Published in final edited form as:
JMLR Workshop Conf Proc. 2013 ; : 606–614.

Learning an Internal Dynamics Model from Control Demonstration

Matthew D. Golub, **Steven M. Chase***, and **Byron M. Yu***
Carnegie Mellon University, 5000 Forbes Ave., Pittsburgh, PA 15213 USA

Matthew D. Golub: MGOLUB@CMU.EDU; Steven M. Chase: SCHASE@CMU.EDU; Byron M. Yu: BYRONYU@CMU.EDU

Abstract

Much work in optimal control and inverse control has assumed that the controller has perfect knowledge of plant dynamics. However, if the controller is a human or animal subject, the subject's internal dynamics model may differ from the true plant dynamics. Here, we consider the problem of learning the subject's internal model from demonstrations of control and knowledge of task goals. Due to sensory feedback delay, the subject uses an internal model to generate an internal prediction of the current plant state, which may differ from the actual plant state. We develop a probabilistic framework and exact EM algorithm to jointly estimate the internal model, internal state trajectories, and feedback delay. We applied this framework to demonstrations by a nonhuman primate of brain-machine interface (BMI) control. We discovered that the subject's internal model deviated from the true BMI plant dynamics and provided significantly better explanation of the recorded neural control signals than did the true plant dynamics.

1. Introduction

Inverse optimal control (IOC) and inverse reinforcement learning (IRL) aim to identify a cost function from demonstrations of successful control (Boyd et al., 1994; Schaal, 1999; Ng & Russell, 2000; Abbeel & Ng, 2004; Ratliff et al., 2006; Coates et al., 2008; Ziebart et al., 2008). These approaches typically require a model of the plant dynamics, which enables prediction of future states given the current state and control input. In previous work, it has been assumed that the controller's internal belief about the plant dynamics matches the actual plant dynamics. However, when the controller is a human or animal subject, this internal belief may differ from the actual plant dynamics (Crapse & Sommer, 2008), especially if the subject has limited experience driving the plant. This mismatch can exist even when demonstrated control is proficient (e.g., the implemented control strategy may be only locally optimal). Because the plant dynamics and cost function jointly determine the optimal control policy, an incorrect assumption about the dynamics model can lead to misestimation of the cost function via IOC or IRL.

Ideally, we would like to use demonstrated control to learn both the subject's internal model of the plant dynamics and the cost function together. This joint estimation is difficult, so previous work has focused on learning the cost function while assuming known plant dynamics. Here, we present and solve the complementary problem of learning the subject's internal model of the plant dynamics while assuming knowledge of the task goals. This

*denotes equal contribution

Proceedings of the 30th International Conference on Machine Learning, Atlanta, Georgia, USA, 2013.

problem is challenging because at each control decision, the subject must generate an internal estimate of the current plant state based on delayed sensory feedback (Miall & Wolpert, 1996), and we cannot directly observe these internal state estimates.

We introduce a probabilistic *internal model estimation* (IME) framework through which inference and learning provide a solution to the current problem in the setting of linear-Gaussian internal model dynamics and quadratic cost functions. In IME, the subject's internal model of the plant dynamics defines trajectories of latent variables representing the subject's moment-by-moment internal estimates of the plant state. We assume knowledge of the control signals sent by the subject, the plant state feedback available to the subject, and target states to which the subject intends to drive the plant during control. Importantly, we make no assumption that the subject's internal dynamics model should match the true plant dynamics.

Beyond the algorithmic advance, the ability to extract a subject's internal model has many potential applications in neuroscience and human-in-the-loop control. Access to a subject's internal model could provide a means for tracking and encouraging skill acquisition in complex tasks, including brain-machine interface (BMI) control, telerobotic surgery or remote control of unmanned-aerial vehicles. In this work we apply the developed methods toward demonstrations of BMI cursor control. BMIs have been developed to assist disabled patients by translating neural activity into control signals for a prosthetic limb or computer cursor (Green & Kalaska, 2011). BMI control is an acquired skill, akin to driving a car or helicopter. Previous studies have shown that subjects improve BMI control performance over time (Taylor et al., 2002; Ganguly et al., 2011). This improvement is likely a result of the subject refining an internal model of the BMI plant dynamics through experience. Access to the subject's internal model of the BMI, through the methods we develop here, may inform the design of future BMI systems and may provide neuroscientists with novel tools for investigating the neural basis of feedback motor control and motor learning.

We begin Section 2 by formalizing the internal model estimation problem. In Section 3 we propose a probabilistic framework for solving the internal model estimation problem. Section 4 details the validation of the framework through application to real neural data underlying BMI control of a computer cursor.

2. Problem Formulation

A standard control model takes the following form:

$$\textbf{Dynamics}: \mathbf{x}_{t+1}=f_1(\mathbf{x}_t,\mathbf{u}_t) \quad (1)$$

$$\textbf{Cost}: J(\{\mathbf{x}_t\},\{\mathbf{u}_t\}) \quad (2)$$

where f_1 represents the subject's belief about the plant dynamics, $\mathbf{x}_t \in \mathbb{R}^n$ is the subject's belief of the plant state at timestep t, $\mathbf{u}_t \in \mathbb{R}^m$ is the control input issued at timestep t, and J is the cost function that encodes task goals and control effort. We distinguish the subject's internal model of the plant dynamics, f_1, from the actual plant dynamics, f_2:

$$\mathbf{y}_{t+1}=f_2(\mathbf{y}_t,\mathbf{u}_t) \quad (3)$$

where $\mathbf{y}_t \in \mathbb{R}^p$ is the actual plant state at timestep t.

Due to sensory feedback delay, the feedback available at timestep t represents the plant state at timestep $t - \tau$, where τ is the feedback delay. To predict the current plant state, the subject

can use f_1 as a forward model, propagating $\mathbf{y}_{t-\tau}$ (or a noise-corrupted function of it) forward in time using knowledge of the plant dynamics and previously issued controls $\mathbf{u}_{t-\tau}, \ldots, \mathbf{u}_{t-1}$. In general, the subject's internal beliefs $\{\mathbf{x}_t\}$ may be inconsistent with the actual plant states $\{\mathbf{y}_t\}$ due to differences between f_1 and f_2 and due to sensory noise.

The problem we seek to solve is:

$$\begin{aligned} \textbf{Given}: &\quad \{\mathbf{y}_t\}, \{\mathbf{u}_t\}, J \\ \textbf{Estimate}: &\quad f_1, \{\mathbf{x}_t\}, \tau \end{aligned}$$

That is, given trajectories of actual plant state and control input, and assuming a cost function, we seek to estimate the subject's internal model of the plant dynamics, the subject's internal estimates of plant state, and the sensory feedback delay.

In the remainder of this paper, we focus on the case where f_1 is linear-Gaussian and J is quadratic over the internal states, in analogy to the well-studied linear-quadratic regulator (Anderson & Moore, 1990). These choices allow us to derive an approximation-free algorithm to solve the internal model estimation problem.

3. Probabilistic framework for internal model estimation

The IME probabilistic model is as follows:

$$\mathbf{x}_{t-\tau}^t | \mathbf{y}_{t-\tau} \sim \mathcal{N}(\mathbf{H}\mathbf{y}_{t-\tau,} \mathbf{W}_0) \quad (4)$$

$$\mathbf{x}_{k+1}^t | \mathbf{x}_k^t, \mathbf{u}_k \sim \mathcal{N}(\mathbf{A}\mathbf{x}_k^t + \mathbf{B}\mathbf{u}_k + \mathbf{b}_{0,} \mathbf{W}) \quad (5)$$

$$\mathbf{G}_t | \mathbf{x}_{t+1}^t \sim \mathcal{N}(\mathbf{C}_t \mathbf{x}_{t+1}^t, \mathbf{V}) \quad (6)$$

At timestep $t \in \{1,\ldots, T\}$, $\mathbf{y}_t \in \mathbb{R}^p$ is the actual plant state, $\mathbf{x}_k^t \in \mathbb{R}^n$ is the subject's internal estimate of the timestep $k \in \{t-\tau, \ldots, t+1\}$ plant state (see below for detailed explanation), $\mathbf{u}_t \in \mathbb{R}^m$ is the subject's control input, and $\mathbf{G}_t \in \mathbb{R}^q$ represents control goals. The parameters are the feedback matrix $\mathbf{H} \in \mathbb{R}^{n \times p}$, the subject's internal model parameters $\{\mathbf{A} \in \mathbb{R}^{n \times n}$, $\mathbf{B} \in \mathbb{R}^{n \times m}$, $\mathbf{b}_0 \in \mathbb{R}^n\}$, the cost matrices $\mathbf{C}_t \in \mathbb{R}^{q \times n}$, noise covariance matrices $\{\mathbf{W}_0 \in \mathbb{R}^{n \times n}$, $\mathbf{W} \in \mathbb{R}^{n \times n}$, $\mathbf{V} \in \mathbb{R}^{q \times q}\}$, and the sensory feedback delay $\tau \in \mathbb{Z}^+$. The IME graphical model for a single timestep feedback delay ($\tau = 1$) is shown in Fig. 1.

