Biological Cybernetics
https://doi.org/10.1007/s00422-022-00922-z

Biological Cybernetics

ORIGINAL ARTICLE

Human motor learning is robust to control-dependent noise

Bo Pang[1] · **Leilei Cui**[1] · **Zhong-Ping Jiang**[1]

/ Accepted: 6 February 2022

Abstract
Noises are ubiquitous in sensorimotor interactions and contaminate the information provided to the central nervous system (CNS) for motor learning. An interesting question is how the CNS manages motor learning with imprecise information. Integrating ideas from reinforcement learning and adaptive optimal control, this paper develops a novel computational mechanism to explain the robustness of human motor learning to the imprecise information, caused by control-dependent noise that exists inherently in the sensorimotor systems. Starting from an initial admissible control policy, in each learning trial the mechanism collects and uses the noisy sensory data (caused by the control-dependent noise) to form an imprecise evaluation of the performance of the current policy and then constructs an updated policy based on the imprecise evaluation. As the number of learning trials increases, the generated policies mathematically provably converge to a (potentially small) neighborhood of the optimal policy under mild conditions, despite the imprecise information in the learning process. The mechanism directly synthesizes the policies from the sensory data, without identifying an internal forward model. Our preliminary computational results on two classic arm reaching tasks are in line with experimental observations reported in the literature. The model-free control principle proposed in the paper sheds more lights into the inherent robustness of human sensorimotor systems to the imprecise information, especially control-dependent noise, in the CNS.

Keywords Robustness · Reinforcement learning · Policy iteration · Sensorimotor control · Arm reaching

1 Introduction

Although goal-directed movements have been extensively studied in the field of sensorimotor control in the past decades, its underlying computational mechanism is still not fully understood yet (Todorov and Jordan 2002; Harris and Wolpert 1998; Franklin et al. 2003; Burdet et al. 2001, 2006; Wolpert et al. 1995; Selen et al. 2009; Zhou et al. 2017; Kadiallah et al. 2011; Mistry et al. 2013; Česonis and Franklin 2020). Various computational models have been proposed to account for sensorimotor control and learning (Shadmehr and Mussa-Ivaldi 2012; Krakauer et al. 2019). One widely accepted theory is that the CNS selects trajectories that minimize a cost function (Flash and Hogan 1985; Uno et al. 1989; Harris and Wolpert 1998; Todorov 2005; Jiang and Jiang 2014; Bian et al. 2020). In particular, the authors of Todorov and Jordan (2002) suggest that the CNS uses a model-based optimal feedback principle to coordinate body movement by minimizing an integral-quadratic cost index that trades off energy consumption with constraints. Such optimal control frameworks have been found to successfully explain diverse phenomena, such as approximately straight movement trajectories and bell-shaped velocity curves (Morasso 1981), variability patterns and flexibility of arm movement trajectories (Todorov and Jordan 2002; Liu and Todorov 2007), adaptation to force fields and visuomotor transforms (Ueyama 2014; Braun et al. 2009), kinematic invariance despite the sacrifice of optimality (Mistry et al. 2013), fast timescale of motor learning (Crevecoeur et al. 2020), to name a few. A common assumption in these studies is that the CNS first identifies the system dynamics and then solves the optimal control problem using the identified model. (This

Communicated by Benjamin Lindner.

This work has been supported in part by the U.S. National Science Foundation under Grants ECCS-1501044 and EPCN-1903781.

✉ Bo Pang
bo.pang@nyu.edu

Leilei Cui
l.cui@nyu.edu

Zhong-Ping Jiang
zjiang@nyu.edu

[1] Department of Electrical and Computer Engineering, New York University, 370 Jay Street, Brooklyn, NY 11201, USA

Published online: 03 March 2022

kind of mechanism is referred to as *model-based* mechanism, according to Haith and Krakauer (2013)). However, currently there is no strong experimental evidence about how the CNS manages to generate an internal representation of the environment, especially for complex environments. To this end, model-free learning approaches, such as reinforcement learning (RL) and adaptive dynamic programming (ADP), are utilized to explain sensorimotor learning behavior (Haith and Krakauer 2013; d'Acremont et al. 2009; Fiete et al. 2007; Jiang and Jiang 2014; Bian et al. 2020). RL and ADP are biologically inspired learning approaches that study how an agent iteratively modifies its control policies toward finding the optimal policy maximizing a cumulative reward function, directly using the observed responses from its interactions with the environment (Sutton and Barto 2018; Bertsekas 2019; Jiang et al. 2020; Jiang and Jiang 2017). Thus, an intermediate internal representation of the environment is not needed in the computation of the optimal control policy anymore. (This kind of mechanism is referred to as a *model-free* mechanism, according to Haith and Krakauer (2013)). The computational models based on RL and ADP also succeed in explaining many experimental observations in sensorimotor control and learning, see (Fiete et al. (2007), d'Acremont et al. (2009), Huang et al. (2011), Izawa and Shadmehr (2011), Shmuelof et al. (2012), Haith and Krakauer (2013), Jiang and Jiang (2014), Vaswani et al. (2015) and Bian et al. (2020)). Some other related but independent works include the iterative model reference adaptive control framework presented in Zhou et al. (2011), and the direct policy updating framework (Hadjiosif et al. 2021). The main difference between these models and the RL/ADP models is that the former is not based on the optimal control framework.

Noise exists at all levels of sensorimotor interactions (Parker et al. 2002; Orbán and Wolpert 2011). Sensory inputs are noisy, which limits the accuracy of our perception. Motor commands are also noisy, which leads to inaccurate movements. The noisy sensory inputs and motor commands further result in imprecise estimation of the state of the environment and our body, and ambiguity of the parameters that characterize the task (Orbán and Wolpert 2011). Although the CNS may use mechanisms in the style of Kalman filter or Bayesian integration to minimize the effects (estimation errors) caused by the noise in sensorimotor interactions (Parker et al. 2002; Körding and Wolpert 2004; Körding and Wolpert 2006; Wolpert 2007; Orbán and Wolpert 2011; Bach and Dolan 2012), the estimation errors can never be completely suppressed (Sternad et al. 2011; Acerbi et al. 2017). Then, a natural question to ask is: How does the CNS manage to learn near-optimal policies or adapt to the new environment, in the presence of estimation errors? Motor adaptation and learning often involve iterative (trial-by-trial) improvement processes (Haith and Krakauer 2013). From the computational perspective, even small errors in the iterative processes may accumulate or be amplified over the iterations, to finally cause divergence or failure of the process (Bertsekas 2011). Thus, it is a nontrivial question why the learning performance of CNS is not affected by the estimation errors, or equivalently, why the learning performance of the CNS is robust to estimation errors in the learning processes. This question is not addressed in most of the model-based mechanisms mentioned previously, since the internal models are often assumed perfect and accurate. Although the effects of the parameter estimation errors for the uncertain force fields are explicitly investigated in the computational models proposed in Mistry et al. (2013) and Crevecoeur et al. (2020), the partial knowledge of the internal model is still needed there. Most of the model-free mechanisms mentioned in the last paragraph have no formal theoretical treatment on this issue either. The computational models proposed in Zhou et al. (2011) and Hadjiosif et al. (2021) are able to adjust the control policies iteratively without formulating an intermediate internal model, by utilizing the estimation errors of the sensory output through model reference adaptive control and direct policy updating, respectively. However, these models do not reflect the objective of minimizing the metabolic cost, which can be naturally embedded in the optimal control framework and is widely deemed to be one of the underlying principles which the CNS obeys to choose the control policies (Todorov and Jordan 2002; Liu and Todorov 2007; Burdet et al. 2001; Franklin et al. 2008; Selen et al. 2009; Franklin and Wolpert 2011).

With the above discussions in mind, in this paper we argue that our recently developed robust reinforcement learning theory (Bian and Jiang 2019; Pang et al. 2021) provides a new model-free adaptive optimal control principle candidate for explaining the robustness features observed in human motor adaptation and learning. Previous theoretical studies of RL and ADP often implicitly assume that the algorithms can be implemented or solved exactly without any estimation errors, which is a strong assumption since model uncertainties and noisy data are common in reality. In Bian and Jiang (2019) and Pang et al. (2021), it is shown by theoretical analysis that the value iteration and the policy iteration, two main classes of learning algorithms in RL and ADP, are robust to errors in the learning process aimed at solving a linear quadratic regulator (LQR) problem. More concretely, in Pang et al. (2021) we prove that the policy iteration algorithm is small-disturbance input-to-state stable. In other words, whenever the estimation error in each iteration is bounded and small, the solutions of the policy iteration algorithm are also bounded and enter a small neighborhood of the optimal solution of the LQR problem. In light of this robustness result, we propose a novel model-free computational model in this paper, named optimistic least-squares policy iteration (O-LSPI), to explain the robustness of the learning and adaptation of the CNS in the arm reaching task. We demonstrate through numeri-

cal studies that although the unmeasurable control-dependent noise in the human arm movement model introduces estimation errors into the learning algorithm, O-LSPI is still capable of finding near-optimal policies in different force fields and producing nearly identical results as observed in experiments conducted by Burdet et al. (2001, 2006) and Franklin et al. (2003).

The rest of this paper is organized as follows: Sect. 2 introduces the robust reinforcement learning, or more precisely the O-LSPI algorithm, as a novel computational principle of human movement, in the context of the general LQR problem with control-dependent stochastic noise. In Sect. 3, the proposed O-LSPI algorithm is applied to the human movement model, to reproduce the arm reaching task in simulation. Section 4 presents some discussions about the proposed mechanism. Section 5 closes the paper with some concluding remarks.

