2008 American Control Conference
Westin Seattle Hotel, Seattle, Washington, USA
June 11-13, 2008

Optimal trade-off between exploration and exploitation

Alex Simpkins†, Raymond de Callafon†, and Emanuel Todorov‡

Abstract— **Control in an uncertain environment often involves a trade-off between exploratory actions, whose goal is to gather sensory information, and "regular" actions which exploit the information gathered so far and pursue the task objectives. In principle both types of action can be modeled by minimizing a single cost function within the framework of stochastic optimal control. In practice however this is difficult, because the control law must be sensitive to estimation uncertainty which violates the certainty-equivalence principle. In this paper we formalize the problem in a way which captures the essence of the exploration-exploitation trade-off and yet is amenable to numerical methods for optimal control. The key to our approach is augmenting the dynamics of the partially-observable plant with the Kalman filter dynamics, thus obtaining a higher-dimensional but fully-observable plant. The resulting control laws compare favorably to other more ad-hoc approaches. Our formalism is also suitable for modeling human behavior in tasks which benefit from active exploration.**

I. INTRODUCTION

Active exploration is a powerful approach for dealing with uncertainty. It is widely used throughout biology - examples include movable eyes, ears and whiskers, fingers used to explore surface properties, and muscle spindles with tunable sensitivity. The importance of active exploration is also increasingly recognized in engineering. The challenge is to incorporate information gains and control actions within the same formalism and define a notion of utility which is equally applicable to both.

In principle, the general framework of stochastic optimal control under uncertainty can be applied. In this framework one uses an optimal/Bayesian estimator which computes the posterior probability density over the state space at each point in time, and an optimal feedback controller which maps probability densities into actions. Although probability densities are infinite-dimensional objects one can rely on finite-dimensional estimators such as the Kalman filter which is optimal in a 2-norm setting. Interestingly the variance of the noise does not affect the optimal control law, thus defeating the purpose of active exploration. This property (known as the separation principle, and controllers based on this are often referred to as certainty equivalence, or CE, controllers) is normally considered a virtue, but from the present perspective it is a deficiency. A modified formalism is needed, where the estimator is still finite-dimensional but the control law is sensitive to uncertainty and generates actions aimed at reducing uncertainty. Here we develop such a formalism.

Active exploration can be classified into two categories. The first includes actions that only affect the flow of sensory feedback, thus having indirect consequences on achieving control objectives. Eye movements are a good example. The second category includes actions that not only affect the flow of sensory feedback but also have direct consequences on control objectives. Finger movements which simultaneously sense and manipulate objects are an example such actions. This second category, which involves an interesting tradeoff between optimization of sensory information and achieving the control objective, is the focus of the present paper.

The concept of control actions reducing uncertainty in unknown parameters through adaptation has previously been studied both from a deterministic and stochastic perspective, and there is a wide literature [18], [6], [2], [9], [8], [12].

From the deterministic side, passive identifiers represent one approach (though dependent on the certainty equivalence principle). Other examples are adaptive backstepping, adaptive Lyapunov design with tuning functions, and modular estimation-based designs. The latter three do not depend on the certainty equivalence principle, but are limited in their ability to remain stable for large nonlinearities. Iterative identification and control [17], [5] is another deterministic approach to producing control while building a model of the plant. However this is an offline optimization and identification; then the parameters of the control remain static once computed. However, they are insufficient here where the parameters have large or rapid variations, and uncertainties may be large.

Exploration and exploitation taking place simultaneously (through online estimation and control) is necessary in these types of problems where the plant is partially-observable and non-stationary. There are adaptive dual control (ADC) techniques which approach this problem by generating a 'cautious' control signal for tracking, with an additional excitation signal that accelerates parameter estimation [18], [2]. A unique characteristic of the present method is that it does not generate a reduced gain (cautious) tracking signal (which ADC controllers do), instead triggering pseudo-random actions when needed which are combined with a nominal tracking signal to produce an optimal action. Additionally, most of the approaches in the ADC literature center around discrete time systems, whereas here we develop an elegant continuous time method (there are some continuous time ADC methods such as the bicritical method but it approximates the original problem, whereas here the original problem is solved). Task-relevant versus task-irrelevant uncertainty, an important but rarely addressed distinction, is also considered in this paper. Finally, it is emphasized that the development of this model for control actions is centered around modeling human exploration/exploitation control actions.