Due to sensory delays, the plant state feedback available at timestep t is outdated by τ timesteps. Accordingly, (4) defines the subject's noisy, partial observation of delayed plant state feedback. Sitting at timestep t, the subject uses this feedback, $\mathbf{y}_{t-\tau}$, to form an internal estimate, $\mathbf{x}_{t-\tau}^t$, of the timestep $t-\tau$ plant state. The noise covariance $\mathbf{W}_0$ accounts for sensory noise.

We define the subject's internal dynamics model in (5) to be a Gaussian linear-dynamical system that propagates the subject's internal estimates of plant state given control input. At timestep t, the subject makes internal estimates, $\mathbf{x}_k^t$, of the past ($k = t-\tau, \ldots, t-1$), current ($k = t$), and future ($k = t + 1$) plant states. This timestep t internal state chain (Fig. 1A) corresponds to a row of latent states in the IME graphical model (Fig. 1B). The state chain begins with $\mathbf{x}_{t-\tau}^t$, the subject's internal belief about the plant state feedback, $\mathbf{y}_{t-\tau}$.

Subsequent internal state estimates, $\{\mathbf{x}^t_{t-\tau+1}, \ldots, \mathbf{x}^t_{t+1}\}$, may be inconsistent with the true plant states, $\{\mathbf{y}_{t-\tau+1}, \ldots, \mathbf{y}_{t+1}\}$, because i) sensory feedback is not yet available for these timesteps, and ii) there may be mismatch between the subject's internal model (5) and the true plant dynamics (3). Internal state transitions not explained by the internal model are accounted for by the noise covariance, **W**. At timestep t+1, the subject receives new plant feedback, $\mathbf{y}_{t-\tau+1}$, and generates revised internal estimates, $\{\mathbf{x}^{t+1}_{t-\tau+1}, \ldots, \mathbf{x}^{t+1}_{t+2}\}$, corresponding to a new row of latent states in the graphical model. The variables, $\mathbf{x}^t_k$, where we fix k and vary $t \in \{k-1, \ldots, k+\tau\}$, correspond to a column of latent states in the graphical model and represent successive revisions of the subject's beliefs about the timestep k plant state, given the sensory feedback available at timestep t. Note that (5) is the IME instantiation of (1).

In (6), we encode the subject's cost function. At timestep t, the subject determines the next control signal to send, $\mathbf{u}_t$, which the subject's internal model predicts will drive the plant to state $\mathbf{x}^t_{t+1}$. The cost matrix, $\mathbf{C}_t$, relates this internal state estimate to the given control goals, $\mathbf{G}_t$. Depending on the application, $\mathbf{C}_t$ may be fully specified in advance, or may contain parameters to be learned. For example, in a trajectory tracking task, the $\mathbf{G}_t$ might encode the (known) desired trajectory, and $\mathbf{C}_t$ might simply extract appropriate components of the subject's internal estimate of the upcoming plant state, $\mathbf{x}^t_{t+1}$. Alternatively, $\mathbf{C}_t$ might compute linear functions of feature counts (Ng & Russell, 2000; Abbeel & Ng, 2004) from the subject's internal state estimates. In Section 4 we describe an application in which the $\mathbf{G}_t$ are constant across timesteps and represent a control goal to be attained by some arbitrary time in the future. In this application, we use $\mathbf{C}_t$ to extract the extent to which the subject is on track to achieve the goal state. Note that (6) relates to (2), but does not incorporate control effort. We focus the scope of this work to problems where either i) the cost function is dominated by the state cost, or ii) we can structure the $\mathbf{C}_t$ in (6) to account for an unknown control cost (see Section 4 for an example).

3.1. Model Fitting

In model fitting, we treat actual plant states, $\{\mathbf{y}_t\}$, control inputs, $\{\mathbf{u}_t\}$, and task goals, $\{\mathbf{G}_t\}$, as observed variables. We treat the internal state estimates, $\{\mathbf{x}^t_{t-\tau}, \ldots, \mathbf{x}^t_{t+1}\}$, as unobserved latent variables. We seek the model parameters, **H**, $\mathbf{W}_0$, **A**, **B**, $\mathbf{b}_0$, **W**, $\{\mathbf{C}_t\}$[1], **V**, and τ, that maximize $P(\{\mathbf{G}_t\}|\{\mathbf{y}_t\}, \{\mathbf{u}_t\})$, the likelihood of the control goals under the distribution induced by (4)–(6).

We derived an exact expectation-maximization (EM) algorithm (Dempster et al., 1977) for a specified feed-back delay, τ (see APPENDIX). In the E-step, we infer posterior distributions over the latent variables, $P(\{\mathbf{x}^t_{t-\tau}, \ldots, \mathbf{x}^t_{t+1}\}|\{\mathbf{y}_t\}, \{\mathbf{u}_t\}, \{\mathbf{G}_t\})$, using the current parameter estimates. In the M-step, we update model parameters given the posterior latent variable distributions. Since the relationships in (4)–(6) are linear-Gaussian, all latent and observed variables are jointly Gaussian. Additionally, given all control inputs, an internal state chain for one timestep is conditionally independent of the internal state chains for all other timesteps

$$\mathbf{x}^{t_1}_{t_1-\tau}, \ldots, \mathbf{x}^{t_1}_{t_1+1} \perp \mathbf{x}^{t_2}_{t_2-\tau}, \ldots, \mathbf{x}^{t_2}_{t_2+1} | \{\mathbf{u}_t\} \quad (7)$$

where $t_1 \neq t_2$. These properties of IME enable an exact and efficient E-step update to the posterior latent variable distributions, and closed-form M-step parameter updates.

[1] In some applications $\mathbf{C}_t$ may be known *a priori* rather than learned.

To identify the feedback delay, τ, we fit IME across a sequence of τ values. As τ increases, the number of parameters remains fixed, and thus increasing τ does not lead to overfitting. For this reason we can simply choose τ_{ML} to be the τ whose corresponding IME fit gives the highest training data likelihood, $P(\{\mathbf{G}_t\}|\{\mathbf{y}_t\}, \{\mathbf{u}_t\})$.

4. Internal model estimation from brain-machine interface control

We demonstrate an application of the IME framework to closed-loop BMI cursor control. A BMI system can be viewed as a feedback control system, whereby a human or animal subject (controller) generates neural activity (control signal) to drive a computer cursor (plant). The experimenter defines the mapping from recorded neural activity to cursor movements (i.e., the actual plant dynamics). The experimenter also defines the task goals in each trial by displaying a visual target to which the subject is instructed to drive the BMI cursor. At any moment in time, the subject does not know the current cursor position due to visual feedback delays in the nervous system, and therefore must choose a control signal to issue based on visual feedback of an outdated cursor position. We assume that the subject always intends to drive the cursor straight to the target from an internal estimate of the current cursor position. We seek to use IME to estimate the subject's internal model of the BMI cursor dynamics along with the subject's timestep-by-timestep internal estimates of the current BMI cursor position.

Previous studies have provided behavioral and neurophysiological evidence that subjects use internal models during motor control (Crapse & Sommer, 2008). In the context of BMI, subjects need to learn to control the cursor, which likely involves refining the internal model with practice controlling the BMI.

4.1. BMI experiments

A 96-channel Utah electrode array was implanted in motor cortex of a Rhesus monkey. In each of 36 experimental sessions, we simultaneously recorded from tens (26 ± 3.44) of neurons, and spike counts were taken in Δt = 33 ms non-overlapping bins. Two-dimensional cursor velocity was linearly decoded from recorded spike counts, and cursor positions were updated according to

$$\mathbf{y}_{t+1}=\mathbf{y}_t+\boldsymbol{\beta}\tilde{\mathbf{u}}_t+\beta_0 \quad (8)$$

$$\tilde{\mathbf{u}}_t=\frac{1}{5}\sum_{k=0}^{4}\mathbf{u}_{t-k} \quad (9)$$

where $\mathbf{y}_t \in \mathbb{R}^2$ is the cursor position displayed at timestep t, $\mathbf{u}_t \in \mathbb{R}^m$ is the raw spike count vector across m simultaneously recorded neuronal units at timestep t, and $\tilde{\mathbf{u}}_t \in \mathbb{R}^m$ is the vector mean spike count over the past 5 timesteps. The decoding parameters $\beta \in \mathbb{R}^{2\times m}$ and $\beta_0 \in \mathbb{R}^2$ were determined via the population vector algorithm (Georgopoulos et al., 1983). Note that (8) describes the actual dynamics of the BMI cursor, corresponding to (3).