2 Theory of robust reinforcement learning

2.1 Problem formulation

Consider linear stochastic systems with control-dependent noises

$$\mathrm{d}x = (Ax + Bu)\mathrm{d}t + B\sum_{k=1}^{q} C_k u \mathrm{d}w_k, \tag{1}$$

where $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times m}$ are constant matrices describing the system dynamics, $u \in \mathbb{R}^m$ is the control signal, $C_k \in \mathbb{R}^{m\times m}$ is the gain matrix of control-dependent noise, and w_k is independent one-dimensional Brownian motion for $k = 1, 2, \ldots, q$. It is assumed that pair (A, B) is controllable. The control-dependent noise in (1) is used to capture the psychophysical observations that the variability of motor errors increases with the magnitude of the movement (Harris and Wolpert 1998; Liu and Todorov 2007). Although the actual human arm system is nonlinear due to the complex behaviors of its components, e.g., tendon and muscles, as demonstrated in Harris and Wolpert (1998), Liu and Todorov (2007), Zhou et al. (2011), Crevecoeur et al. (2020) and Mistry et al. (2013), for the simple arm reaching task considered in this paper (introduced in detail in the next section), the nonlinear dynamics can be linearized (Khalil 2002) and approximated well using the linear dynamics (1).

Following Todorov and Jordan (2002) and Liu and Todorov (2007), the optimal control problem is to find an optimal control policy to minimize the following cost with respect to the nominal system of (1) without the control-dependent noise

$$J(x(0), u) = \int_0^\infty \left(x^{\mathrm{T}} Q x + u^{\mathrm{T}} R u\right) \mathrm{d}t, \tag{2}$$

where $Q \in \mathbb{S}^n$ and $R \in \mathbb{S}^m$ are positive definite constant weighting matrices, with $\mathbb{S}^n$ denoting the set of all real symmetric matrices of order n. It is well known (Liberzon 2012, Section 6.2.2) that the optimal control policy is $u^* = -K^* x$, where $K^* = R^{-1}B^{\mathrm{T}}P^*$ and $P^* \in \mathbb{S}^n$ is the unique positive definite solution of the algebraic Riccati equation (ARE)

$$A^{\mathrm{T}}P + PA + Q - PBR^{-1}B^{\mathrm{T}}P = 0. \tag{3}$$

In addition, K^* is stabilizing in the sense that $A - BK^*$ is Hurwitz or its eigenvalues have negative real parts. Define

$$\mathcal{A}(K) = I_n \otimes (A - BK)^{\mathrm{T}} + (A - BK)^{\mathrm{T}} \otimes I_n + \sum_{k=1}^{q} (BC_kK)^{\mathrm{T}} \otimes (BC_kK)^{\mathrm{T}}.$$

As it can be directly checked (Kleinman 1969), the system (1) in closed loop with $u = -Kx$ is mean-square stable in the sense of Willems and Willems (1976 Definition 1.) if $\mathcal{A}(K)$ is Hurwitz. In particular, if K is stabilizing and the gain matrices C_k of the control-independent noise are small enough, then $\mathcal{A}(K)$ is Hurwitz.

2.2 Policy iteration

Notice that (3) is a nonlinear matrix equation in P, which is hard to be directly solved. Policy iteration is an iterative method to find P^* by solving successively a sequence of transformed linear matrix equations.

For any $P \in \mathbb{S}^n$ and any $K \in \mathbb{R}^{m\times n}$, define

$$G(P) \triangleq \begin{bmatrix} Q + A^{\mathrm{T}}P + PA & PB \\ B^{\mathrm{T}}P & R \end{bmatrix} = \left[\begin{array}{c|c} [G(P)]_{xx} & [G(P)]_{ux}^{\mathrm{T}} \\ \hline [G(P)]_{ux} & [G(P)]_{uu} \end{array}\right].$$

and

$$\mathcal{H}(G(P), K) = \begin{bmatrix} I_n & -K^{\mathrm{T}} \end{bmatrix} G(P) \begin{bmatrix} I_n \\ -K \end{bmatrix}.$$

The following policy iteration method was originally presented in Kleinman (1968).

Algorithm 1 ***(Kleinman's Policy Iteration)***

(1) Choose a stabilizing control gain K_1, and let $i = 1$.

(2) (Policy evaluation) Evaluate the performance of control gain K_i, by solving

$$\mathcal{H}(G_i, K_i) = 0 \tag{4}$$

for $P_i \in \mathbb{S}^n$, where $G_i \triangleq G(P_i)$.

(3) (Policy improvement) Get the improved policy by

$$K_{i+1} = [G_i]_{uu}^{-1}[G_i]_{ux}. \tag{5}$$

(4) Set $i \leftarrow i+1$ and go back to Step (2).

The following properties were proved in Kleinman (1968).

(i) $A - BK_i$ is Hurwitz for all $i = 1, 2, \ldots$.
(ii) $P_1 \geq P_2 \geq P_3 \geq \cdots \geq P^*$.
(iii) $\lim_{i\to\infty} P_i = P^*$, $\lim_{i\to\infty} K_i = K^*$.

2.3 Robust policy iteration

Clearly, the implementation of Kleinman's policy iteration algorithm relies upon the exact knowledge of the system matrices A and B. In the absence of the exact knowledge of A and B, only an estimate of G in Algorithm 1 obtained from data can be used, which leads to the following inexact, yet implementable, policy iteration algorithm:

Algorithm 2 ***(Inexact Policy Iteration)***

(1) Choose a stabilizing control gain $\hat{K}_1$, and let $i = 1$.
(2) (Inexact policy evaluation) Obtain $\hat{G}_i \in \mathbb{S}^{m+n}$ as an approximation of $G(\hat{P}_i)$, where $\hat{P}_i$ is the solution of

$$\mathcal{H}(G(\hat{P}_i), \hat{K}_i) = 0. \tag{6}$$

(3) (Policy update) Construct a new control gain

$$\hat{K}_{i+1} = [\hat{G}_i]_{uu}^{-1}[\hat{G}_i]_{ux}. \tag{7}$$

(4) Set $i \leftarrow i+1$ and go back to Step (2).

With the error $\Delta G_i \triangleq \hat{G}_i - G(\hat{P}_i)$ in each iteration, the sequences $\{\hat{G}_i\}_{i=1}^{\infty}$ and $\{\hat{K}_i\}_{i=1}^{\infty}$ generated by Algorithm 2 would be different from the sequences $\{G_i\}_{i=1}^{\infty}$ and $\{K_i\}_{i=1}^{\infty}$ generated by Algorithm 1. Thus, a natural question to ask is: Is policy iteration robust to the errors in the learning process? In other words, in the presence of error ΔG_i, when will $\hat{K}_i$ still converge to a small neighborhood of K^*? In our recent work (Pang et al. 2021), we provide an answer to this question, as shown in the following theorem.

Theorem 1 *For any given stabilizing control gain $\hat{K}_1$ and any $\epsilon > 0$, there exists $\delta(\epsilon, \hat{K}_1) > 0$, such that if $Q > 0$ and $\|\Delta G\|_\infty < \delta$,*

(i) *$[\hat{G}_i]_{uu}$ is invertible, $\hat{K}_i$ is stabilizing, $\|\hat{K}_i\|_F < M_0$, $\forall i \in \mathbb{Z}_+$, $i > 0$, where $M_0(\delta) > 0$.*
(ii) *$\limsup_{i\to\infty} \|\hat{K}_i - K^*\|_F < \epsilon$.*
(iii) *$\lim_{i\to\infty} \|\Delta G_i\|_F = 0$ implies $\lim_{i\to\infty} \|\hat{K}_i - K^*\|_F = 0$.*

Intuitively, Theorem 1 implies that in Algorithm 2, if the error signal ΔG is bounded and not too large, then the generated control policy $\hat{K}_i$ is also bounded and will ultimately be in a neighborhood of the optimal policy K^* whose size is proportional to the l^∞-norm of the error signal. The smaller the error is, the better the ultimately generated policy is. In other words, the algorithm described in Algorithm 2 is not sensitive to small errors in the learning process.

2.4 Optimistic least-squares policy iteration

This subsection presents a specific method to construct the estimation $\hat{G}_i$ in Step (2) of Algorithm 2 from the input/state data generated by system (1) (sensory data generated in the sensorimotor interactions), without the knowledge of system matrices A, B, gain matrices $\{C_k\}_{k=1}^q$ and the control-dependent noise. Thus, the resulting Algorithm 3, named optimistic least-squares policy iteration (O-LSPI), is a novel model-free computational mechanism and an instantiation of Algorithm 2.

The O-LSPI is based on the following lemma.

Lemma 1 *For any stabilizing control gain K, its associated P_K satisfying (4) is the unique stable equilibrium of linear dynamical system*

$$\dot{P} = \mathcal{H}(G(P), K), \quad P(0) \in \mathbb{S}^n, \tag{8}$$

and $\lim_{t\to\infty} G(P(t)) = G(P_K)$.

Proof Vectorizing (8), we have

$$\begin{aligned}\mathrm{vec}(\dot{P}) = &\left(I_n \otimes (A - BK)^{\mathrm{T}} + (A - BK)^{\mathrm{T}} \otimes I_n\right) \\ &\times \mathrm{vec}(P) + \mathrm{vec}(Q + K^{\mathrm{T}}RK).\end{aligned} \tag{9}$$

Since $(A - BK)$ is Hurwitz, obviously this linear dynamical system admits a unique stable equilibrium P_K. □

Lemma 1 implies that in Algorithm 1, instead of solving (4), one can solve the ODE (8). This is actually the continuous-time version of the optimistic policy iteration in Tsitsiklis (2002) and Bertsekas (2011) for finite state and action spaces (thus the name "optimistic"). Lemma 1 is a well-known result in control theory (Mori et al. 1986), where (4) and (8) are in fact the algebraic Lyapunov matrix equation and the Lyapunov matrix differential equation, respectively.