II. EXAMPLE OF EXPLORATION TASK

We develop our method via an example, which is later used to compare the optimal solution to experimental data obtained from human subjects. The task can be thought of as tracking a screen target using an uncertain computer mouse:

This work was supported by the US National Science Foundation

†Department of Mechanical and Aerospace Engineering, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0411, email: csimpkin@ucsd.edu, callafon@ucsd.edu

‡Department of Cognitive Science, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0515, email: todorov@cogsci.ucsd.edu

we do not know how the mouse works and the way the mouse works changes all the time. More precisely, let $h(t) \in \Re^{n_h}$ denote a state variable which is fully observable and controllable (say the position of the hand), and $s(t) \in \Re^{n_s}$ denote a target which needs to be tracked. The unknown mapping from h-space to s-space has a hidden state $m(t) \in \Re^{n_s n_h}$ which undergoes Brownian motion. The mapping is assumed to be a linear projection from hand to target space

$$h \rightarrow M[m]\, h. \tag{1}$$

Here M is a linear operator which reshapes the vector $m(t)$ into an n_s-by-n_h projection matrix.

III. Formulation as a stochastic optimal control problem

The dynamics are assumed linear-Gaussian:

$$dh = u(t)dt, \qquad ds = d\omega_s, \qquad dm = d\omega_m, \tag{2}$$

in which $u(t) \in \Re^{n_h}$ is the control signal which is generated in order to modify the hand state and $\omega_s(t)$ and $\omega_m(t)$ are Brownian motion processes with covariances Ω_s and Ω_m respectively.[1] We assume that $h(t)$ and $s(t)$ are directly observable, while $m(t)$ is not directly observable and needs to be estimated. The observation process is modeled as

$$dy = M[m(t)]\, h(t)dt + d\omega_y, \tag{3}$$

where $y(t) \in \Re^{n_s}$ corresponds to the integral of the (noisy) cursor position. dy is the n_s-dimensional observation/measurement vector, and $d\omega_y$ is the corresponding measurement error. In this observation model $m(t)$ is the unknown state and $h(t)$ plays the role of an observation "matrix". In usual estimation problems the latter would be fixed, but here the subject can control it directly by changing hand position. This makes apparent the exploratory nature of hand movements in our setting. To put this observation model in a more familiar form, define the n_s-by-$n_s n_h$ matrix $H[h(t)]$ such that $H[h(t)]\, m(t) = M[m(t)]\, h(t)$. With this notation, we rewrite (3) as

$$dy = H[h(t)]\, m(t)dt + d\omega_y. \tag{4}$$

Now suppose the prior over the initial state of the mapping is Gaussian, with mean $\widehat{m}(0)$ and covariance $\Sigma(0)$. Then the posterior over $m(t)$ remains Gaussian for all $t > 0$. Given the additive, Gaussian white noise model, the optimal estimate of the mean and error covariance of the map is propagated by the well-known Kalman-Bucy filter [1], [14]:

$$\begin{aligned} d\widehat{m} &= K\left(dy - H[h(t)]\, \widehat{m}(t)dt\right), \\ K &= \Sigma(t) H[h(t)]^{\mathsf{T}}\, \Omega_y^{-1}, \\ d\Sigma &= \Omega_m dt - K(t) H[h(t)]\, \Sigma(t)dt, \end{aligned} \tag{5}$$

with measurement equation (4), and $H[h(t)]$ appropriately sized. The properties of the noise and disturbances do not change over time. The mean and covariance of the state estimate is $\widehat{m}(t)$ and $\Sigma(t)$, respectively, and $d\omega_y$ is a white, zero-mean Gaussian random process as well, with covariance Ω_y. $d\omega_m$ and $d\omega_y$ are assumed to be uncorrelated[2]. The Kalman filter can be written in innovations form by expressing $\widehat{m}(t)$ as another stochastic process:

$$d\widehat{m} = K d\omega_{\widehat{m}}. \tag{6}$$

Here $\omega_{\widehat{m}}(t)$ is a standard Brownian motion process with unit covariance. The advantage of the innovations form is that we are now dealing with a fully observable system where $\widehat{m}(t)$ and $\Sigma(t)$ act as state variables. The latter is a symmetric matrix, therefore it is uniquely defined by its upper-triangular part. Let $\sigma(t) \in \Re^{n_s n_h (n_s n_h + 1)/2}$ be the vector of upper-triangular elements of $\Sigma(t)$. Similarly to M and H above, we will define the linear operators f and F which transform between the vector and matrix representations of the covariance, namely $\sigma(t) = f[\Sigma(t)]$ and $\Sigma(t) = F[\sigma(t)]$. We can now define the composite state vector of our system which includes the mean and covariance (a measure of uncertainty) of the estimates:

$$x(t) = \left[h(t); s(t); \widehat{m}(t); \sigma(t)\right], \tag{7}$$

and write its stochastic dynamics in the control-affine form [15]

$$dx = \left(a(x) + Bu\right) dt + C(x)\, d\omega. \tag{8}$$

In the next three equations explicit time-dependence is temporarily omitted for clarity (e.g. $h(t) \rightarrow h$). The uncontrolled dynamics $a(x)$ are needed to represent the evolution of the covariance matrix:

$$a(x) = \begin{bmatrix} 0 \\ 0 \\ 0 \\ f\left[\Omega_m - F[\sigma] H[h]^{\mathsf{T}} \Omega_y^{-1} H[h] F[\sigma]\right] \end{bmatrix}. \tag{9}$$

The controlled dynamics Bu capture the evolution of the hand state:

$$B = \begin{bmatrix} I & 0 & 0 & 0 \end{bmatrix}^T. \tag{10}$$

The noise-scaling matrix $C(x)$ captures the dependence of the innovation process on the filter gain matrix, as well as the covariance of the target drift:

$$C(x) = \begin{bmatrix} 0 & & & \\ & \sqrt{\Omega_s} & & \\ & & F[\sigma] H[h]^{\mathsf{T}} \Omega_y^{-1} & \\ & & & 0 \end{bmatrix}. \tag{11}$$

Here $\sqrt{\Omega_s}$ denotes the symmetric matrix square root, and $\omega(t)$ is a vector of standard Brownian motion processes with unit covariance.

The main idea behind our work is to define a sensible cost function for the control task, and then induce an indirect cost over exploratory actions by considering how they affect uncertainty. Since we have a tracking task, the obvious state-dependent cost rate is

$$\left\| M[m(t)]\, h(t) - s(t) \right\|^2.$$

[1] In this study we apply deterministic hand dynamics and assume a first-order model where the control signal corresponds directly to hand velocity. A more sophisticated biomechanical arm model (in the case of modeling biomechanical systems) can be used but is not necessary in order to capture the trade-off of interest.

[2] Note that $K(t)$ is the filter gain matrix. It is clear that K is a deterministic function of the other quantities and does not have to be propagated through time explicitly. Unlike usual estimation problems where $K(t)$ can be precomputed, here it needs to be computed online because we do not know in advance how the state $h(t)$ and thereby the observation matrix $H[h(t)]$ will evolve over time.

This cost rate depends on the true state of the mapping $m(t)$ which is not part of our composite state vector x. However we know that $m(t)$ has a Gaussian distribution with mean $\widehat{m}(t)$ and covariance $F[\sigma(t)]$, and both $\widehat{m}(t)$ and $\sigma(t)$ are part of $x(t)$. Therefore we can compute the cost rate by taking an expectation over $m(t)$. Using the identity $H[h(t)]m(t) = M[m(t)]h(t)$, we have (again omitting time-dependence):

$$q(x) = E\left(\|M[m]h - s\|^2\right), \tag{12}$$
$$= \|M[\widehat{m}]h - s\|^2 + \mathrm{tr}\left(H[h]^{\mathsf{T}} F[\sigma] H[h]\right),$$

where $tr(\cdot)$ denotes the trace. We also incorporate a quadratic control cost, to obtain the following cost rate:

$$\ell(x,u) = q(x) + \tfrac{1}{2}\|u\|^2. \tag{13}$$

Thus we have transformed our partially-observable tracking problem to the fully-observable non-linear stochastic optimal control problem defined by equations (8) and (13).

The cost terms defined in $q(x)$ each have a unique significance. The first term in (12) represents a simple tracking cost, quantifying the control objective to minimize the distance between the cursor and target. The second term in (12) represents an uncertainty cost. This comes into play only when the system moves into a region where there is larger uncertainty. The advantage of this term is the triggering of random 'exploratory' actions which by nature reduce the uncertainty and allow the first cost term to once again dominate. The final term in (13) is a quadratic control cost to penalize overly large control.

The cost function for the set of problems discussed in this paper is taken to be an infinite horizon discounted cost, since there is no expected final time for the behavior.