In each experimental trial, the subject modulated neural activity to drive the BMI cursor (radius, 8 mm) to a target (radius, 8 mm) that appeared on the perimeter of a circular workspace (radius, 85 mm). A trial was deemed successful and terminated as soon as the cursor visibly overlapped with the target for 50 ms. A trial was deemed a failure if the subject did not acquire the target within 2 s. The subject typically failed less than 5% of trials. Experimental details were previously described in Chase et al. (2012).

4.2. IME formulation for BMI control

The specific IME formulation we applied to the BMI control data is as follows:

$$\mathbf{x}_{t-\tau}^{t} = \mathbf{y}_{t-\tau} \quad (10)$$

$$\mathbf{x}_{k+1}^{t} | \mathbf{x}_{k}^{t}, \mathbf{u}_{k} \sim \mathcal{N}(\mathbf{A}\mathbf{x}_{k}^{t} + \mathbf{B}\mathbf{u}_{k} + \mathbf{b}_{0}, \sigma_{\mathbf{x}}^{2}\mathbf{I}) \quad (11)$$

$$\mathbf{G}_{t} | \mathbf{x}_{t}^{t}, \mathbf{x}_{t+1}^{t} \sim \mathcal{N}(\mathbf{x}_{t+1}^{t} + c_{t}(\mathbf{x}_{t+1}^{t} - \mathbf{x}_{t}^{t}), \sigma_{\mathbf{G}}^{2}\mathbf{I}) \quad (12)$$

where (10)–(12) are specific instances of (4)–(6). A single-timestep slice of this IME graphical model is shown in Fig. 2.

Since the BMI experiments used a position-only state, we define the subject's internal state estimates to reside in the same 2-dimensional position space to allow for interpretable **A**, **B**, $\mathbf{b}_0$, and $\{\mathbf{x}_k^t\}$. For simplicity, we assume the BMI subject has noiseless access to delayed visual feedback of cursor position. These choices are formalized in (10), where $\mathbf{x}_{t-\tau}^{t} \in \mathbb{R}^2$ is the internal estimate made by the subject at timestep t of the BMI cursor position at timestep $t-\tau$, and $\mathbf{y}_{t-\tau} \in \mathbb{R}^2$ is the actual BMI cursor position at timestep $t-\tau$.

In (11), $\mathbf{x}_k^t \in \mathbb{R}^2$ is the subject's internal estimate made at timestep t of the timestep $k \in \{t-\tau, \ldots, t+1\}$ cursor position, and the control input $\mathbf{u}_t \in \mathbb{R}^m$ is neural activity recorded at timestep t, as defined in Section 4.1. Note that the form of the subject's internal dynamics model in (11) matches the form of the actual cursor dynamics from (8), allowing for direct comparison between the two.

The cost function in (12) encodes the subject's intention to drive the cursor straight toward the current visual target position, $\mathbf{G}_t \in \mathbb{R}^2$, from the subject's up-to-date estimate of cursor position, $\mathbf{x}_t^t$, as illustrated in Fig. 3. Within a particular BMI trial, all $\{\mathbf{G}_t\}$ take the same value. The straight-to-target aiming intention is implemented by the neural control input, $\mathbf{u}_t$, which the subject's internal model predicts will bring the cursor to position $\mathbf{x}_{t+1}^t$. The mean of the distribution in (12) lies on the line defined by the two points, $\mathbf{x}_t^t$ and $\mathbf{x}_{t+1}^t$. The length of the line segment between $\mathbf{x}_{t+1}^t$ and $\mathbf{G}_t$ is controlled by c_t, which is determined by the data. The noise variance, $\sigma_{\mathbf{G}}^2$, accounts for motor commands that deviate from straight-to-target aiming.

In model fitting for this IME formulation, we seek the subject's internal model parameters $\mathbf{A} \in \mathbb{R}^{2\times 2}$, $\mathbf{B} \in \mathbb{R}^{2\times m}$, and $\mathbf{b}_0 \in \mathbb{R}^2$, the noise variances $\sigma_{\mathbf{x}}^2 \in \mathbb{R}^+$ and $\sigma_{\mathbf{G}}^2 \in \mathbb{R}^+$, the length scale factors $c_t \in \mathbb{R}^+$, and the visual feedback delay $\tau \in \mathbb{Z}^+$ that maximize $P(\{\mathbf{G}_t\}|\{\mathbf{y}_t\}, \{\mathbf{u}_t\})$, the likelihood of the actual target positions under the distribution induced by (10)–(12). To identify these maximum likelihood parameters, we derived an exact EM algorithm (see APPENDIX). We initialized the EM algorithm with randomly drawn parameters and applied multiple random restarts to avoid local optima.

Note that the number of c_t parameters varies with the amount of data. Although they might be more naturally treated as latent variables, the c_t were treated as parameters to preserve the joint Gaussian relationship between latent and observed variables. As a result, we optimize the c_t rather than integrate over them during model fitting. When cross-validating IME predictions, we will not have access to c_t for the held-out data, but this does not pose a

problem because, as discussed in Section 4.3, we can readily evaluate goodness-of-fit on held out data without requiring these c_t.

4.3. Results

We fit and evaluated IME models using 10-fold cross validation. For each trial that was held out during training, we asked to what extent IME predictions of the subject's internal estimates of cursor position were indicative of straight-to-target aiming. We define aiming error at timestep *t* to be the absolute angle by which the cursor would miss the target had the cursor continued straight along the line connecting $\mathbf{x}_t^t$ and $\mathbf{x}_{t+1}^t$. For this evaluation, we inferred the subject's internal estimates of cursor position to be

$$\mathrm{E}[\mathbf{x}_{k+1}^t|\mathbf{y}_{t-\tau},\mathbf{u}_{t-\tau},\ldots,\mathbf{u}_k]=\mathbf{A}\mathrm{E}[\mathbf{x}_k^t|\mathbf{y}_{t-\tau},\mathbf{u}_{t-\tau},\ldots,\mathbf{u}_{k-1}]+\mathbf{B}\mathbf{u}_k+\mathbf{b}_0 \quad (13)$$

for $k \in \{t-\tau, \ldots, t\}$, where $\mathbf{A}$, $\mathbf{B}$, $\mathbf{b}_0$ and τ were fit to the training data, and the recursive expectations were initialized by (10). Note that the expectation in (13) depends only on the visual feedback available at timestep *t*, neural commands, and the subject's internal model parameters. Critically, (13) does not depend on the target position, $\mathbf{G}_t$, which will be used to evaluate the output of the internal model estimated by IME. Additionally, (13) does not require us to compute $\{c_t\}$ for held-out trials.

To determine the feedback delay, τ, we fit IME models for $\tau \in \{0, \ldots, 9\}$ and assessed model fit by examining the data likelihood, $P(\{\mathbf{G}_t\}|\{\mathbf{y}_t\}, \{\mathbf{u}_t\})$. Recall that τ controls the number of latent variables, $\mathbf{x}$, in each row of the graphical model in Fig. 1, and that the number of model parameters does not depend on the setting of τ. Fig. 4A shows the training data log-likelihood for a single BMI session across all evaluated choices of τ, and Fig. 4B gives the values of τ that maximized each training fold's data log-likelihood. A feedback delay of 3 timesteps (100 ms) most often gave the best model fit. This result agrees with reaction times we measured from BMI cursor trajectories and previously reported motor-cortical latencies to visual stimuli (Schwartz et al., 1988).

On many BMI trials, the subject drove the cursor roughly straight to the target, whereas cursor trajectories were more circuitous on other trials. It is an open question as to why the subject does not always drive the cursor straight to the target. In Figs. 5A and 5B, we show a BMI trial with a circuitous cursor trajectory, along with cross-validated IME-inferred internal estimates of cursor position from (13). In Fig. 5A, the subject's aiming intention at timestep *t*, as defined by IME predictions of the subject's internal state estimates $\mathbf{x}_t^t$ and $\mathbf{x}_{t+1}^t$, points straight toward the target. We would evaluate this aiming command as having a 0° angular error because the cursor would have hit the target had it continued along that intended aiming direction. For comparison, we computed velocity motor commands, $\mathbf{v}_t$, through the actual cursor dynamics:

$$\mathbf{v}_t=\frac{1}{\Delta t}(\boldsymbol{\beta}\mathbf{u}_t+\beta_0) \quad (14)$$

which correspond to the single-timestep (Δt = 33ms) contribution to the current cursor velocity, implicit in (8). In Fig. 5A, if the subject was aiming from a perfect prediction of the current cursor position, $\mathbf{y}_t$, (e.g., by using an internal model that exactly matched the cursor dynamics), the motor command decoded from the actual cursor dynamics would miss the target by 10°.