Now, we show how $\hat{G}_i$ in Algorithm 2 can be estimated by CNS directly using the sensory data, i.e., input/state data

collected from system (1), based on (8) and least squares. Suppose in the i-th iteration, a control policy

$$u_i = -\hat{K}_i x + y \tag{10}$$

is applied to the system (1) to collect data, where $y \in \mathbb{R}^m$ is the exploration noise. For any $\bar{P}_i \in \mathbb{S}^n$, Ito's formula (Pavliotis 2014, Lemma 3.2) yields

$$\begin{aligned} \mathrm{d}(x^{\mathrm{T}}\bar{P}_i x) &= 2x^{\mathrm{T}}\bar{P}_i(Ax+Bu)\mathrm{d}t + u^{\mathrm{T}}\Sigma(\bar{P}_i)u\mathrm{d}t \\ &\quad + 2x^{\mathrm{T}}\bar{P}_i B\sum_{k=1}^{q} C_k u \mathrm{d}w_k, \end{aligned}$$

where $\Sigma(\bar{P}_i) = \sum_{k=1}^{q} C_k^{\mathrm{T}} B^{\mathrm{T}} \bar{P}_i B C_k$. Define $t_j = j\Delta t$, where $j = 0, 1, \ldots, M$, $\Delta t > 0$ and M is a positive integer. Integrating the above equation from t_j to t_{j+1}, we have

$$\begin{aligned} &x^{\mathrm{T}}(t_{j+1})\bar{P}_i x(t_{j+1}) - x^{\mathrm{T}}(t_j)\bar{P}_i x(t_j) \\ &= \int_{t_j}^{t_{j+1}} z^{\mathrm{T}}\theta(\bar{P}_i) z \mathrm{d}t + \int_{t_j}^{t_{j+1}} 2x^{\mathrm{T}}\bar{P}_i B\sum_{k=1}^{q} C_k u \mathrm{d}w_k, \end{aligned} \tag{11}$$

where $z = \left[x^{\mathrm{T}}, u^{\mathrm{T}}\right]^{\mathrm{T}}$ and

$$\begin{aligned} \theta(\bar{P}_i) &\triangleq \begin{bmatrix} A^{\mathrm{T}}\bar{P}_i + \bar{P}_i A & \bar{P}_i B \\ B^{\mathrm{T}}\bar{P}_i & \Sigma(\bar{P}_i) \end{bmatrix} \\ &= \left[\begin{array}{c|c} [\theta(\bar{P}_i)]_{xx} & [\theta(\bar{P}_i)]_{ux}^{\mathrm{T}} \\ \hline [\theta(\bar{P}_i)]_{ux} & [\theta(\bar{P}_i)]_{uu} \end{array}\right]. \end{aligned}$$

Taking the expectation on both sides of (11) yields

$$(X_{j+1} - X_j)^{\mathrm{T}} \operatorname{svec}(\bar{P}_i) = Z_j^{\mathrm{T}} \operatorname{svec}(\theta(\bar{P}_i)), \tag{12}$$

where for any $Y \in \mathbb{S}^m$

$$\begin{aligned} \operatorname{svec}(Y) &= [y_{11}, \sqrt{2}y_{12}, \ldots, \sqrt{2}y_{1m}, y_{22}, \sqrt{2}y_{23}, \\ &\quad \ldots, \sqrt{2}y_{m-1,m}, y_{m,m}]^{\mathrm{T}} \in \mathbb{R}^{\frac{1}{2}m(m+1)}, \\ X_j &= \mathbb{E}[\operatorname{svec}(x(t_j)x^{\mathrm{T}}(t_j))], \\ Z_j &= \mathbb{E}\left[\int_{t_j}^{t_{j+1}} \operatorname{svec}(zz^{\mathrm{T}})\mathrm{d}t\right]. \end{aligned}$$

Rewriting the above linear equations (12) for $j = 0, \ldots, M-1$ into a compact form, we obtain

$$\Phi_{i,M} \operatorname{svec}(\theta(\bar{P}_i)) = \Psi_{i,M} \operatorname{svec}(\bar{P}_i), \tag{13}$$

where

$$\begin{aligned} \Phi_{i,M} &= [Z_0, Z_1, \ldots, Z_{M-1}]^{\mathrm{T}}, \\ \Psi_{i,M} &= [X_1 - X_0, X_2 - X_1, \ldots, X_M - X_{M-1}]^{\mathrm{T}}, \end{aligned}$$

and i in the subscriptions of Φ and Ψ is used to emphasize that we are using (10) as the control policy. The following assumption is made.

Assumption 1 $\Phi_{i,M}$ has full column rank.

Remark 1 Assumption 1 is in the spirit of persistent excitation condition in adaptive control (Åström and Wittenmark 1995). Similar assumptions are needed in other RL methods, see Jiang and Jiang (2014, 2017), Kamalapurkar et al. (2018), Kiumarsi et al. (2018), Bian et al. (2016, 2020), Bian and Jiang (2019), Pang and Jiang (2020, 2021), Pang et al. (2019, 2020). Assumption 1 makes the data-based differential equation (18) a good approximation of the model-based differential equation (8), which is a key component in the convergence analysis of the O-LSPI in the sequel. The presence of exploration noise y is necessary for Assumption 1; otherwise, u will always be linearly dependent on x.

Under Assumption 1, (13) can be rewritten as

$$\operatorname{svec}(\theta(\bar{P}_i)) = \Phi_{i,M}^{\dagger}\Psi_{i,M} \operatorname{svec}(\bar{P}_i), \tag{14}$$

where $\Phi_{i,M}^{\dagger}$ denotes the Moore–Penrose pseudoinverse of $\Phi_{i,M}$. Notice that (8) can be rewritten as

$$\dot{\bar{P}}_i = \mathcal{H}(\underbrace{\theta(\bar{P}_i) - 0_n \oplus [\theta(\bar{P}_i)]_{uu} + Q \oplus R}_{=G(\bar{P}_i)}, K) \tag{15}$$

with $\bar{P}_i(0) \in \mathbb{S}^n$, where

$$\begin{aligned} 0_n \oplus [\theta(\bar{P}_i)]_{uu} &= \begin{bmatrix} 0_n & 0_{n\times m} \\ 0_{m\times n} & [\theta(\bar{P}_i)]_{uu} \end{bmatrix}, \\ Q \oplus R &= \begin{bmatrix} Q & 0_{n\times m} \\ 0_{m\times n} & R \end{bmatrix}. \end{aligned}$$

If (14) is inserted into (15), then (15) only depends on the data-based matrices $\Phi_{i,M}$ and $\Psi_{i,M}$, i.e., the precise knowledge of system matrices A, B and $\{C_k\}_{k=1}^q$ is not needed. However, the expectations in $\Phi_{i,M}$ and $\Psi_{i,M}$ are not known directly. Thus, they need to be estimated from the data. Suppose there are in total N trajectories of state and input data of length t_M that are collected along the solutions of system (1) with control policy (10). Then, we can construct approximation $\hat{\theta}(\bar{P}_i)$ of $\theta(\bar{P}_i)$ using

$$\operatorname{svec}(\hat{\theta}(\bar{P}_i)) = \hat{\Phi}_{i,M,N}^{\dagger}\hat{\Psi}_{i,M,N} \operatorname{svec}(\bar{P}_i), \tag{16}$$

where

$$\hat{\Phi}_{i,M,N} = [\hat{Z}_{0,N}, \hat{Z}_{1,N}, \ldots, \hat{Z}_{M-1,N}]^{\mathrm{T}},$$
$$\hat{\Psi}_{i,M,N} = [\hat{X}_{1,N} - \hat{X}_{0,N}, \hat{X}_{2,N} - \hat{X}_{1,N}, \ldots, \hat{X}_{M,N} - \hat{X}_{M-1,N}]^{\mathrm{T}},$$
$$\hat{Z}_{j,N} = \frac{1}{N}\sum_{l=1}^{N}\int_{t_j}^{t_{j+1}} \mathrm{svec}\left(z^{(l)}[z^{(l)}]^{\mathrm{T}}\right)\mathrm{d}t,\ z^{(l)} = \begin{bmatrix} x^{(l)} \\ u^{(l)} \end{bmatrix},$$
$$\hat{X}_{j,N} = \frac{1}{N}\sum_{l=1}^{N}\mathrm{svec}\left(x_{t_j}^{(l)}[x_{t_j}^{(l)}]^{\mathrm{T}}\right), \quad j = 0, \ldots, M-1.$$

and $x^{(l)}$, $u^{(l)}$ are the l-th state trajectory and input trajectory, respectively. By the strong law of large numbers, almost surely

$$\lim_{N\to\infty}\hat{\Phi}_{i,M,N} = \Phi_{i,M}, \quad \lim_{N\to\infty}\hat{\Psi}_{i,M,N} = \Psi_{i,M}. \tag{17}$$

This implies that the solution of the following ordinary differential equation

$$\dot{\check{P}}_i = \mathcal{H}(\hat{\theta}(\check{P}_i) - 0_n \oplus [\hat{\theta}(\check{P}_i)]_{uu} + Q \oplus R, K), \tag{18}$$

with $\check{P}_i(0) \in \mathbb{S}^n$ would be close to the solution of (15), if N is large enough and $\bar{P}(0) = \check{P}(0)$. Thus by Lemma 1, $\check{P}_i(s)$ and $\hat{\theta}(\check{P}_i(s))$ would be close to $\hat{P}_i$ and $\theta(\hat{P}_i)$ in Algorithm 2, respectively, for N and $s > 0$ large enough. In view of the relationship between $\theta(\cdot)$ and $G(\cdot)$ in (15), an estimation $\hat{G}_i$ of $G(\hat{P}_i)$ can be constructed as

$$\hat{G}_i = \hat{\theta}(\check{P}_i(s)) - 0_n \oplus [\hat{\theta}(\check{P}_i(s))]_{uu} + Q \oplus R \tag{19}$$

in Algorithm 2. The O-LSPI algorithm is summarized in Algorithm 3.

Algorithm 3 ***(Optimistic Least-Squares Policy Iteration)***

(1) (Initialization) Choose a stabilizing control gain $\hat{K}_1$, number of trajectories N, time step $\Delta t > 0$, length of samples M, length of policy evaluation $s > 0$, number of iterations $\bar{I}$. Let $i = 1$.

(2) Collect input/state data to construct data matrices $\hat{\Phi}_{i,M,N}$ and $\hat{\Psi}_{i,M,N}$ in (16), by applying control policy (10) to (1).

(3) (Inexact policy evaluation) Obtain $\hat{G}_i$ defined in (19) by solving the initial value problem of equation (18) on $[0, s]$ with initial value $\check{P}_i(0) = 0_n$.

(4) (Policy update) Construct a new control gain

$$\hat{K}_{i+1} = [\hat{G}_i]_{uu}^{-1}[\hat{G}_i]_{ux}. \tag{20}$$

(5) Set $i \leftarrow i + 1$. If $i < \bar{I}$, go back to Step (2).