IV. Solution methods

One approach to approximating the solutions to continuous optimal control problems is to discretize them. However discretization methods such as in [10] and [3] are only feasible in low-dimensional spaces, while the problems we are dealing with tend to be rather high-dimensional. In particular, the dimensionality of the augmented state x is

$$n_x = n_h + n_s + n_s n_h + n_s n_h (n_s n_h + 1)/2. \tag{14}$$

For $n_h = 2$ and $n_s = 1$, which is the simplest redundant problem and corresponds to the experiments described below, we have $n_x = 2+1+2+3 = 8$. For $n_h = 3$ and $n_s = 2$, corresponding to a mapping from 3D hand space to a 2D screen, we have $n_x = 3+2+6+21 = 32$. Thus we have to focus on continuous function approximation methods - which may lack theoretical guarantees in terms of convergence and error bounds, but in practice turn out to have very appealing properties.

The method begins with the continuous stochastic dynamical equation defined as in (8)-(13).

Consider an infinite horizon discounted cost formulation, with discount factor $\alpha > 0$. The optimal value function for our problem satisfies the Hamilton-Jacobi-Bellman (HJB) equation for stochastic systems:

$$\alpha v(x) = \min_u \Big\{ q(x) + \frac{1}{2}\|u\|^2 + (a(x)+Bu)^T v_x \tag{15}$$
$$+ \frac{1}{2} tr(C(x)C(x)^T v_{xx}) \Big\},$$

where the subscripts denote partial derivatives. The minimization in (15) can be performed in closed form to yield the optimal feedback control law

$$\pi(x) = -B(x) v_x(x). \tag{16}$$

Substituting (16) into (15) and dropping the min operator we arrive at the minimized HJB equation

$$\alpha v(x) = q(x) + a(x)^T v_x(x) \tag{17}$$
$$+ \frac{1}{2} tr(C(x)C(x)^T v_{xx}(x)) - \frac{1}{2}\|\pi(x)\|^2.$$

Using (16) and (17) we now construct a function approximation scheme based on the collocation method [4] to approximate a continuous time optimal control law. We begin with a general linear (in the parameters, nonlinear in the state) function approximator

$$v(x,w) = \sum_i w_i \phi^i(x) = \phi^T(x) w, \tag{18}$$

where $\{\phi^i\}$ is a set of predefined features, and w_i are corresponding to-be-determined weights. Function approximation is a broad topic, and many choices are available for the set $\{\phi^i(x)\}$. The reader is referred to [13] for a survey of techniques. It is possible to approximate any given function to any desired accuracy given a sufficient number of terms [11]. However, with sensible choices of terms an equivalent quality of fit is obtained using many less terms. Since we have a tracking task, we first choose to include the two cost terms. We then choose the set of all quadratic terms of the form x_r, and $x_r x_s$ to generally fit the function, with a set of Gaussians to introduce corrections about the quadratic.

Before we can substitute (18) into (17), the first and second derivatives of $v(x)$ must also be computed:

$$v_x(x,w) = \sum_i w_i \phi^i_x(x) = \phi^T_x(x) w, \tag{19}$$

$$v_{xx}(x,w) = \sum_i w_i \phi^i_{xx}(x) = \phi^T_{xx}(x) w. \tag{20}$$

The idea is to reduce this nonlinear partial differential equation to an equation of the form

$$\mathbf{M} w = \mathbf{d}, \tag{21}$$

which we can solve as a linear least squares problem, then recompute $\mathbf{d}$ and iterate until appropriate convergence criteria are met.

We get to (21) by rewriting (17) as

$$\alpha v(x) - a(x)^T v_x(x) - \frac{1}{2} tr(C(x)C(x)^T v_{xx}(x)) \tag{22}$$
$$= q(x) - \frac{1}{2}\|\pi(x)\|^2,$$

Then substituting (18), (19), and (20) into (23) and simplifying for w, defining $\mathbf{M}$ and $\mathbf{d}$

$$\mathbf{M} = \Big\{ M_{j,i} = \Big(\alpha\phi^i(x_j)^T - a(x_j)^T \phi^i_x(x_j)^T \tag{23}$$
$$- \frac{1}{2} tr\{C(x_j)C(x_j)^T \phi^i_{xx}(x_j)^T\} \Big), \forall i,j \Big\},$$

$$\mathbf{d} = \Big\{ d_j = q(x_j) - \frac{1}{2}\|\pi(x_j)\|^2, \forall j \Big\}. \tag{24}$$

Function Approximation Scheme (FAS) initialization: The function approximation scheme is initialized in the following way:

1) Generate a set of vectors $\{x_j\}$ and Gaussian centers $\{c_{i_g}\}$ which span the space of interest (in this case the boundaries of the space are chosen to be equivalent to the experimental boundaries) and $n_j > n_i$. Make sure that at least one x_j equals each c_i. If $S \in \Re^r$ is the state space,

$$x_j = rand_j\{S\}, \qquad c_{i_g} = rand_{i_g}\{S\}. \tag{25}$$

2) Compute $\phi^i(x_j)$, $\phi^i_x(x_j)$, $\phi^i_{xx}(x_j)$ and store the results for all i, j.
3) Initialize w^0 in a sensible way :

$$w^0 = \begin{cases} 0 & i \neq i_{cost} \\ 1 & i = i_{cost} \end{cases}, \tag{26}$$

where i_{cost} are the two locations of the cost terms included in the function approximator. Then we initialize our control policy as

$$\pi^0(x_j) = -B\phi_x(x_j)^T w^0. \tag{27}$$

FAS Iteration: Given a set of control actions $\pi(x_j)$ for every j, the update from iteration k to k+1 is as follows:

1) Substitute the constraints $\phi^i(x_j)$, $\phi^i_x(x_j)$, and $\phi^i_{xx}(x_j)$ into (34), and $\pi^k(x_j)$ into (24) to obtain one constraint on **w** for every j.
2) Find the least-squares solution to (21) and assign it to w^{k+1}. For the new setting of the function approximator, compute $\pi^{k+1}(x_j)$ for every j using (16)
3) Stop if the stopping criterion is met. Many criteria are possible, and the one used in the present results is

$$e^k = \frac{1}{n_j}\|Mw^k - d\|_2^2, \quad de^k = e^k - e^{k-1}, \tag{28}$$

$$if(\{e^k < \gamma\}|\{de^k < \beta\}) \rightarrow break, \tag{29}$$

where $\gamma = 10^{-3}$ and $\beta = 10^{-5}$ are tolerances . We also test for divergence:

$$if(\{(de^k) > \lambda\}|\{isnan(de^k) == true\}) \rightarrow break, \tag{30}$$

where $isnan$ is a test for invalid numbers, and $\lambda = 10^{-3}$ is a positive constant which is arbitrary, but can be on the order of one.

V. Results

A. Optimality and convergence of the function approximation scheme

The FAS converged after only a few iterations. It is not guaranteed to converge for very poor initializations of w_j, however the algorithm is somewhat forgiving. $|e_{nom} - e_i|$ for initializations spanning a full order of magnitude in each parameter (w_{track}, and $w_{explore}$) was consistently $< 7.846e-4$ in each case, and $\|w^{nom} - w^i\| < 1e-15$ where w^{nom} represents the set of weights in the first solution of the group.

Thus the FAS converges to a similar solution in each case given the same states and Gaussian centers. With a smaller convergence criterion, less variability in the computed weights is observed. In all cases, 1-5 iterations were required for convergence, and typical Bellman error was $< 10^{-3}$, the tolerance selected.

The number of Gaussians to include could be determined by posing the problem as an optimization to minimize Bellman error. However, it is known that minimizing the Bellman error ($\mathbf{M}w - \mathbf{d}$) does not necessarily correspond to the best control performance possible for stochastic systems and thus the resulting 'optimal' control policy can be deceiving. A more appropriate criterion here is to use the performance criterion - the cost function history during simulation. The optimal number of Gaussians over a range of 400-1000 was 712, with a variance (elongated appropriately in each axis) of 0.0161.

B. Simulations

The FASC was compared to several ad-hoc controls, and to human subject data for the same task. The other controllers were 1)a CE-based controller with a Kalman filter estimating the map parameters, 2) a non-adaptive controller (NAC) which did not estimate uncertain parameters and used only the error between cursor and target to drive actions, and 3)a controller driven by white noise (RANDC). The RANDC was used primarily to compare estimation performance.

To perform the simulations, the continuous time nonlinear ordinary differential equation (8) was integrated with explicit Euler integration and for comparison, predictor corrector methods [7] both with a stepsize of 10^{-3}sec. When necessary, experimental data was interpolated linearly to match the step size.

Map estimation and uncertainty reduction: The most information can be gained by moving in such a way as to cover as much of the unknown space as possible in a random way. Indeed, it seems clear that replacing the controller with a white noise signal generator of arbitrary gain could yield the closest estimate of the map.

However, exploring the entire subspace is neither necessary nor desired in our case. The measure of uncertainty can be broken down into relevant parts to the task and irrelevant parts in the following way. The task-relevant uncertainty is measured by the trace term in (12), whereas the irrelevant uncertainty can be quantified by finding the orthogonal complement of $h(t)$, defined by $h(t)^\perp$. Then $h(t)^\perp$ is substituted in the trace term in (12), and the two uncertainty histories are computed over several trials (Fig. 2(e), 2(f), and Table I). These measures were compared for the FASC, CE and RANDC (the latter using a gain of 50, which approximates human bandwidth).

