fixed at -1 but is allowed to alter its x-position. We assume an otherwise unconstrained observer and the criterion being minimised is (cf. (17))

$$J\left(A_{1:N}\right) := \mathbf{E}_{(\pi_0, A_{1:N})}\left\{\left(\psi(X_N) - \langle \pi_N, \psi\rangle\right)^2\right\}$$

w.r.t. $A_{1:N} \in R^N$. All the simulations below were for $\sigma_X = 0.5$, $\sigma_Y = 1$ and $N = 3$. The posterior density $\pi_{0:N}$ was estimated using MC with $L = 1000$.

We first demonstrate the convergence of the algorithm presented in Section III-B.2 to the optimal CV constants as well as the reduction of variance in the estimate of the OTP criterion gradient. Figure 1 depicts the result of the simulation for $\tilde{A}_{1:3} = [0,0,0]^T$ and constant step-size $\alpha_k = 0.1$. The simulations were performed for 4500 iterations. Note that the iterates $(b_1^{(k)}, b_2^{(k)}, b_3^{(k)}) \to (0.8, 0.9, 0.9)$ and then oscillate. The oscillations are due to the constant step-size. The vastly different convergence rates are explained below.

The 4500 realisations of target trajectory $X_{0:N}$ and observations $Y_{1:N}$ (generated according to the law $\left(\Pi_{i=1}^N q\left(Y_i|x_i, \tilde{A}_i\right) p\left(x_i|x_{i-1}\right)\right) \pi_0(x_0)$), as well as the MC approximation $\hat{\pi}_{0:N}$ used to generate Figure 1 were stored. For a fixed $b \in R$ and realisation k, define (cf. (30))

$$\begin{aligned}
\nabla J_i^{b,(k)} \quad := \quad &\int \left[\left(\left(\psi(x_N) - \langle \hat{\pi}_N, \psi\rangle\right)^2 - b\right) S(Y_i, x_i, \tilde{A}_i)\right] \\
&\times \hat{\pi}_{0:N}(x_{0:N}) dx_{0:N}. \qquad (37)
\end{aligned}$$

The empirical variance, $var(b,i) := \text{var}\{\nabla J_i^{b,(k)} : k = 1, \ldots, 4500\}$, $i = 1,2,3$, is plotted on Figure 2 for different values of $b \in [0, 1.5]$ on the horizontal axis. As indicated on Figure 2, $var(b,1)$, $var(b,2)$ and $var(b,3)$ are minimised at 0.8, 0.9 and 0.9 respectively, which are precisely the converged values of iterates $(b_1^{(k)}, b_2^{(k)}, b_3^{(k)})$. Also, the shallow curve $var(b,3)$ explains the slow convergence of $b_3^{(k)}$ in Figure 1. At the values converged values of $(b_1^{(k)}, b_2^{(k)}, b_3^{(k)})$, the variance in the gradient estimate was reduced by more than 100 fold. Specifically, $var(0.8,1)/var(0,1) \approx 260$, $var(0.9,2)/var(0,2) \approx 170$ and $var(0.9,3)/var(0,3) \approx 150$.

We now demonstrate the convergence of the observer trajectory iterates generated by the algorithm presented in Section IV. Without motion constraints, we execute the following iteration instead of (34):

$$A_{1:N}^{(k)} = A_{1:N}^{(k-1)} - \beta_k \nabla J^{(k)}.$$

The simulations below were performed for $N = 3$, $A_{1:N}^{(0)} = [0, \ldots, 0]^T$ and constant step-sizes $\alpha_k = .0.3$ and $\beta_k = 0.1$. The iterates $A_{1:N}^{(k)}$ are depicted in Figure 3. In Figure 4, the corresponding CV constant iterates $(b_1^{(k)}, b_2^{(k)}, b_3^{(k)})$ are plotted. The initial values $(b_1^{(0)}, b_2^{(0)}, b_3^{(0)})$ were $(0.8, 0.9, 0.9)$, which were the optimal CV constants for estimator of $\nabla J(A_{1:3}^{(0)} = [0,0,0]^T)$ identified from the previous simulation. The final values of iterates $(b_1^{(k)}, b_2^{(k)}, b_3^{(k)})$ were the optimal values for $\nabla J(A_{1:3} = [1.5, 3, 4.5]^T)$. The optimal

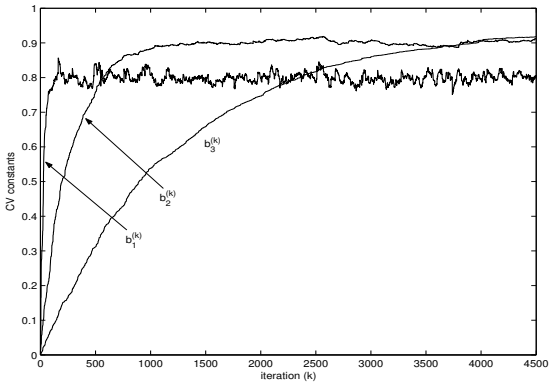

Fig. 1. Convergence of $(b_1^{(k)}, b_2^{(k)}, b_3^{(k)})$ in (31).