(6) Use $u_{\bar{I}} = -\hat{K}_{\bar{I}}x$ as the approximately optimal control policy.

It is worth emphasizing again that Algorithm 3 does not need the knowledge of system matrices A and B, gain matrices $\{C_k\}_{k=1}^q$ and the control-dependent noise w_k. Based on Theorem 1, the convergence of Algorithm 3 is given in the following theorem, whose proof is given in "Appendix 1."

Theorem 2 *For any given stabilizing control gain $\hat{K}_1$ and $\epsilon_1 > 0$, there exist an integer N_0, an integer $\bar{I}$ and a constant $s_0 > 0$, such that if Assumption 1 is satisfied for all $i = 1, \ldots, \bar{I}$, then for any $N > N_0$ and $s > s_0$, almost surely*

$$\|\hat{K}_{\bar{I}} - K^*\|_F < \epsilon_1,$$

and $\hat{K}_i$ is stabilizing for all $i = 1, \ldots, \bar{I}$.

2.5 An example: double integrator

To verify the effectiveness of the O-LSPI algorithm and its convergence result Theorem 2, consider a double integrator disturbed by control-dependent stochastic noise,

$$\mathrm{d}x = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}x\mathrm{d}t + \begin{bmatrix} 0 \\ 1 \end{bmatrix}(u\mathrm{d}t + 0.1u\mathrm{d}w_1). \tag{21}$$

It is assumed that the system parameters in (21) are unknown and the stochastic noise w_1 is unmeasurable, but an initial stabilizing control gain $\hat{K}_1 = [11, 9]$ is available. Setting $Q = I_2$ and $R = 1$, we run O-LSPI with parameters $N = 10^4$, $\Delta t = 0.05$, $M = 7$, $s = 10$, $\bar{I} = 10$, and exploration noise

$$y = \sum_{j=1}^{100}\sin(\eta_j t),$$

where $\{\eta_j\}_{j=1}^{100}$ are drawn independently and identically from the Gaussian distribution with mean -250 and standard deviation 500. The simulation results are shown in Fig. 1, where the relative errors of $\hat{K}_i$ and $\hat{P}_i$ with respective to their optimal values K^* and P^* are plotted, respectively. One can see that the relative errors converge to a small neighborhood of zero, which implies that $\hat{K}_i$ and $\hat{P}_i$ converge to a small neighborhood of their optimal value, respectively.

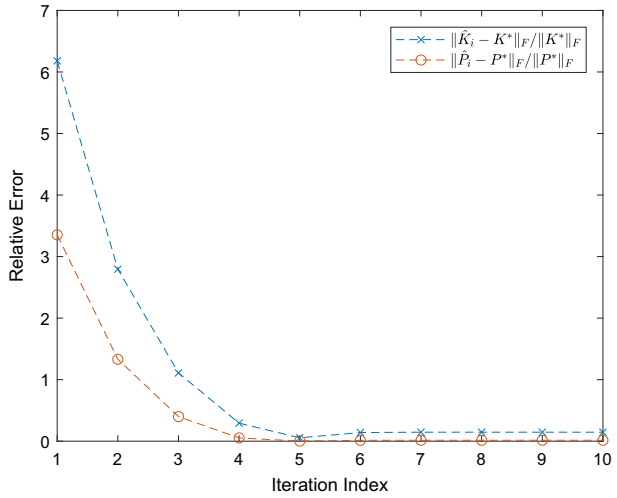

Fig. 1 Simulation results on the double integrator

3 Numerical studies in motor learning and control

3.1 Human arm movement model

We consider the sensorimotor control tasks studied in Harris and Wolpert (1998), Burdet et al. (2001, 2006) and Franklin et al. (2003), where human subjects make point-to-point reaching movements in the horizontal plane. The objective is to reproduce similar results to those observed from experiments in Burdet et al. (2001, 2006) and Franklin et al. (2003), using the proposed O-LSPI algorithm.

In our numerical experiment, the dynamics of the arm are simplified to a point-mass model (Liu and Todorov 2007) as shown below,

$$\begin{aligned} \dot{p} &= v, \\ m\dot{v} &= a - bv + f, \\ \tau\dot{a} &= u - a + C_1 u\xi_1 + C_2 u\xi_2, \end{aligned} \tag{22}$$

where $p = [p_x, p_y]^{\mathrm{T}}$, $v = [v_x, v_y]^{\mathrm{T}}$, $a = [a_x, a_y]^{\mathrm{T}}$, $u = [u_x, u_y]^{\mathrm{T}}$, $f = [f_x, f_y]^{\mathrm{T}}$ are the two-dimensional hand position, velocity, actuator state, control input and external force generated from the force fields, respectively; m denotes the mass of the hand; b is the viscosity constant; τ is the time constant; ξ_1 and ξ_2 are Gaussian white noises (Pavliotis 2014); and

$$C_1 = \begin{bmatrix} c_1 & 0 \\ c_2 & 0 \end{bmatrix}, \qquad C_2 = \begin{bmatrix} 0 & c_2 \\ 0 & c_1 \end{bmatrix}$$

are gain matrices of the control-dependent noise (Harris and Wolpert 1998). To fit this model into the optimal control problem formulated in Sect. 2.1, (22) is rewritten in the form of state-space model,

$$\mathrm{d}x = (Ax + Bu)\mathrm{d}t + B(C_1 u\mathrm{d}w_1 + C_2 u\mathrm{d}w_2) + Df\mathrm{d}t \tag{23}$$

where w_1 and w_2 are one-dimensional standard Brownian motions, and

$$x = \begin{bmatrix} p \\ v \\ a \end{bmatrix}, \quad A = \begin{bmatrix} 0_2 & I_2 & 0_2 \\ 0_2 & -\frac{b}{m} I_2 & \frac{1}{m} I_2 \\ 0_2 & 0_2 & -\frac{1}{\tau} I_2 \end{bmatrix},$$

$$B = \begin{bmatrix} 0_2 \\ 0_2 \\ \frac{1}{\tau} I_2 \end{bmatrix}, \quad D = \begin{bmatrix} 0_2 \\ I_2 \\ 0_2 \end{bmatrix}.$$

The weighting matrices in cost function (2) are chosen as

$$Q = \begin{bmatrix} 2000 & -40 & 0 & 0 & 0 & 0 \\ -40 & 1000 & 0 & 0 & 0 & 0 \\ 0 & 0 & 20 & -1 & 0 & 0 \\ 0 & 0 & -1 & 20 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.01 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.01 \end{bmatrix},$$

$$R = 0.01 I_2.$$

The term f in (23) is used to model possible external disturbances exerted on the hand (Liu and Todorov 2007) from the force fields. Three kinds of disturbances are considered here: the null field (NF) (Burdet et al. 2001, 2006; Franklin et al. 2003), $f \equiv 0$, where no external disturbances are exerted; the velocity-dependent force field (VF) (Burdet et al. 2006; Franklin et al. 2003),

$$f = \chi \begin{bmatrix} 13 & -18 \\ 18 & 13 \end{bmatrix} \begin{bmatrix} v_x \\ v_y \end{bmatrix}$$

where $\chi \in [2/3, 1]$ is a constant that can be adjusted to the subject's strength; the divergent force field (DF) (Burdet et al. 2001, 2006; Franklin et al. 2003),

$$f = \begin{bmatrix} \beta & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} p_x \\ 0 \end{bmatrix} \tag{24}$$

where $\beta > 0$ and a negative elastic force disturbance perpendicular to the target direction is produced.

Table 1 Parameters of the arm movement model

Parameters	Description	Value
m	Hand mass	1.3 kg
b	Viscosity constant	10 N s/m
τ	Time constant	0.05 s
$c1$	Noise magnitude	0.075
$c2$	Noise magnitude	0.025

The model parameters used in our simulations are given in Table 1.

3.2 Sensorimotor control in velocity-dependent force field

In the experiments conducted by Franklin et al. (2003) and Burdet et al. (2006), subjects sat in a chair and moved the parallel-link direct drive air-magnet floating manipulandum (PFM) in a series of forward reaching movements performed in the horizontal plane. Subjects performed reaching movements from a start circle to a target circle with total distance 0.25 m. They firstly practiced in the NF until they achieved enough successful trials. Trials were considered successful if they ended inside the target within the prescribed time 0.6 ± 0.1 s. Then, VF was activated without informing the subjects in advance. Subjects practiced in VF until achieving enough successful trials. Next they took a short break and performed several trials in the NF. These trials were called after-effects and were recorded to confirm that subjects adapted to the force field. It was observed (Franklin et al. 2003; Burdet et al. 2006) that initial trials of the subjects in VF were distorted drastically, but subjects made straighter movements gradually. After learning through enough trials, the trajectories were relatively straight and consistently reached the final target position. Inspection of the stiffness, which was defined as graphical depiction of the elastic restoring force corresponding to the unit displacement of the hand for the subject in the force fields (Burdet et al. 2001, 2006), revealed that after adaptation endpoint stiffness was selectively modified to the direction of the instability. See Gomi and Kawato (1996) and Franklin et al. (2003) for more details.

In this subsection, we apply O-LSPI to the human arm movement model (23) to reproduce the experimental results in Franklin et al. (2003) and Burdet et al. (2006).