What is clear is that the FASC reduces the relevant uncertainty the most of all methods compared. The FASC emphasizes reducing the h-space uncertainty (which is relevant in this task, effectively ignoring irrelevant uncertainty). The CE controller does not achieve as much overall reduction in uncertainty as the FASC since its design does not include an exploratory component[3]. Only the white noise control reduces both types of uncertainty; it does not differentiate between the two types of uncertainty. However, the RANDC does not achieve as large of a reduction of the task-relevant uncertainty as the other two methods. By distributing control action over the whole space, without unlimited gain,

[3]Additionally, the CE gain affects this measure heavily - low gain reduces uncertainty less, but high gains lead to instability which yield poor numerical results.

TABLE I

$\sqrt{Norm}$ OF UNCERTAINTY QUANTITIES PLOTTED IN FIG. 2(E) AND 2(F), WHICH IS A STANDARD DEVIATION QUANTITY.

	h-space	$h^{\perp}$-space
FASC	8.8	36.7
CE	10.2	81.2
RAND	13.7	14.7

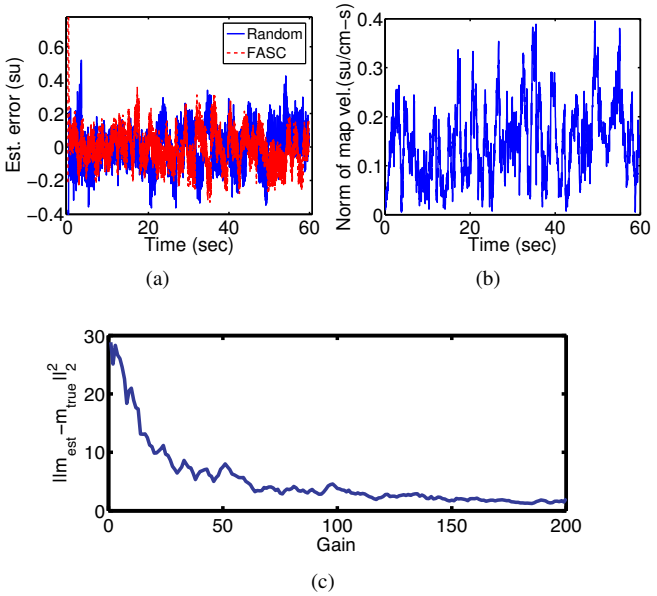

Fig. 1. (a) Estimation error for FAS scheme, with a noise level of 0.1su, and estimation error for random signal generator, with a gain of 50. (b) $||\dot{m}(t)||_2^2$, giving a measure of where the map is changing more rapidly. (c) Map estimation error (over 1 trial), $\|m_{est} - m_{true}\|_2^2$ vs. maximum amplitude of white noise 'controller.' The estimation improves with increasing white noise amplitude.