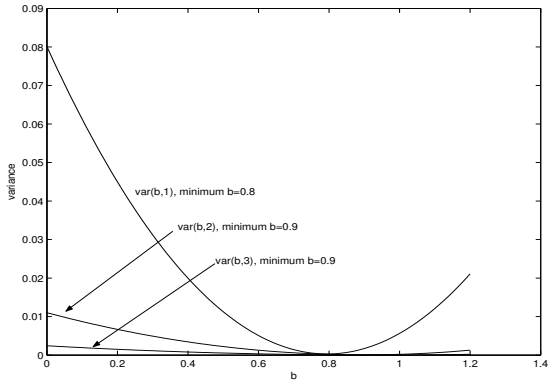

Fig. 2. Empirical variance (37) for different values of CV constants.

observer locations are $(1.5, 3, 3.5)$, which is precisely the expected trajectory of the target.

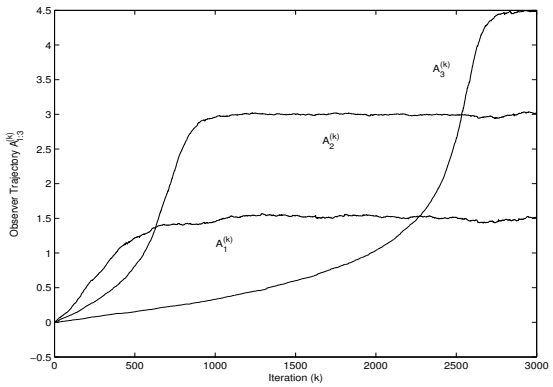

Fig. 3. Convergence of observer trajectory.

VI. Conclusion

We have presented a stochastic approximation algorithm for optimal observer trajectory planning. Our approach only requires the target motion to be Markovian. The only restriction on the observation process is that its law (3) admits a "known" density and the density is differentiable w.r.t. the second conditioning argument (A_k). Under these general conditions, the filtering density π_k cannot be evaluated in closed form. Hence, we resorted to Sequential Monte Carlo methods. We derived the standard score-gradient estimator for

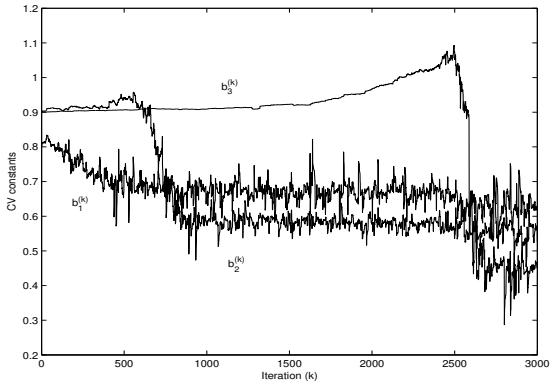

Fig. 4. Tracking of optimal CV constants.

the performance criterion. We proposed an adaptive control variate method to reduce the variance of $\widehat{\nabla J}$ by introducing a second stochastic iteration that estimates the optimal control variate for $\widehat{\nabla J}$. We then demonstrated in simple numerical examples that the variance of $\widehat{\nabla J}$ is reduced by about two orders of magnitude and convergence of the observer trajectory to the minimising solution is rapid. In future work, we aim to investigate open-loop feedback control for adaptive trajectory optimisation.

VII. References

[1] D.P. Bertsekas, Dynamic programming and optimal control. Belmont: Athena Scientific, 1995.

[2] A.Doucet, J.F.G. de Freitas and N.J. Gordon Sequential Monte Carlo methods in practice. New York: Springer, 2001.

[3] A. Doucet, V. Tadić and S. Singh, "Particle methods for average cost control in nonlinear non-Gaussian state-space models," Technical report, Department of Electrical and Electronic Engineering, University of Melbourne, June 2002.

[4] D.J. Kershaw and R.J. Evans, "Optimal waveform selection for target tracking," IEEE Trans. Inform. Theory, vol. 40, no. 5, pp. 1536–1550, Sep. 1994.

[5] O. Hernandez-Lerma, Adaptive Markov Control Processes. New York: Springer, 1989.

[6] A. Logothetis, A. Isaksson and R.J. Evans "An information theoretic approach to observer path design for bearings-only tracking," in IEEE Conf. Decision Control, 1997, pp. 3132–3137.

[7] W.S. Lovejoy, "A survey of algorithmic methods for partially observed Markov decision processes," Annals Oper. Res., vol. 28, pp. 47–66, 1991.

[8] L. Meier, J. Perschon and R.M. Dressler, "Optimal control of measurement systems," IEEE Trans. Automat. Contr., vol. 12, no. 5, pp. 528–536, Oct. 1967.

[9] G.Ch. Pflug, Optimization of stochastic models: the interface between simulation and optimization. Boston: Kluwer, 1996.

[10] R.D. Smallwood and E.J. Sondik, "The optimal control of partially observable Markov processes over a finite horizon," Oper. Res., vol. 21, pp. 1071–1088, 1973.

[11] O. Tremois and J.-P. LeCadre, "Optimal observer trajectory in bearings-only tracking for manoeuvring sources," IEE Proc. Radar, Sonar Navig., vol. 146, no. 1, pp. 31–39, Feb. 1999.