The experiments in NF are firstly simulated. The O-LSPI starts with an initial stabilizing control gain $\hat{K}_1 \in \mathbb{R}^{2\times 6}$, such that $A - B\hat{K}_1$ is Hurwitz. Such a control gain can be found by robust control theory (Zhou and Doyle 1998), if some upper and lower bounds of the elements in A and B are available and the pair (A, B) is stabilizable. Indeed, the first several trials in the NF can be interpreted as the searching for an initial stabilizing control gain, by estimating the bounds of the parameters b, m and τ and using robust control techniques. If the CNS figures out that $b \in [-8, 12]$, $m \in [1, 1.5]$ and $\tau \in [0.03, 0.07]$ in the first several trials, an initial control gain can be chosen as

$$\hat{K}_1 = \begin{bmatrix} 100 & 0 & 10 & 0 & 10 & 0 \\ 0 & 100 & 0 & 10 & 0 & 10 \end{bmatrix}.$$

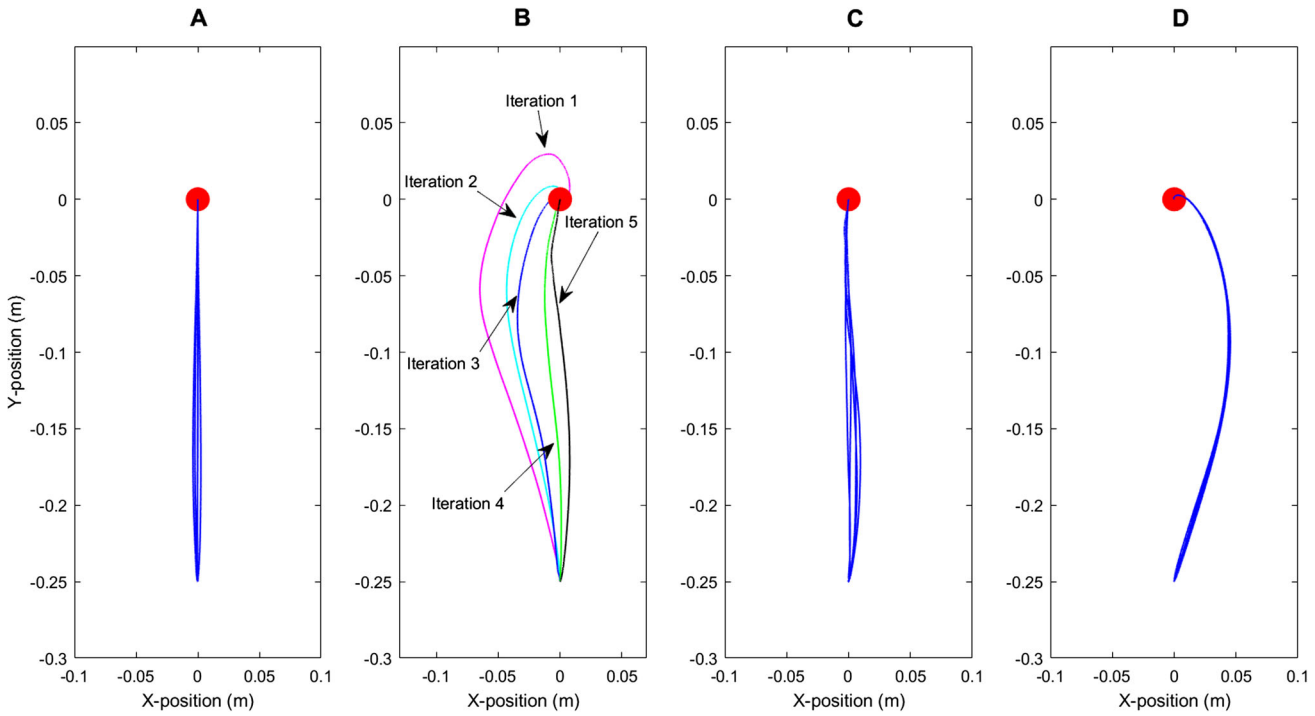

Fig. 2 Simulated movement trajectory generated by O-LSPI. **a** Five movement trajectories of the subject after learning in the NF. **b** The first five consecutive movement trajectories of the subject in the VF. **c** Five consecutive movement trajectories of the subject after 30 trials in the VF. **d** Five after-effect trials in the NF

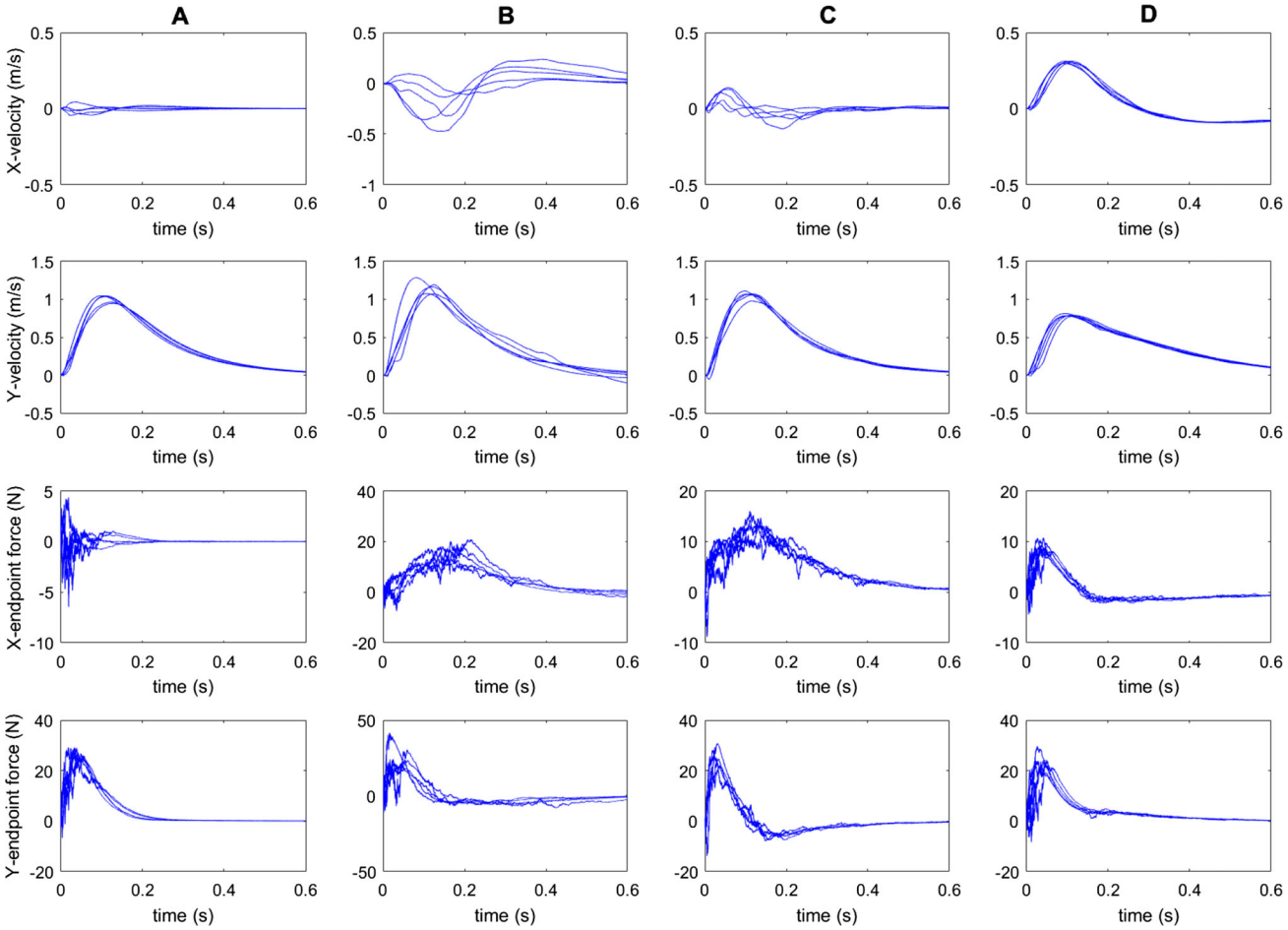

Fig. 3 Simulated velocity and endpoint force curves generated by O-LSPI. **a** Simulated curves of the subject after learning in the NF. **b** Simulated curves of the subject in several trials when firstly exposed to the VF. **c** Simulated curves of the subject after 30 trials in the VF. **d** After-effect trials in the NF. x-velocity and y-velocity curves are shown in the first and second rows, respectively. Bell-shaped velocity curves are clearly observed in y-velocity curves. x-endpoint force and y-endpoint force curves are shown in the third and fourth rows, respectively. A comparison of the first and third figures in the x-endpoint force suggests that subjects adapted to the VF by generating compensation force to counteract the force produced by VF. The shapes of these curves resemble closely the experimental results reported in Franklin et al. (2003); Burdet et al. (2006)

Then during the i-th trial, we collect input/state data generated by control policy (10) and construct estimation $\hat{G}_i$ from the data to update the control policy.[1] After 30 trials, the control policy is updated to

$$\hat{K}_{30} = \begin{bmatrix} 446.21 & -6.98 & 49.80 & -0.94 & 1.41 & -0.02 \\ -5.46 & 315.00 & -0.84 & 43.38 & -0.01 & 1.31 \end{bmatrix} \tag{25}$$

which is nearly optimal since the optimal policy in NF is

$$\hat{K}^*_{\mathrm{NF}} = \begin{bmatrix} 447.19 & -5.91 & 49.70 & -0.94 & 1.41 & -0.01 \\ -4.27 & 316.17 & -0.91 & 43.34 & -0.01 & 1.30 \end{bmatrix}.$$

Next, the experiments in VF are simulated. Since the velocity force field is activated without notifying the subjects, the first trial in VF was under the same control policy as learnt in NF. After the first trial, the CNS can realize that it is facing with a new environment different from NF. So O-LSPI is applied starting from the second trial. The initial

[1] To construct good enough data matrices $\hat{\Phi}_{i,M,N}$ and $\hat{\Psi}_{i,M,N}$ in Step (2) of Algorithm 3, stochastic differential equation (23) needs to be solved over a long time interval, which is time-consuming on an ordinary laptop. We avoid this difficulty by directly computing $\Phi_{i,M}$ and $\Psi_{i,M}$ based on (23) and setting $\hat{\Phi}_{i,M,N} = \Phi_{i,M} + \omega_1$ and $\hat{\Psi}_{i,M,N} = \Psi_{i,M} + \omega_2$, where each element in ω_1 and ω_2 are drawn from the uniform distribution. All the simulations in this section are conducted in this way. Elements in ω_1 and ω_2 are drawn from uniform distribution over $[-10, 10]$ in NF and VF, and over $[-1, 1]$ in DF, respectively.