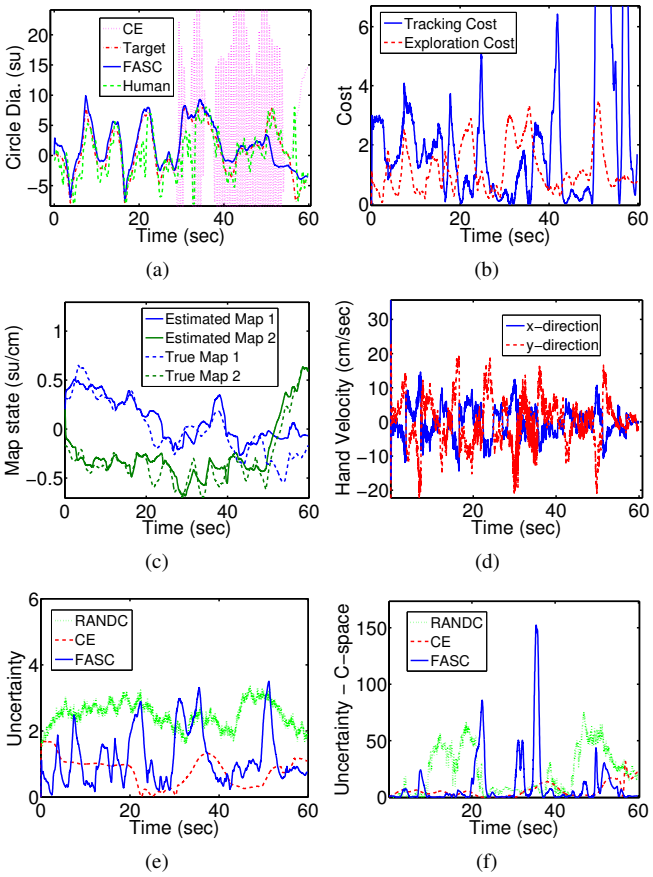

Fig. 2. (a) A section of a 60 second trial displaying human subject data, FAS, and proportional feedback controller tracking. (b) represents the two portions of the cost - the tracking and exploration costs. Plot (c) shows the true and estimated map. (d) shows the FAS control actions. (e) and (f) show two measures of uncertainty - (e) is in h-space, and is given by the trace term in 12, while (f) is the same, but in $h^{\perp}$-space.

the RANDC wastes much of its resources reducing task-irrelevant uncertainty. When considering the mean of the estimation error, Fig. 1(c) shows that as the amplitude of the white noise is increased, the averaged 2-norm of the mean of the estimation error decreases. Mathematically this makes sense. If the map were stationary, then the amplitude would control the speed of convergence of the estimator to the true map. Since the map is not stationary, the amplitude of the random control inputs must be large enough that the current map state is well estimated before it changes significantly. Fig. 2(a) and 1(a) show that the FAS controller leads to a comparable cursor estimation error (Sec. III) as a white noise signal, while still tracking the target in a stable manner, and as we have seen, reducing relevant uncertainty.

Controller performance: The controller's performance can be characterized by the behavior of the two components of the cost function (Fig. 2(b)), and by summing the total average cost per trial. The the NAC performed worst ($cost_{total} = 6.3e5$), followed by the RANDC (this does not diverge as far as the NAC, so the overall tracking cost is lower) with a gain of 50 ($cost_{total} = 1.6e5$), the CE-based control ($cost_{total} = 4.7e4$ or $4.2e7$, depending on stability), and then the best performance was achieved by the FASC ($cost_{total} = 1.5e4$). This is attributed to the cost term that triggers random movements when entering state space with large uncertainty, which happens periodically, as can be seen in Fig. 2(b). In Fig. 2(b) it is clear that exploration cost increases when tracking cost is low, and then tracking cost increases during the resulting exploratory actions, leading to a decrease in uncertainty and thus exploration cost. This whole behavior is periodic since the map and target velocity continuously change, requiring new information input at varying speeds (i.e. when $m(t)$ has a lower velocity - Fig. 1(b) - tracking can be achieved with less exploration, and as uncertainty increases due to rapid changes in $m(t)$, tracking suffers during exploratory movements).

The estimator parameters converge to a close enough approximation of the true mapping to achieve good tracking. A poor estimate would result in large control and possible instability. The FAS improves estimates with larger control inputs, so this acts to counter uncertainties, instabilities, and errors. Fig. 2(d) shows the FAS control actions, which appear to be near mirror images at certain time periods (e.g. - 0-15sec). Comparing that figure with Fig. 2(c) it is clear that the control in one axis becomes negative when the map is negative, and a resulting desired cursor output is positive.

C. Comparisons to human subjects data

See Fig. 3 and caption for more specifics of the experimental task described and modeled in section II. The human

Fig. 3. Experimental setup. The subject sits in front of a monitor, holding a 3D absolute position sensor (Polhemus Liberty). Here $n_h = 2, n_s = 1$. The center-stationary green circle is the target, whose diameter ($s(t)$) fluctuates according to damped Brownian motion, and there is a concentric 'cursor' (red circle) whose diameter is $z(t) = m(t)^T h(t)$, with $m(t)$ fluctuating according to an independent damped Brownian motion. The subject is then told to move however possible to try to keep $z(t) = s(t)$. Trials are 1 min., started by a key-press and sampling rate is 50Hz.

subjects, as shown in Fig. 2(a), were able to track the target. When the map's velocity changed or for some other reason they lost the map, they exhibited sudden large exploratory movements and then tracked for another period of time. Often both the model and human exhibited exploratory movements at similar times.

It is notable that the FASC outperforms the human subject in terms of tracking the target. This is not surprising since a human subject has delays between perception, processing and action, as well as biomechanical limitations. The model is not approximating any of those processes, and so should outperform humans. Though we cannot measure (given the current experimental setup) the human exploratory cost function, we can compute the associated tracking cost. The average human tracking cost for 4 subjects over 24 trials was $2.6e4$. This is higher than the FAS control ($1.1e4$), but within the same order, whereas the control which does not adapt or explore had an average cost of $6.2e5$ which was the highest of control systems compared. The basic feedback controller, when given a carefully tuned gain, had an average tracking cost of $3.3e4$, which is also comparable with humans, but when poorly initialized, or presented with highly rapid map changes (which were part of the experiments), the CE control diverged frequently (Fig. 2(a), around 30sec where a rapid map change takes place, more complete loss of stability at 50sec). The CE control did not exhibit exploratory movements. Instead if the gain was too low it failed to track sufficiently or to excite the system enough to estimate $m(t)$, and with a sufficient gain, deviations which occurred due to map estimation errors led to explosive instability.