Despite the cursor's circuitous trajectory in this example trial, IME predictions suggest that the subject was indeed aiming straight toward the target throughout the trial (Fig. 5B). IME-

extracted internal state estimates diverge from the true cursor trajectory because the extracted internal model differs from the true cursor dynamics (8). The key insights here are that i) the subject appears to be aiming from where it *believes* the cursor is at each point in time, as determined by the internal model, and ii) the subject's neural commands, when evaluated through the internal model, are consistent with straight-to-target movements.

In Fig. 5C, we show a representative trial where the subject drove the BMI cursor along a more direct path to the target. On this trial and many similar to it, IME predictions of the subject's internal state estimates tend to agree quite closely with the actual cursor movement. This agreement contrasts with the disagreement between IME predictions and cursor movements in Figs. 5A and 5B, highlighting the redundancy in the mapping from high-dimensional neural activity to low-dimensional cursor kinematics. Certain patterns of neural activity may produce straight-to-target movements through both the subject's internal model *and* the cursor dynamics, whereas other patterns may produce straight-to-target movements through the internal model, but *not* through the cursor dynamics. On both trials in Fig. 5, IME-extracted internal models reveal the subject's internal state estimates tending toward the target, regardless of where the actual cursor moves.

Across 5,760 trials from 36 BMI experiments, we computed cross-validated angular aiming errors from IME-inferred internal state estimates (as in the red angular error from Fig. 5A) and from actual cursor positions (as in the black angular error from Fig. 5A). IME-based aiming errors were about 5 times smaller than the analogous errors in demonstrated cursor trajectories (Fig. 6). Recall that the straight-to-target aiming assumption is only applied to the training data during model fitting and not when evaluating test data. That being said, neural commands in the test data were more consistent with straight-to-target aiming under IME-extracted internal models than under the actual cursor dynamics. Internal models extracted by IME therefore explain the subject's neural commands better than the actual cursor dynamics.

5. Discussion

The developed IME framework describes the internal state estimation process, whereby the subject integrates feedback with previously issued control signals to generate an up-to-date estimate of the plant state. IME specifically addresses complicating realities inherent in learning from demonstration, including i) sensory feedback delays and ii) mismatch between the true plant model and the subject's internal model. This mismatch, often referred as *model bias*, is typically treated as an obstacle to be overcome, for example by incorporating model uncertainty in policy search (Deisenroth & Rasmussen, 2011) or through policy improvement in the presence of model inaccuracies (Abbeel et al., 2006). In this work, we seek to explicitly identify the subject's internal model, rather than attempt to factor it out.

The internal model estimation problem that we address is fundamentally different from the problem of learning the actual plant dynamics (i.e., system identification). The internal model estimation problem cannot be solved by simply relating control signals to state transitions. Actual plant state trajectories only enter IME as delayed feedback, and as such, the actual plant dynamics play only an indirect role when extracting the subject's internal model. Additionally, the objective in this work is not to design better plant dynamics, but rather to estimate the subject's knowledge of the current plant dynamics, whatever they may be. The ability to extract the subject's internal model depends critically on the consistency of control relative to the task goals and internal model, and not on the extent to which the subject's demonstrations achieve those goals through the actual plant dynamics.

Ground truth cannot be known in the BMI application. For comparison against our IME-based results, we present aiming errors resulting from evaluating recorded neural commands through the true cursor dynamics. We chose this comparison because, for optimal control, the subject should learn an internal model that matches the actual plant dynamics. Although the subject is proficient in BMI tasks, our results suggest a mismatch between the subject's internal model and the actual cursor dynamics. IME provides the opportunity to study the brain's deviations from optimality, which may ultimately enable researchers to design specific training paradigms to improve a subject's learning.

While IME requires specification of a cost function, we have demonstrated notable flexibility in this requirement. Unlike optimal control in the linear-quadratic Gaussian setting (Anderson & Moore, 1990), IME does not require pre-specification of a control time horizon. Additionally, by learning the c_t in (12), internal model estimation becomes agnostic to control effort. This feature is advantageous because the form of control costs is often unknown, especially in the BMI setting (although see (Körding & Wolpert, 2004)). In some cases, these control costs may be known a *priori*, and we are interested in extending IME to handle these scenarios (e.g., when a full quadratic cost function of state and control is specified).

One potential limitation of IME is the prescription of a static linear-Gaussian form to the internal model. Some applications may require more expressive internal model dynamics, which could be accommodated by a nonparametric representation of the internal model (Deisenroth et al., 2012). Additionally, the ability to extract a time-varying internal model may be needed in settings where the subject updates an internal model as control experience accrues.

Acknowledgments

This work was supported by NSF IGERT Fellowship (MDG), NIH-NICHD-CRCNS-R01-HD-071686 (BMY), and PA Department of Health Research Formula Grant SAP#4100057653 under the Commonwealth Universal Research Enhancement program (SMC). We thank A. Schwartz for access to BMI data. We thank P. Abbeel and Z. Kolter for insightful comments on a preliminary version of this paper.

APPENDIX

EM algorithm for the IME probabilistic framework described in Section 3:

Expectation Step

$$\boldsymbol{\mu}_{t-\tau}^{t} = \mathbf{H}\mathbf{y}_{t-\tau}$$
$$\boldsymbol{\mu}_{k+1}^{t} = \mathbf{A}\boldsymbol{\mu}_{k}^{t} + \mathbf{B}\mathbf{u}_{k} + \mathbf{b}_{0}$$
$$\boldsymbol{\Sigma}_{\mathbf{x}_{t-\tau}^{t},\mathbf{x}_{t-\tau}^{t}} = \mathbf{W}_{0}$$
$$\boldsymbol{\Sigma}_{\mathbf{x}_{k+1}^{t},\mathbf{x}_{k+1}^{t}} = \mathbf{A}\boldsymbol{\Sigma}_{\mathbf{x}_{k}^{t},\mathbf{x}_{k}^{t}}\mathbf{A}' + \mathbf{W}$$
$$\boldsymbol{\Sigma}_{\mathbf{x}_{k}^{t},\mathbf{x}_{k+d+1}^{t}} = \boldsymbol{\Sigma}_{\mathbf{x}_{k}^{t},\mathbf{x}_{k+d}^{t}}\mathbf{A}'$$

$$\boldsymbol{\Sigma}_{\mathbf{x}}=\begin{bmatrix} \boldsymbol{\Sigma}_{\mathbf{x}_{t-\tau}^{t},\mathbf{x}_{t-\tau}^{t}} & \cdots & \boldsymbol{\Sigma}_{\mathbf{x}_{t-\tau}^{t},\mathbf{x}_{t+1}^{t}} \\ \vdots & \ddots & \vdots \\ \boldsymbol{\Sigma}_{\mathbf{x}_{t+1}^{t},\mathbf{x}_{t-\tau}^{t}} & \cdots & \boldsymbol{\Sigma}_{\mathbf{x}_{t+1}^{t},\mathbf{x}_{t+1}^{t}} \end{bmatrix}$$

$$\boldsymbol{\mu}_{t}=\begin{bmatrix} \boldsymbol{\mu}_{t-\tau}^{t} \\ \vdots \\ \boldsymbol{\mu}_{t+1}^{t} \end{bmatrix}$$

$$\boldsymbol{\Gamma}_{t}=\mathbf{C}_{t}\boldsymbol{\mu}_{t+1}^{t}$$

$$\boldsymbol{\Sigma}_{\mathbf{G}_{t}}=\mathbf{C}_{t}\boldsymbol{\Sigma}_{\mathbf{x}_{t+1}^{t},\mathbf{x}_{t+1}^{t}}\mathbf{C}_{t}'+\mathbf{V}$$

$$\boldsymbol{\Sigma}_{\mathbf{x},\mathbf{G}_{t}}=\begin{bmatrix} \boldsymbol{\Sigma}_{\mathbf{x}_{t-\tau}^{t},\mathbf{x}_{t+1}^{t}} \\ \vdots \\ \boldsymbol{\Sigma}_{\mathbf{x}_{t+1}^{t},\mathbf{x}_{t+1}^{t}} \end{bmatrix}\mathbf{C}_{t}'$$

$$\tilde{\mu}_{t}=\boldsymbol{\mu}_{t}+\boldsymbol{\Sigma}_{\mathbf{x},\mathbf{G}_{t}}\boldsymbol{\Sigma}_{\mathbf{G}_{t}}^{-1}(\mathbf{G}_{t}-\boldsymbol{\Gamma}_{t})$$

$$\tilde{\boldsymbol{\Sigma}}_{\mathbf{x}}=\boldsymbol{\Sigma}_{\mathbf{x}}-\boldsymbol{\Sigma}_{\mathbf{x},\mathbf{G}_{t}}\boldsymbol{\Sigma}_{\mathbf{G}_{t}}^{-1}\boldsymbol{\Sigma}_{\mathbf{x},\mathbf{G}_{t}}'$$