control gain we use for the second trial is obtained by tripling the gains in (25) in the NF. This is to mimic the experimental observation (Franklin et al. 2008) that muscle activities increased dramatically after the first trial. After 30 trials, the control policy is updated to

$$\hat{K}_{30} = \begin{bmatrix} 426.00 & 100.54 & 66.30 & -12.56 & 1.65 & -0.04 \\ -155.75 & 299.92 & 5.93 & 61.57 & -0.04 & 1.59 \end{bmatrix}$$

which is nearly optimal since the optimal policy in NF is

$$\hat{K}^*_{\mathrm{VF}} = \begin{bmatrix} 419.42 & 101.30 & 65.99 & -12.37 & 1.65 & -0.04 \\ -155.19 & 299.56 & 6.05 & 61.80 & -0.04 & 1.59 \end{bmatrix}.$$

The simulated movement trajectories, the velocity curves and the endpoint force curves are shown in Figs. 2 and 3. It can be seen that the simulated movement trajectories in the NF are approximately straight lines, and the velocity curves along the y-axis are bell-shaped. The subject successfully reaches the target in the first trial in VF, but the trajectory is heavily distorted to the upper-left side, since the subject is still using the near-optimal control policy learnt in NF when firstly exposed to VF. Although VF produces a stable interaction with the arm, the near-optimal control policy learnt in NF is far from being optimal in the new environment VF. The O-LSPI takes effect from the second trial in VF. Motor adaptation can be observed by comparing the first five consecutive trials, where the movement trajectories are getting straighter and straighter. After 30 trials, the movement trajectories return to be straight lines, and the velocity curves become bell-shaped again. This implies that after 30 trials in the VF, the CNS can learn well the optimal control policy using data, without knowing or using the precise system parameters, and the unmeasurable control-dependent noise. The stiffness ellipses after 30 trials are shown in Fig. 4. The stiffness in the VF increases significantly in the direction of the external force, compared with the stiffness in the NF. Finally, our numerical study shows clearly the after-effects of the subject behavior when the VF is suddenly deactivated. These observations are a clear testament that motor adaptation to VF indeed occurs. One can find through comparison that our simulation results in Figs. 2, 3 and 4 are consistent with the experimental observations in Franklin et al. (2003) and Burdet et al. (2006).

The relative estimation errors of $\hat{G}_i$ in Algorithm 3 are shown in Fig. 5. One can see that the errors are as large as 16% in the learning process. This implies that the CNS is able to adapt to VF with imperfect or noisy information.

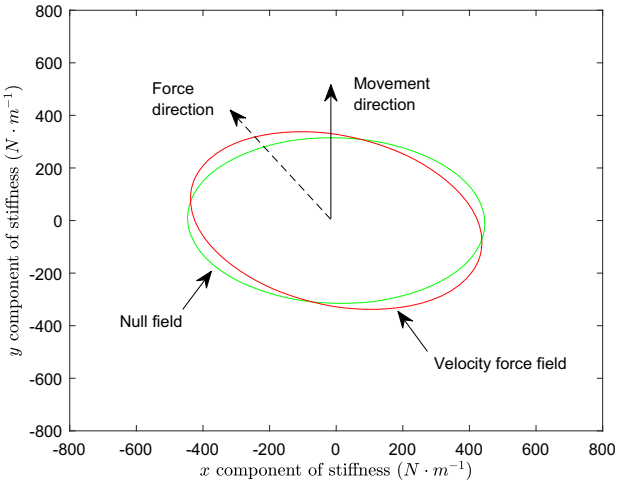

Fig. 4 Illustration of the stiffness geometry to the VF. Compared with the stiffness in the NF (green), the stiffness in the VF (red) increased significantly in the direction of the external force

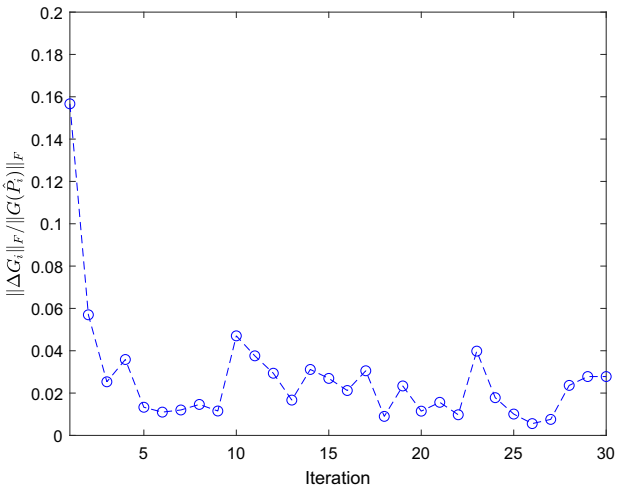

Fig. 5 Relative estimation error ΔG_i between $\hat{G}_i$ and its true value $G(\hat{P}_i)$ in O-LSPI in the VF. Although the estimation errors exist in the learning process and can be as large as 16% of the true value measured by the Frobenius norm, the subject is still able to adapt to the VF

3.3 Sensorimotor control in divergent force field

The effects of the divergent field to the subjects in the arm reaching movement are also studied in Burdet et al. (2001, 2006) and Franklin et al. (2003), whose experimental results are reproduced using our proposed computational method in this subsection. We set $\beta = 300$ in (24).

The simulation of movements in the NF before the DF is applied is the same with that presented in the last subsection. However, with $\beta = 300$ the DF produces an unstable interaction with the arm, so that the near-optimal policy learned in the NF is not stabilizing anymore. Therefore, when we apply the same near-optimal control policy learnt in the NF to generate the movements for the first five trials in the DF, unstable

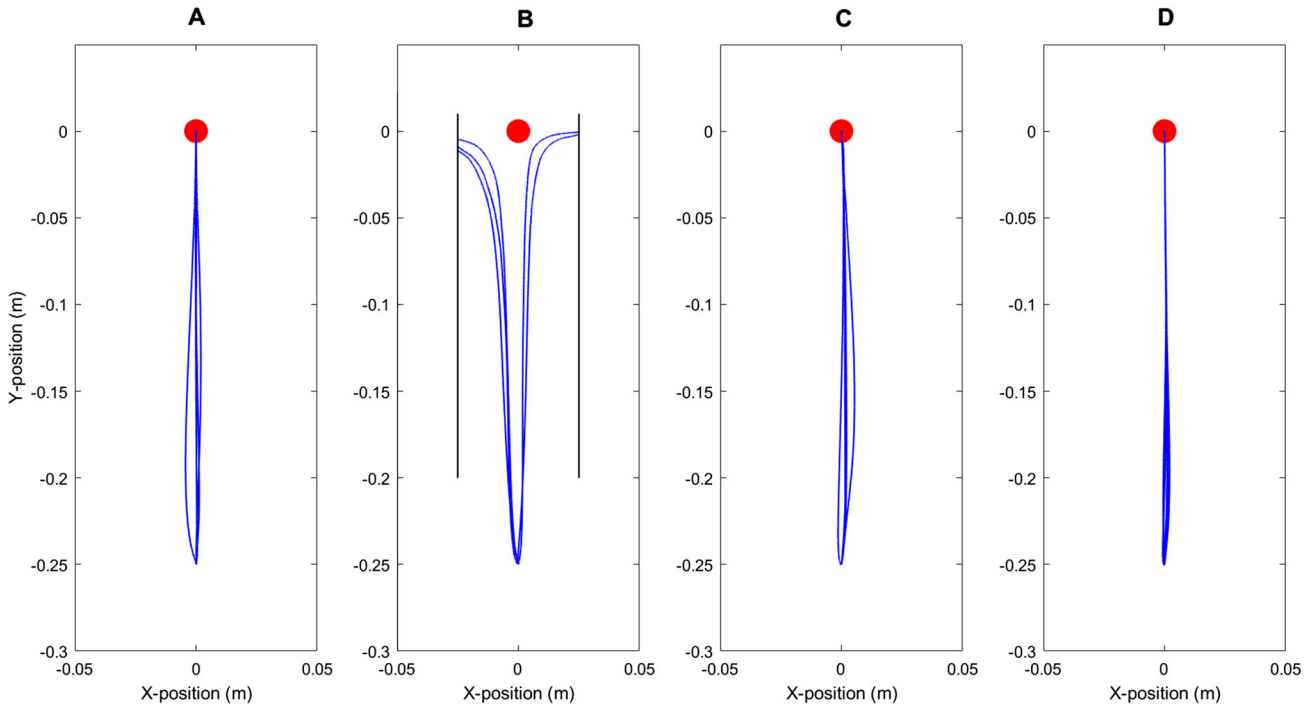

Fig. 6 Simulated movement trajectory generated by O-LSPI. **a** Five movement trajectories of the subject after learning in the NF. **b** Five independent movement trajectories of the subject when firstly exposed to DF. For safety reasons, the DF is turned off when the trajectory deviates more than 2.5 cm from the y-axis. The black lines indicate this safety zone. **c** Five consecutive movement trajectories of the subject after 30 trials in the DF. **d** Five after-effect trials in the NF

behaviors are observed, as shown in Fig. 6b. In this case, it is hypothesized that the CNS re-learns a new initial stabilizing controller using the robust control theory (see Remark 2 for an example), such that

$$A - B\hat{K}_1 = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ \frac{\beta}{m} & 0 & -\frac{b}{m} & 0 & \frac{1}{m} & 0 \\ 0 & 0 & 0 & -\frac{b}{m} & 0 & \frac{1}{m} \\ 0 & 0 & 0 & 0 & -\frac{1}{\tau} & 0 \\ 0 & 0 & 0 & 0 & 0 & -\frac{1}{\tau} \end{bmatrix} - \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ \frac{1}{\tau} & 0 \\ 0 & \frac{1}{\tau} \end{bmatrix} \hat{K}_1$$

is Hurwitz. Here, we increase the first entry in the first row of the control gain in (25) by 600 and set the resultant stabilizing control gain to be the initial stabilizing control gain. After 30 trials in the DF, the O-LSPI has updated the control gain to

$$\hat{K}_{30} = \begin{bmatrix} 1462.15 & 4.70 & 81.59 & -0.14 & 1.87 & 0.01 \\ -7.06 & 310.12 & -0.88 & 43.09 & 0.00 & 1.31 \end{bmatrix},$$

which is near-optimal, since the corresponding optimal control gain is

$$K^*_{\mathrm{DF}} = \begin{bmatrix} 1481.89 & -4.86 & 82.19 & -0.76 & 1.88 & -0.01 \\ -6.65 & 316.19 & -0.799 & 43.34 & -0.01 & 1.30 \end{bmatrix}.$$

The simulated movement trajectories, the velocity curves and the endpoint force curves are shown in Figs. 6 and 7. It can be seen that the simulated movement trajectories in the NF are approximately straight lines, and the velocity curves along the y-axis are bell-shaped. Due to the control-dependent noise, the movement trajectories differ slightly from trial to trial. Since DF produces an unstable interactions with the arm, unstable behaviors are observed in the first several trials when the subject is first exposed to the DF. Then, O-LSPI is applied, and after 30 trials, the movement trajectories become approximately straight as in the NF, which implies that the CNS has learned to adapt to the DF. The stiffness ellipses after 30 trials are shown in Fig. 8. It is clear that the stiffness in the DF increases significantly in the direction of the divergent force, and the change of stiffness along the movement direction is not significant, compared with the stiffness in the NF. Finally, the after-effects of the subject behavior are simulated when the DF is suddenly deactivated. The after-effects movement trajectories are even straighter than the trajectories in the NF. The reason is that the CNS has learned to compensate more to the displacement along the x-axis. One can find through comparison that our simulation results in Figs. 6, 7 and 8 are consistent with the experimental observations in Burdet et al. (2001, 2006) and Franklin et al. (2003).