VI. Conclusion

We introduced a method of approximating optimal control laws in problems which involve a trade-off between exploration and exploitation. This was done by treating the Kalman-Bucy filter dynamics as part of the plant dynamics, incorporating the estimation uncertainty in the state vector, and applying a basis function approximation scheme to the Hamilton-Jacobi-Bellman equation. Our method produced control laws which outperformed other controllers. It not only succeeded in achieving low tracking error, but also did so with nonlinear time-varying system parameters and large uncertainties. This approach shows promise in situations where active exploration can reduce uncertainty. Additionally, this method distinguishes between and reduces task-relevant versus task-irrelevant uncertainty. That is advantageous in real-world situations where there are limited control resources.

The method was also compared to human data from an uncertain tracking task similar to the one being modeled. Although these comparisons are very preliminary, we already see interesting similarities which will be pursued in future work. In fact our motivation for developing the method is to study human behavior. Stochastic optimal control combined with Bayesian estimation is emerging as the leading theoretical framework for understanding sensorimotor function in the brain [16]. However model-data comparisons are presently limited to the few simple tasks where we can compute what behavior is optimal. New methods such as the one developed here, which allow us to extend optimality principles to more interesting tasks, can accelerate progress in the field of sensorimotor control. Of course such methods also are likely to find engineering applications.

References

[1] B. Anderson and J. Moore. *Optimal Filtering*. Prentice Hall, 1979.

[2] Y. Bar-Shalom. Stochastic dymamic programming: Caution and probing. *IEEE Transactions on Automatic Control*, 26(5):1184–1195, 1981.

[3] D. Bertsekas and S. Schreve. *Stochastic Optimal Control: The Discrete Time Case*. Athena Scientific, 1995.

[4] L. Collatz. *The Numerical Treatment of Differential Equations*. Springer-Verlag, 1966.

[5] R. de Callafon and P. Van den Hof. Multivariable feedback relevant system identification of a wafer stepper system. *IEEE Transactions on Control Systems Technology*, 9(2):281–390, March 2001.

[6] A. Feldbaum. Dual control theory i-iv. *Automation and Remote Control*, 21:874-880, 21:1033-1039, 22:1-12, 22:109-121, 1960-61.

[7] J. Ferziger. *Numerical Methods for Engineering Application*. John Wiley and Sons, New York, NY, 2nd edition edition, 1998.

[8] P. Ioanno and J. Sun. *Robust Adaptive Control*. Prentice Hall, 1996.

[9] M. Krstic, I. Kanellakopoulos, and P. Kokotovic. *Nonlinear and Adaptive Control Design*. John Wiley and Sons, 1995.

[10] H. Kushner and P. Dupuis. *Numerical Methods for Stochastic Control Problems in Continuous Time*. Springer-Verlag, New York, New York, 1992.

[11] L. Ljung. *System Identification: Theory for the User*. Prentice Hall, PTR, Upper Saddle River, NJ, 2nd edition edition, 1999.

[12] L. Scardoviet, M. Baglietto, and T. Parisini. Active state estimation for nonlinear systems: A neural approximation approach. *IEEE Transactions on Neural Networks*, 18(4):1172–1184, July 2007.

[13] J. Si, A. Barto, W. Powell, and D. Wunsch, editors. *Handbook of Learning and Approximate Dynamic Programming*. IEEE Press on Computational Intelligence, 2004.

[14] H. W. Sorenson, editor. *Kalman Filtering: Theory and Application*. IEEE Press, 1985.

[15] R. Stengel. *Stochastic Optimal Control: Theory and Application*. John Wiley and Sons, 1986.

[16] E. Todorov. Optimality principles in sensorimotor control. *Nature Neuroscience*, 7:907–915, 2004.

[17] P. Van den Hof and R. Schrama. Idenfication and control - closed loop issues. *Automatica*, 31(12):1751–1770, 1995.

[18] B. Wittenmark. Adaptive dual control methods: An overview. In *5th IFAC symposium on Adaptive Systems in Control and Signal Processing*, pages 67–73, 1995.