Maximization Step

$$\mathbf{H}=\left(\sum_{t=1}^{T}\tilde{\mu}_{t-\tau}^{t}(\mathbf{y}_{t-\tau})'\right)\left(\sum_{t=1}^{T}\mathbf{y}_{t-\tau}(\mathbf{y}_{t-\tau})'\right)^{-1}$$

$$\mathbf{W}_{0}=\frac{1}{T}\sum_{t=1}^{T}(\tilde{\boldsymbol{\Sigma}}_{\mathbf{x}_{t-\tau}^{t},\mathbf{x}_{t-\tau}^{t}}+\tilde{\mu}_{t-\tau}^{t}(\tilde{\mu}_{t-\tau}^{t})'-\mathbf{H}\mathbf{y}_{t-\tau}(\tilde{\mu}_{t-\tau}^{t})')$$

$$\mathbf{P}_{k}^{t}=[\tilde{\boldsymbol{\Sigma}}_{\mathbf{x}_{k+1}^{t},\mathbf{x}_{k}^{t}}+\tilde{\mu}_{k+1}^{t}(\tilde{\mu}_{k}^{t})'\tilde{\mu}_{k+1}^{t}\mathbf{u}_{k}'\tilde{\mu}_{k+1}^{t}]$$

$$\mathbf{Q}_{k}^{t}=\begin{bmatrix} \tilde{\boldsymbol{\Sigma}}_{\mathbf{x}_{k}^{t},\mathbf{x}_{k}^{t}}+\tilde{\mu}_{k}^{t}(\tilde{\mu}_{k}^{t})' & \tilde{\mu}_{k+1}^{t}\mathbf{u}_{k}' & \tilde{\mu}_{k}^{t} \\ \mathbf{u}_{k}(\tilde{\mu}_{k}^{t})' & \mathbf{u}_{k}\mathbf{u}_{k}' & \mathbf{u}_{k} \\ (\tilde{\mu}_{k}^{t})' & \mathbf{u}_{k}' & 1 \end{bmatrix}$$

$$\mathbf{M}=[\mathbf{A}\mathbf{B}\mathbf{b}_{0}]=\left(\sum_{t=1}^{T}\sum_{k=t-\tau}^{T}\mathbf{P}_{k}^{t}\right)\left(\sum_{t=1}^{T}\sum_{k=t-\tau}^{T}\mathbf{Q}_{k}^{t}\right)^{-1}$$

$$\mathbf{W}=\frac{1}{\tau\times T}\sum_{t=1}^{T}\sum_{k=t-\tau}^{T}\left(\tilde{\boldsymbol{\Sigma}}_{\mathbf{x}_{k}^{t},\mathbf{x}_{k}^{t}}+\tilde{\mu}_{k}^{t}(\tilde{\mu}_{k}^{t})'-\mathbf{M}\mathbf{P}_{k}^{t\,\prime}\right)$$

$$\mathbf{C}=\left(\sum_{t=1}^{T}\mathbf{G}_{t}(\tilde{\mu}_{t+1}^{t})'\right)\left(\sum_{t=1}^{T}\tilde{\boldsymbol{\Sigma}}_{\mathbf{x}_{k+1}^{t},\mathbf{x}_{k+1}^{t}}+\tilde{\mu}_{t+1}^{t}(\tilde{\mu}_{t+1}^{t})'\right)^{-1}$$

$$\mathbf{V}=\frac{1}{T}\sum_{t=1}^{T}(\mathbf{G}_{t}\mathbf{G}_{t}'-\mathbf{C}_{t}\tilde{\mu}_{t+1}^{t}\mathbf{G}_{t}')$$

For the IME variant from Section 4, update the above Estep with the following:

$$\mathbf{H}=\mathbf{I}$$
$$\mathbf{W}_0=\mathbf{0}$$
$$\mathbf{W}=\sigma_{\mathbf{x}}^2\mathbf{I}$$
$$\mathbf{V}=\sigma_{\mathbf{G}}^2\mathbf{I}$$
$$\mathbf{C}_t=[(1+c_t)\mathbf{I}-c_t\mathbf{I}]$$
$$\mathbf{\Gamma}_t=\mathbf{C}_t\begin{bmatrix}\boldsymbol{\mu}_{t+1}^t\\ \boldsymbol{\mu}_t^t\end{bmatrix}$$
$$\mathbf{\Sigma}_{\mathbf{G}_t}=\mathbf{C}_t\begin{bmatrix}\mathbf{\Sigma}_{\mathbf{x}_{t+1}^t,\mathbf{x}_{t+1}^t} & \mathbf{\Sigma}_{\mathbf{x}_{t+1}^t,\mathbf{x}_t^t}\\ \mathbf{\Sigma}_{\mathbf{x}_t^t,\mathbf{x}_{t+1}^t} & \mathbf{\Sigma}_{\mathbf{x}_t^t,\mathbf{x}_t^t}\end{bmatrix}\mathbf{C}_t'+\mathbf{V}$$
$$\mathbf{\Sigma}_{\mathbf{x},\mathbf{G}_t}=\begin{bmatrix}\mathbf{\Sigma}_{\mathbf{x}_{t-\tau}^t,\mathbf{x}_{t+1}^t} & \mathbf{\Sigma}_{\mathbf{x}_{t-\tau}^t,\mathbf{x}_t^t}\\ \vdots & \vdots\\ \mathbf{\Sigma}_{\mathbf{x}_{t+1}^t,\mathbf{x}_{t+1}^t} & \mathbf{\Sigma}_{\mathbf{x}_{t+1}^t,\mathbf{x}_t^t}\end{bmatrix}\mathbf{C}_t'$$

and in the M-step, take:

$$\sigma_{\mathbf{x}}^2=\frac{1}{2}tr(\mathbf{W})\sigma_{\mathbf{G}}^2=\frac{1}{2}tr(\mathbf{V})$$
$$\alpha_{1,t}=(\tilde{\mu}_{t+1}^t-\tilde{\mu}_t^t)(\mathbf{G}_t-\tilde{\mu}_{t+1}^t)'+\tilde{\mathbf{\Sigma}}_{\mathbf{x}_t^t,\mathbf{x}_{t+1}^t}-\tilde{\mathbf{\Sigma}}_{\mathbf{x}_{t+1}^t,\mathbf{x}_{t+1}^t}$$
$$\alpha_{2,t}=(\tilde{\mu}_{t+1}^t-\tilde{\mu}_t^t)(\tilde{\mu}_{t+1}^t-\tilde{\mu}_t^t)'+\tilde{\mathbf{\Sigma}}_{\mathbf{x}_{t+1}^t,\mathbf{x}_{t+1}^t}+\tilde{\mathbf{\Sigma}}_{\mathbf{x}_t^t,\mathbf{x}_t^t}-2\tilde{\mathbf{\Sigma}}_{\mathbf{x}_{t+1}^t,\mathbf{x}_t^t}$$
$$c_t=\max(0,\frac{tr(\alpha_{1,t})}{tr(\alpha_{2,t})})$$

References

Abbeel, P.; Ng, AY. Apprenticeship learning via inverse reinforcement learning; Proc. 21st International Conf. on Machine Learning; 2004. p. 1-8.

Abbeel, P.; Quigley, M.; Ng, AY. Using inaccurate models in reinforcement learning; Proc. 23rd International Conf. on Machine learning; 2006. p. 1-8.

Anderson, BDO.; Moore, JB. Optimal control: linear quadratic methods. Upper Saddle River, NJ, USA: Prentice-Hall, Inc; 1990.

Boyd, S.; El Ghaoui, L.; Feron, E.; Balakrishnan, V. Studies in Applied Mathematics. Philadelphia, PA: 1994. Linear Matrix Inequalities in System and Control Theory.

Chase SM, Kass RE, Schwartz AB. Behavioral and neural correlates of visuomotor adaptation observed through a brain-computer interface in primary motor cortex. J. Neurophysiol. 2012; 108(2):624–644. [PubMed: 22496532]

Coates, A.; Abbeel, P.; Y Ng, AY. Learning for control from multiple demonstrations; Proc. 25th International Conf. on Machine Learning; 2008. p. 144-151.