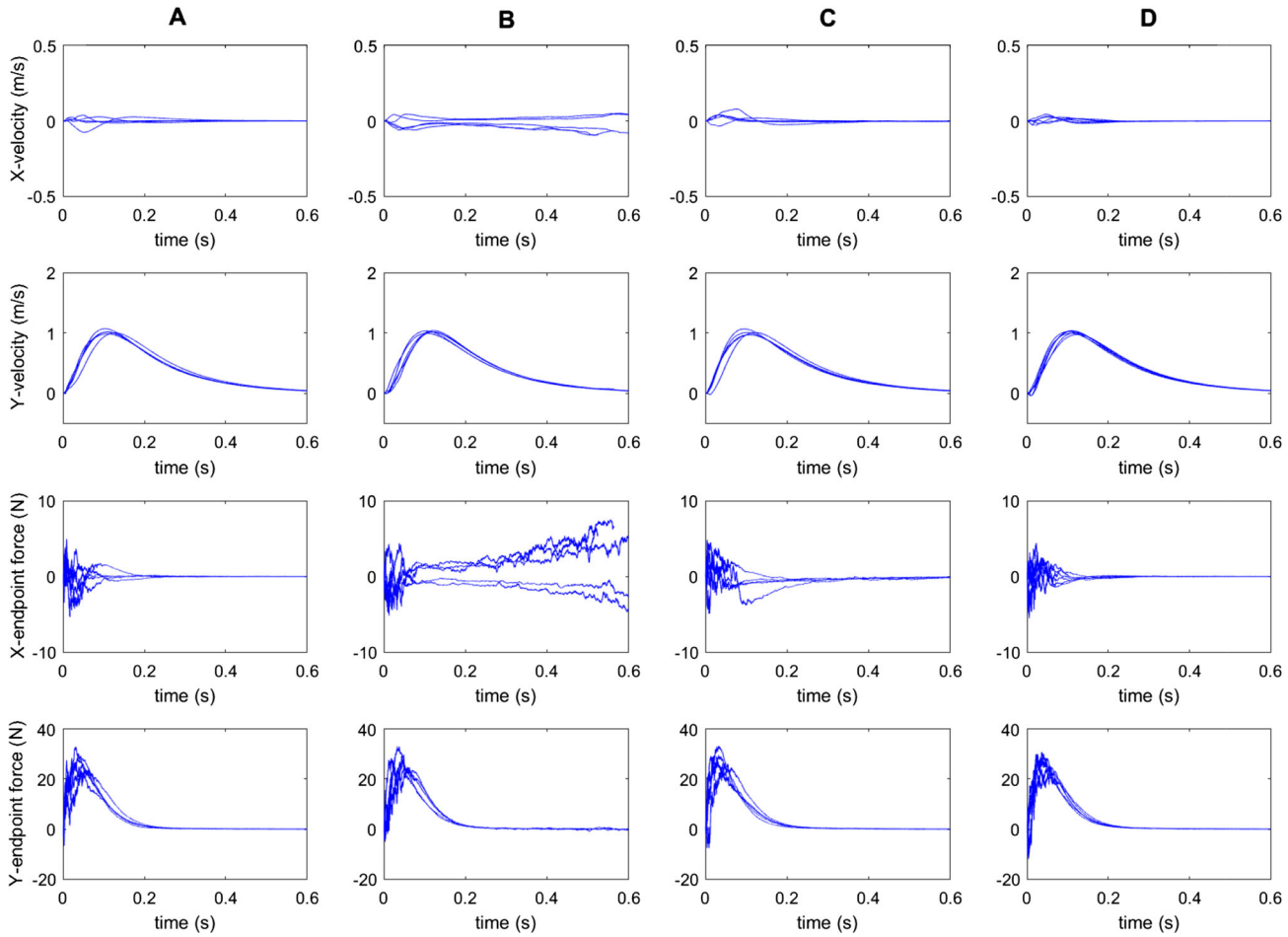

Fig. 7 Simulated velocity and endpoint force curves generated by O-LSPI. **a** Simulated curves of the subject after learning in the NF. **b** Simulated curves of the subject in several trials when firstly exposed to the DF. **c** Simulated curves of the subject after 30 trials in the DF. **d** After-effect trials in the NF. x-velocity and y-velocity curves are shown in the first and second rows, respectively. Bell-shaped velocity curves are clearly observed in y-velocity curves. x-endpoint force and y-endpoint force curves are shown in the third and fourth rows, respectively. The third figure in the x-endpoint force suggests that subjects adapted to the DF by generating compensation force in the x-direction to counteract the force produced by DF. The shapes of these curves resemble closely the experimental results reported in Burdet et al. (2001, 2006); Franklin et al. (2003)

The relative estimation errors of $\hat{G}_i$ in Algorithm 3 are shown in Fig. 9. The errors can be as large as 90% of the true values in the learning process. This implies that the CNS is able to adapt to DF in the presence of the large imperfect or noisy information.

Remark 2 By robust control theory (Zhou and Doyle 1998), an initial stabilizing controller can be found provided that the bounds of the parameters of a linear system are known. For example, suppose the CNS has access to the bounds of the unknown system parameters:

$$\begin{aligned} &0.5 \le m \le 5, \quad 2 \le b \le 20, \\ &0.02 \le \tau \le 0.5, \quad 200 \le \beta \le 500, \end{aligned} \tag{26}$$

in the DF. Then, with these bounds at hand, a direct application of robust control theory (an application of the Robust Control Toolbox (Balas et al. 2007) in terms of the code implementation) yields:

$$K = \begin{bmatrix} 15854 & -17.23 & 477.10 & -7 & 29.76 & -270.88 \\ -11.65 & 3.89 & -1.43 & 4.55 & -0.04 & 58.57 \end{bmatrix}$$

that stabilizes the hand in DF for all possible parameters satisfying the conditions in (26). In this way, an initial stabilizing (but generally not optimal) control gain required by the O-LSPI algorithm can be found, without the exact knowledge of the system dynamics.

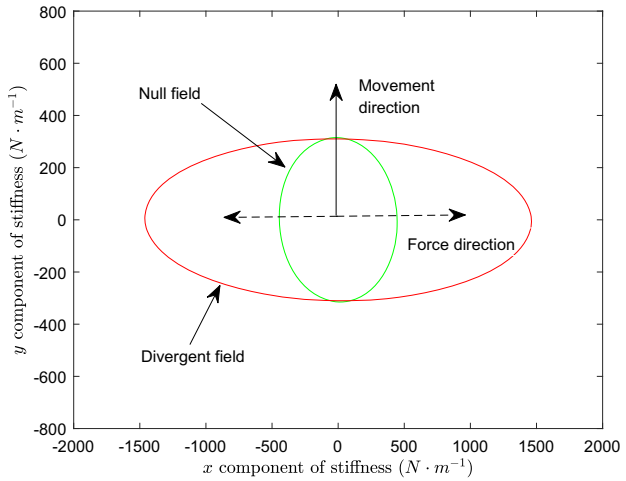

Fig. 8 Illustration of the stiffness geometry to the DF. Compared with the stiffness in the NF (green), the stiffness in the DF (red) increased significantly in the direction of the external force

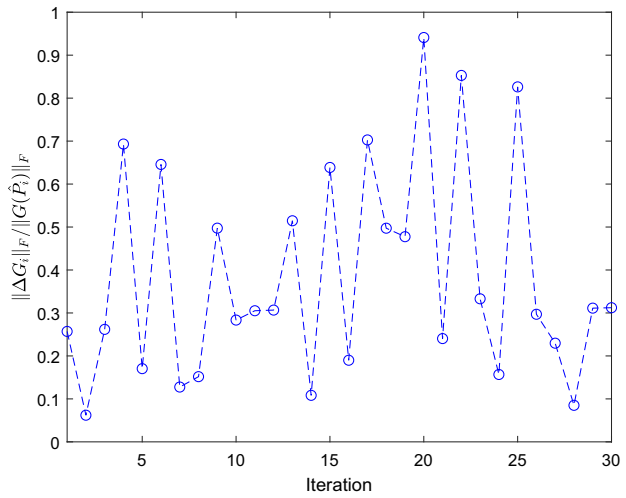

Fig. 9 Relative estimation error ΔG_i between $\hat{G}_i$ and its true value $G(\hat{P}_i)$ in O-LSPI in the DF. Although the estimation errors exist in the learning process and can be as large as 90% of the true value measured by the Frobenius norm, the subject is still able to adapt to the DF

Remark 3 It should be mentioned that if the bounds of the parameters are completely unknown, the value iteration, an alternative reinforcement learning algorithm, can be used to find the optimal LQR control policies and reproduce the experimental results in the divergent field (Bian et al. 2020), without any initial stabilizing control policy. Please see Bian and Jiang (2019), Pang and Jiang (2020) and Bian et al. (2020) for details.