Crapse TB, Sommer MA. Corollary discharge across the animal kingdom. Nat. Rev. Neurosci. 2008; 9(8):587–600. [PubMed: 18641666]

Deisenroth, MP.; Rasmussen, CE. Pilco: A model-based and data-efficient approach to policy search; Proc. 28th International Conf. on Machine Learning; 2011. p. 465-472.

Deisenroth MP, Turner RD, Huber MF, Hanebeck UD, Rasmussen CE. Robust filtering and smoothing with gaussian processes. IEEE Trans. Automatic Control. 2012; 57(7):1865–1871.

Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. J. Royal Stat. Soc., Series B. 1977; 39(1):1–38.

Ganguly K, Dimitrov DF, Wallis JD, Carmena JM. Reversible large-scale modification of cortical networks during neuroprosthetic control. Nature Neurosci. 2011; 14(5):662–667. [PubMed: 21499255]

Georgopoulos AP, Caminiti R, Kalaska JF, Massey JT. Spatial coding of movement: a hypothesis concerning the coding of movement direction by motor cortical populations. Exp. Brain Res. Suppl. 1983; 7:327–336.

Green AM, Kalaska JF. Learning to move machines with the mind. Trends Neurosci. 2011; 34(2):61–75. [PubMed: 21176975]

Körding KP, Wolpert DM. The loss function of sensorimotor learning. Proc. Natl. Acad. Sci. 2004; 101(26):9839–9842. [PubMed: 15210973]

Miall RC, Wolpert DM. Forward models for physiological motor control. Neural Networks. 1996; 9(8):1265–1279. [PubMed: 12662535]

Ng, AY.; Russell, S. Algorithms for inverse reinforcement learning; Proc. 17th International Conf. on Machine Learning; 2000. p. 663-670.

Ratliff, ND.; Bagnell, JA.; Zinkevich, Martin. Maximum margin planning; Proc. 23rd International Conf. on Machine Learning; 2006. p. 729-736.

Schaal S. Is imitation learning the route to humanoid robots? Trends in Cognitive Sciences. 1999; 3(6): 233–242. [PubMed: 10354577]

Schwartz AB, Kettner RE, Georgopoulos AP. Primate motor cortex and free arm movements to visual targets in 3-dimensional space. 1. relations between single cell discharge and direction of movement. J. Neurosci. 1988; 8(8):2913–2927. [PubMed: 3411361]

Taylor DM, Helms Tillery SI, Schwartz AB. Direct cortical control of 3D neuroprosthetic devices. Science. 2002; 296:1829–1832. [PubMed: 12052948]

Ziebart, BD.; Maas, AL.; Bagnell, JA.; Dey, AK. Maximum entropy inverse reinforcement learning; Proc. 23rd AAAI Conf. on Artificial Intelligence; 2008. p. 1433-1438.

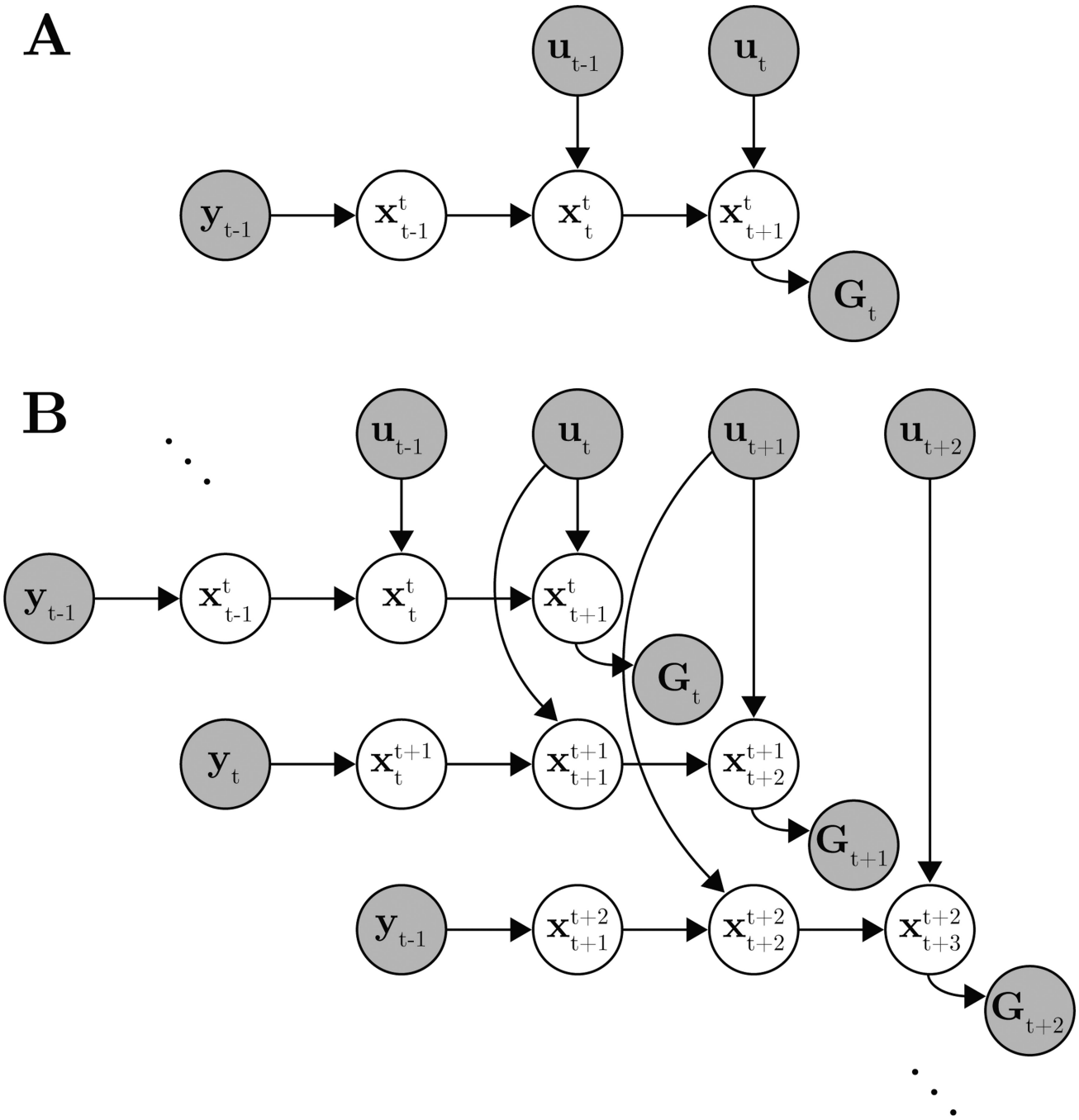

Figure 1.

IME graphical model with a single timestep feedback delay ($\tau = 1$). Observed variables are indicated as shaded nodes, and unshaded nodes are latent variables. (A) Timestep *t* slice of IME. Delayed sensory feedback of plant state, $\mathbf{y}_{t-1}$, and previously issued control input, $\mathbf{u}_{t-1}$, propagate through the subject's internal dynamics model to generate internal estimates of plant state, $\mathbf{x}_{t-1}^{t}$ and $\mathbf{x}_{t}^{t}$. With belief $\mathbf{x}_{t}^{t}$ of the current plant state, the subject generates a new control signal, $\mathbf{u}_t$. The internal model predicts the resulting future state, $\mathbf{x}_{t+1}^{t}$, which should agree with the current task goal, $\mathbf{G}_t$. (B) The full IME graphical model. The timestep *t* slice from (A) is embedded in the upper left corner.

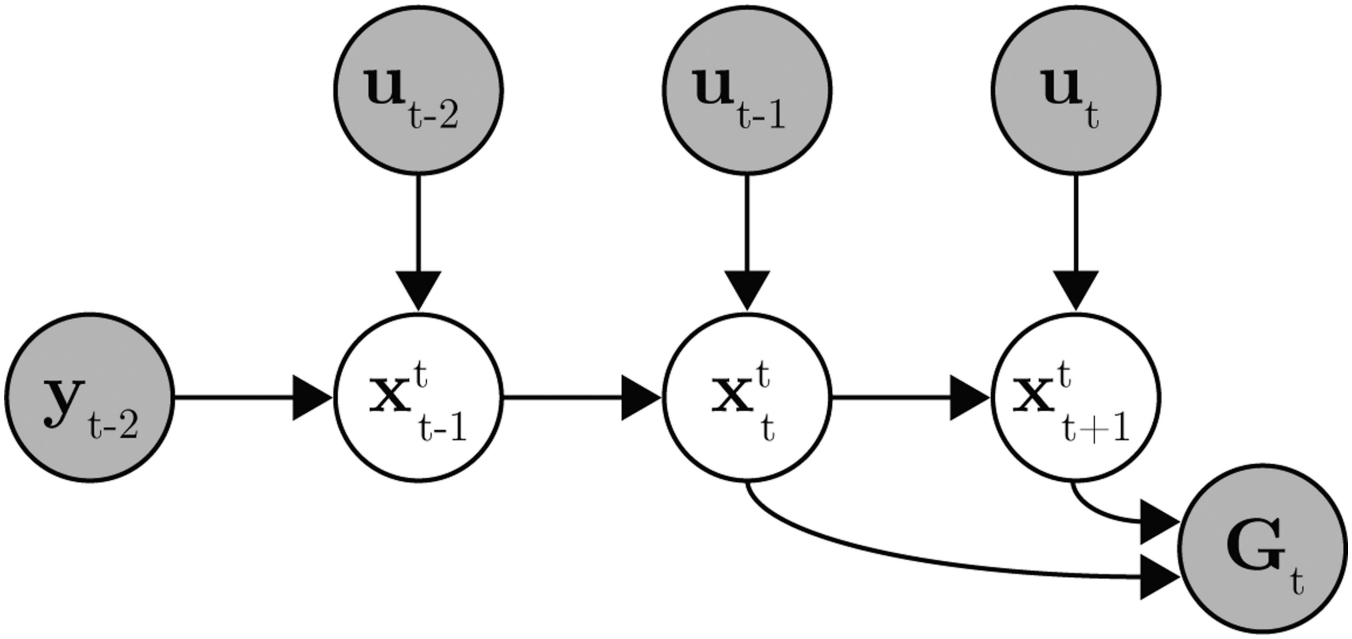

Figure 2.
Single-timestep slice of a two-timestep feedback delay ($\tau = 2$) IME graphical model, as used in the BMI application.

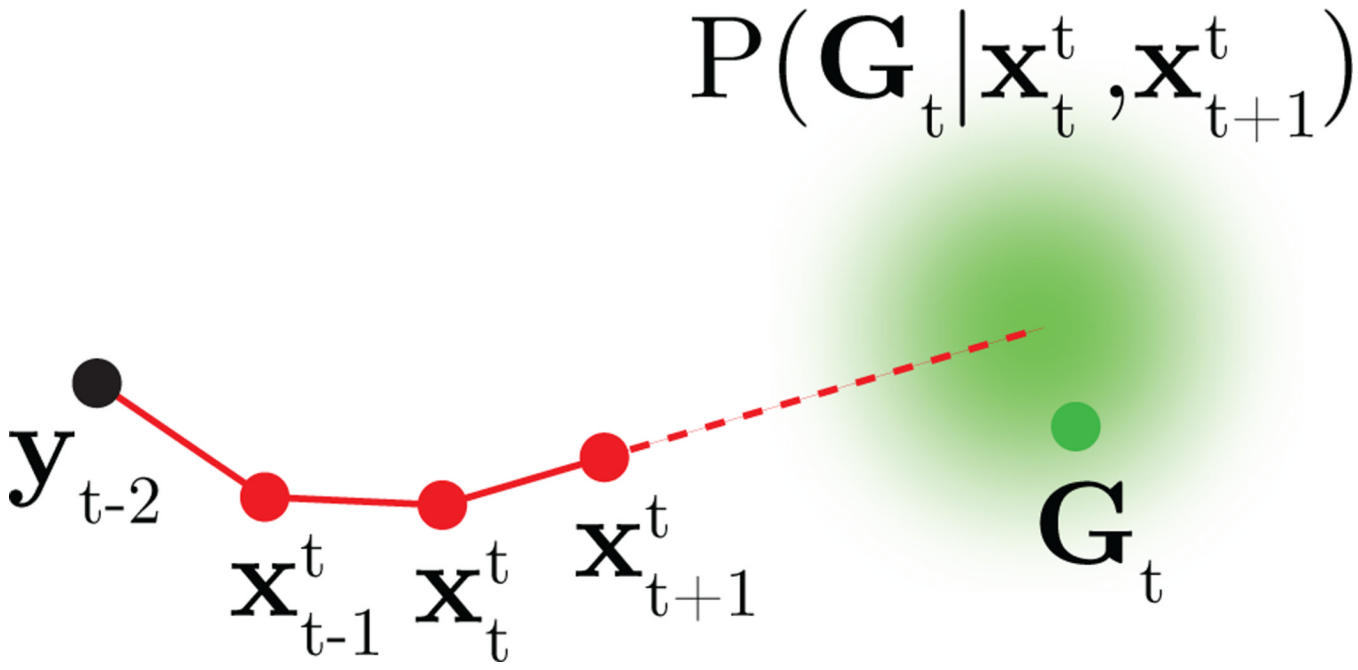

Figure 3.

Task goals encoded by the IME probabilistic model. For illustration, we consider a two-timestep visual feedback delay. At timestep *t*, the subject's internal estimate of the cursor position two timesteps ago matches the actual cursor position (black point). The subject's internal state estimates (red points) evolve according to (11). If the subject intends to drive the cursor straight to the target position, $\mathbf{G}_t$ (green point), from the subject's estimate, $\mathbf{x}_t^t$, of the current cursor position, then according to (12), the target should lie near the line defined by the two points, $\mathbf{x}_t^t$ and $\mathbf{x}_{t+1}^t$ (dashed red line).

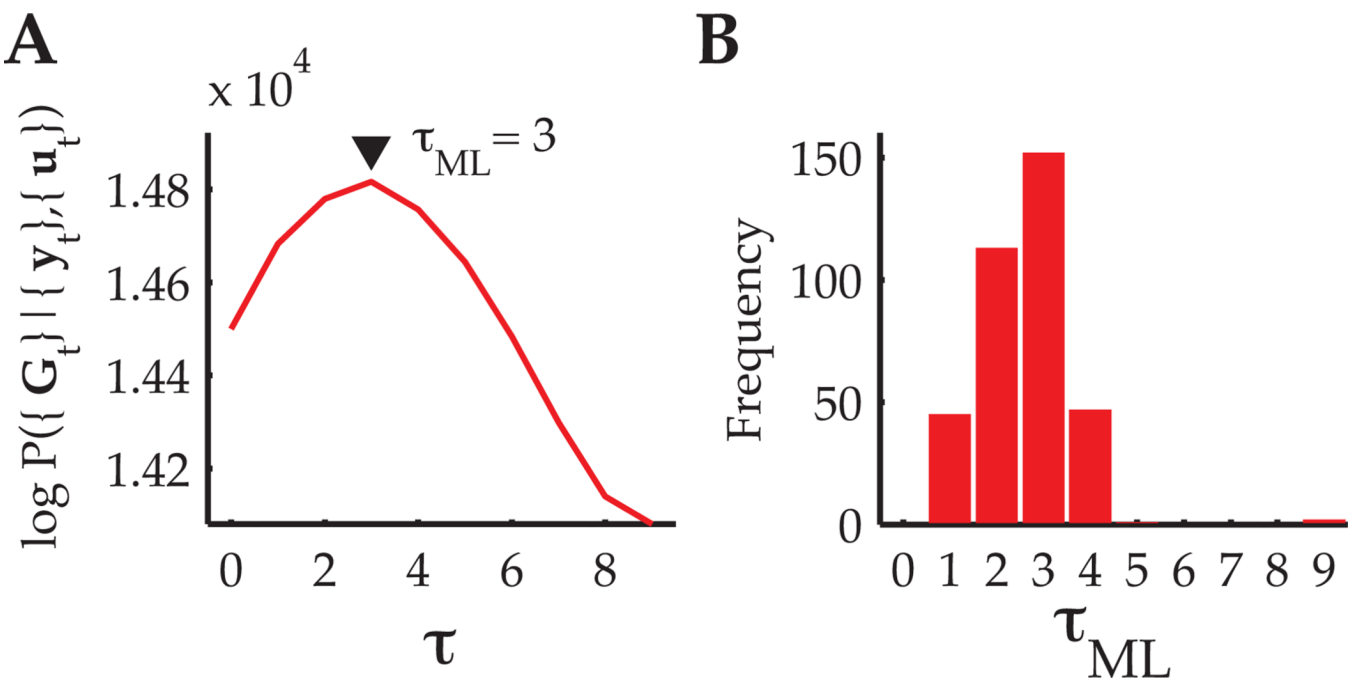

Figure 4.

Model selection for determining the feedback delay, τ. (A) Log-likelihood as a function of τ over training data from a single BMI experiment. For this representative experiment, the feedback delay that maximized the data likelihood was τ_{ML} = 3 timesteps (100 ms). (B) Histogram of τ_{ML} values across 36 experiments (10 folds each).

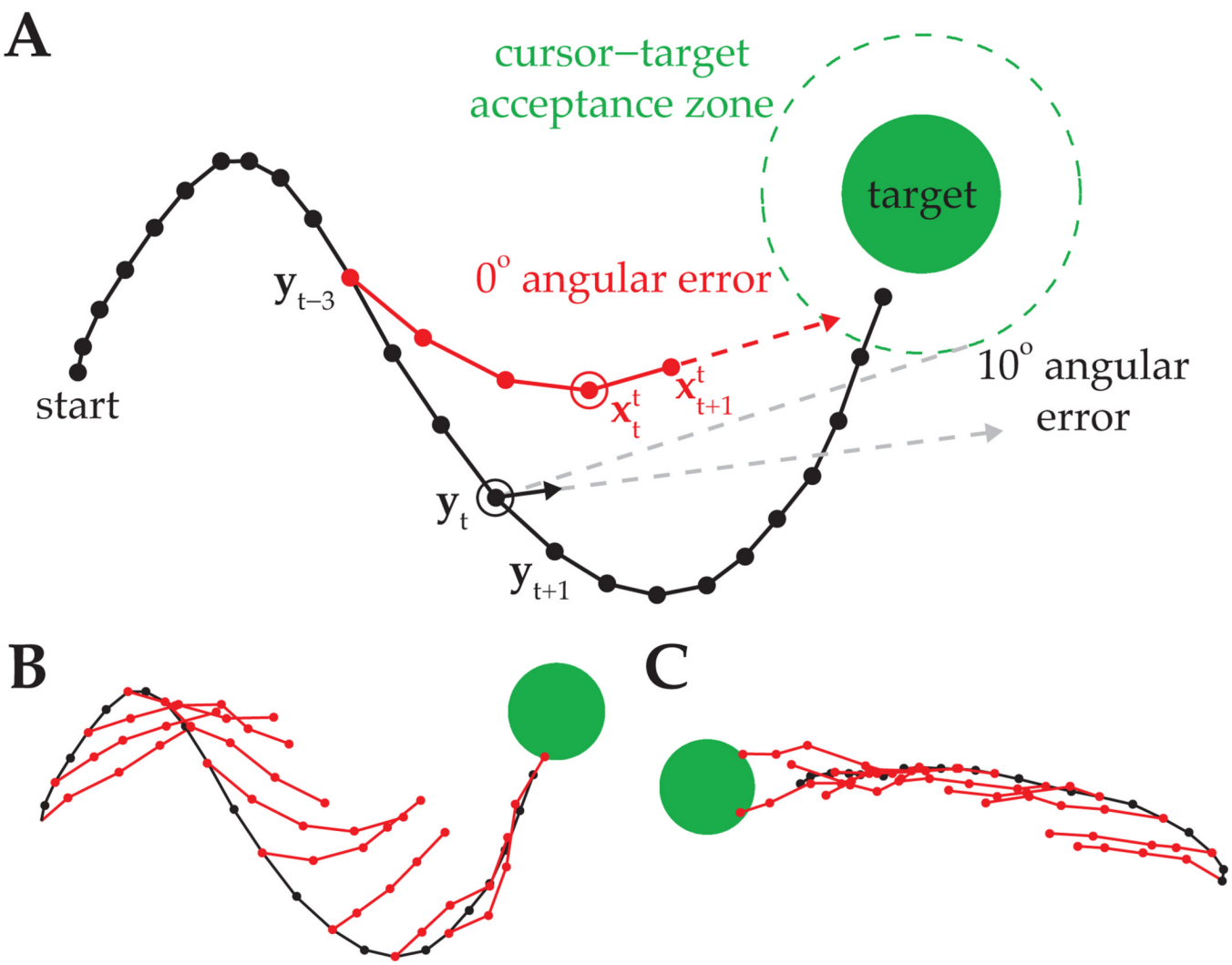

Figure 5.

Evaluation of IME predictions. (A) Cursor trajectory (black line) from a BMI trial that was not used in model fitting. Red points are IME-inferred internal estimates of cursor position as determined by (13). Black arrow is the single-timestep velocity command, $\mathbf{v}_t$, as computed through the actual cursor dynamics (14). Task success occurs when the cursor visibly overlaps with the target, which happens when the center of the cursor enters the cursor-target overlap zone (dashed green circle). (B) IME-inferred internal state chains across the entire BMI cursor trajectory from (A). (C) BMI cursor trajectory and IME-inferred internal state chains from a different trial.

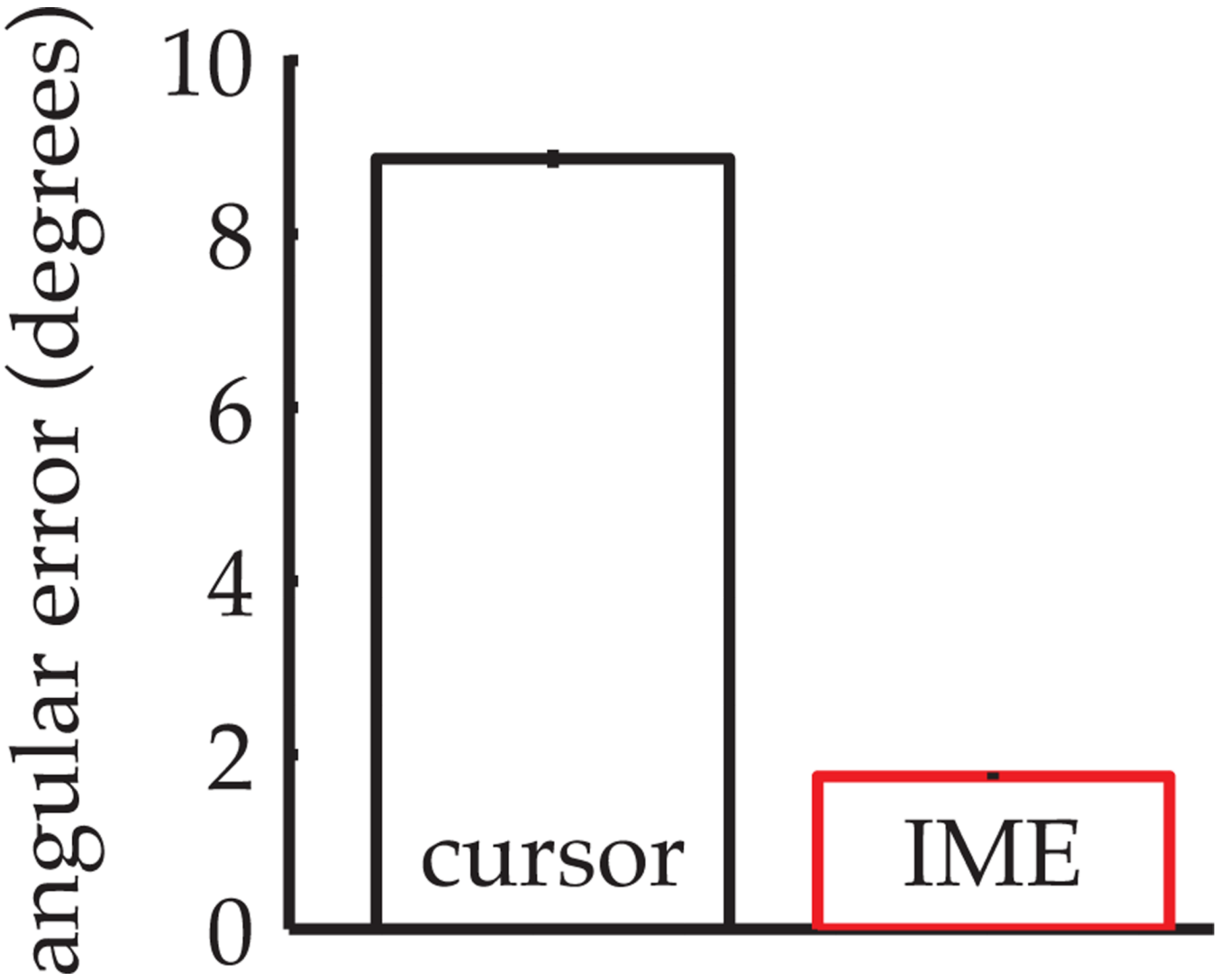

Figure 6.

Angular aiming errors based on internal models estimated by IME (red bar) compared to errors based on internal models which exactly match the cursor dynamics (black bar). Angular errors were first averaged within each trial, then averaged across all 160 trials from each of 36 BMI experiments. Error bars indicate +/− SEM (*n*=5,760).