3.4 Fitts's law

In this subsection, we further validate our computational model using the Fitts's law (Fitts 1954; Schmidt et al. 2018). Fitts's law is one of the widespread formal rules in the study of human behavior. The law dictates that the movement duration t_f required to rapidly reach a target area is a function of the distance d to the target and the size of the target γ. There are multiple versions of Fitts's law (Schmidt et al. 2018), two of which are used in our validation. The first one is the logarithmic law (log law)

$$t_f = \alpha_0 + \alpha_1 \log_2 \left(\frac{2d}{\gamma} \right),$$

where α_0 and α_1 are two constants. The second one is the power law

$$t_f = \alpha_0 \left(\frac{d}{\gamma} \right)^{\alpha_1}.$$

In our case, γ is the diameter of the target circle, and d is the distance from the starting point to the center of the target. We generate trials using the after-learning control polices in the NF, VF and DF, respectively. The movement duration t_f is defined as the time when the hand cursor enters the target. The data fitting results are shown in Fig. 10 and Table 2. It can be seen that our simulation results are consistent with the predictions of Fitts's law.

4 Discussion

4.1 Model-free learning

Most of the previous computational models for sensorimotor control assume that the CNS has the exact knowledge of the motor system and the environment that it is interacting with Wolpert et al. (1995), Todorov (2005), Liu and Todorov (2007), Todorov and Jordan (2002), Harris and Wolpert (1998), Mussa-Ivaldi et al. (1985), Yeo et al. (2016), Česonis and Franklin (2020), Zhou et al. (2017), Mistry et al. (2013), Ueyama (2014), Cluff and Scott (2015) and Gaveau et al. (2014). Then, the optimal control policies are computed based on this assumption. In contrast, our proposed computational model Algorithm 3 is a model-free approach which does not need accurate model information or estimate the unknown model parameters. Algorithm 3 informs that near-optimal policies are derived using the sensory data and are robust to the control-dependent noise. The numerical experiments in the last section show that out proposed model can generate typical movement trajectories, stiffness ellipses observed in previous experiments in different settings. As one of the key differences with the sensorimotor models mentioned above, our proposed computational mechanism suggests that, when confronted with unknown environments and imprecise dynamics, the CNS may update and improve

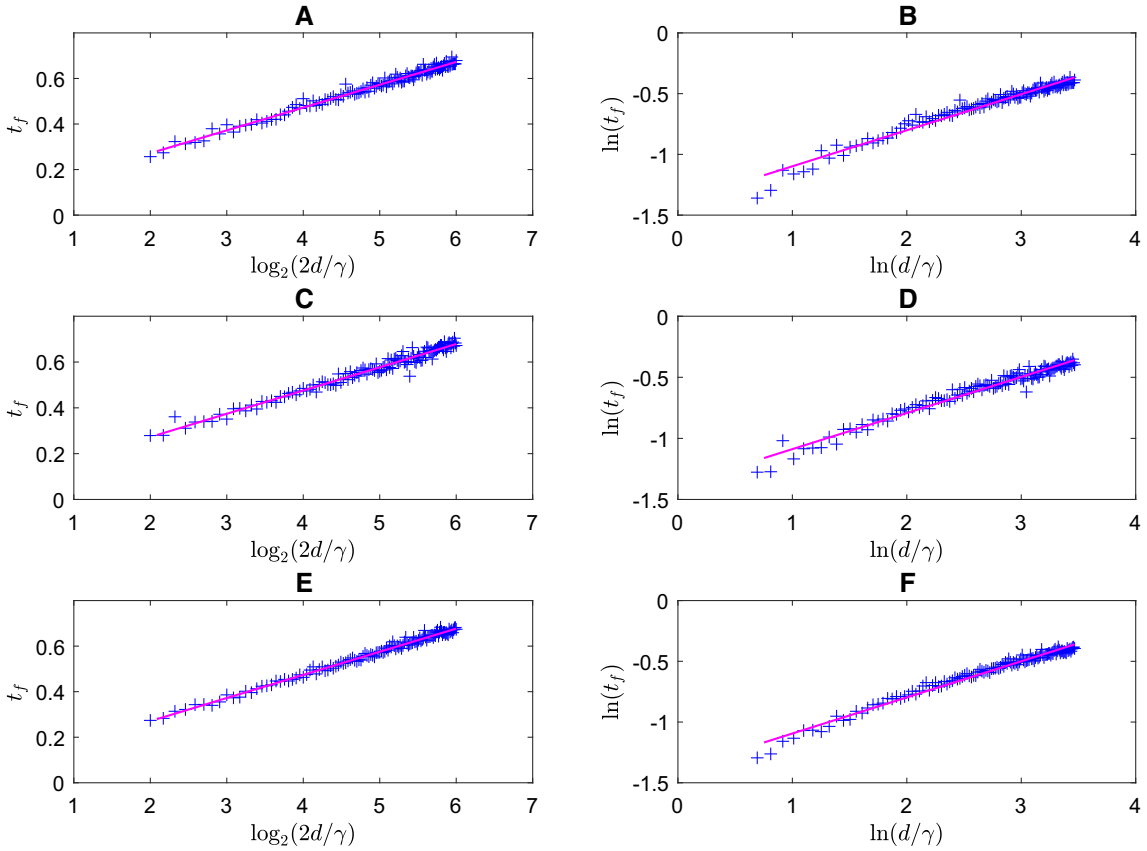

Fig. 10 Log and power versions of Fitts's law. Crosses in the first row, the second row and the third row represent after-leaning movement durations simulated in the NF, the VF and the DF, respectively. Solid lines in A, C, E are least-squares fits using the log Fitts's law, and solid lines in B, D, F are the least-squares fits using the power Fitts's law

Table 2 Parameters in the log law and power law estimated by least squares

Parameters	NF	VF	DF
α_0 (Log law)	0.0715	0.0696	0.0676
α_1 (Log law)	0.1002	0.1015	0.1015
α_0 (Power law)	− 1.3950	− 1.3833	− 1.3919
α_1 (Power law)	0.2969	0.2954	0.2975

its command signals for movement through learning and repeated trials using sensory data.

Two model-based computational mechanisms that do not require an accurate internal model are proposed in Mistry et al. (2013) and Crevecoeur et al. (2020). Assuming that the parameters in the uncertain force fields are unknown, modified and adaptive linear quadratic Gaussian control is proposed in Mistry et al. (2013) and Crevecoeur et al. (2020), respectively, to explain the experimental phenomena observed there. However, the accurate models for the arms and the rest of the environments are still needed to be known. Furthermore, the parameters in the uncertain force fields need to be estimated explicitly. By contrast, our proposed mechanism assumes that all the parameters in the model are unknown, and the control policies are directly generated from the sensory data, and no parameter in the arm and environment models is explicitly estimated.

Model-free approaches based on the optimal control framework are also developed in Jiang and Jiang (2014, 2015), Bian et al. (2016, 2020) for sensorimotor control. Although these model-free approaches successfully reproduce the experimental observations in arm reaching task, they assume implicitly that the control-dependent noises $\{w_k\}_{k=1}^{\infty}$ in (23) are measurable and explicitly use the noise-corrupted data in the computation of iterative estimates of the optimal policy. In this paper, we conjecture that the CNS makes decisions without direct access to any noise-dependent data, so the name of model-free is more relevant to the proposed computational model. Interestingly, both theoretical and numerical studies have shown that, without cancelling exactly the noise-dependent terms, human motor learning and control is inherently robust to the small noise occurring in the learning process.

Model-free approaches not based on the optimal control framework are proposed in Zhou et al. (2011) and Hadjiosif et al. (2021). The model in Zhou et al. (2011) synthesizes iterative learning control and model reference adaptive control to reproduce the behavior observed in the human motor adaptation, so that the motion command is carried out without the need for inverse kinematics and the knowledge of disturbance in the environment. It is assumed in Zhou et al. (2011) that the CNS aims at letting the arm track an ideal trajectory of a reference model. Although this mechanism does not require internal models of the human arm and the environment, the reference model needs to be identified from the experimental data or specially designed. The model in Hadjiosif et al. (2021) takes the derivative of the movement errors with respect to the parameters in control policies, such that the control policies are directly updated by gradient descent without the knowledge of an forward model. Although this model successfully characterizes some properties of the implicit adaption under mirror-reversed visual feedback, the relationship between the sensory output and the control policy is simply modeled as a static function, rather than a dynamical system used in our paper [see Eq. (1)]. Besides, both the models in Zhou et al. (2011) and Hadjiosif et al. (2021) only aim at minimizing sensory output errors, without minimizing the metabolic cost, which is considered in our model and is deemed to be one of the fundamental principles in sensorimotor systems by many previous studies (Todorov and Jordan 2002; Liu and Todorov 2007; Burdet et al. 2001; Franklin et al. 2008; Selen et al. 2009; Franklin and Wolpert 2011).

4.2 Robust reinforcement learning

With the control-dependent noise unmeasurable, we can only solve the data-based approximate ODE (18) of the model-based precise ODE (8), which is the main factor that causes the discrepancies ΔG_i between $\hat{G}_i$ and its true value $G(\hat{P}_i)$ in Algorithm 3. The simulation results in the last section show that even if the errors ΔG_i are present and can be as large as 90% of the true value (see Figs. 5 and 9), O-LSPI still successfully reproduces movement trajectories, velocity and endpoint force curves, stiffness geometries similar to those observed in experiments in Burdet et al. (2001, 2006) and Franklin et al. (2003). This is consistent with our theoretical result Theorem 1 on the robustness of reinforcement learning algorithms and suggests that the CNS is able to learn to adapt to the new environment and external disturbances in an error-tolerant, robust way. Such observations and implications are not included or explored in the previously proposed computational mechanisms based on optimal control theory in Todorov and Jordan (2002), Harris and Wolpert (1998), Burdet et al. (2001, 2006), Franklin et al. (2003), Wolpert et al. (1995), Selen et al. (2009), Zhou et al. (2017), Kadiallah et al. (2011), Mistry et al. (2013), Česonis and Franklin (2020), Haith and Krakauer (2013), d'Acremont et al. (2009), Fiete et al. (2007), Jiang and Jiang (2014) and Bian et al. (2020).

It is worth emphasizing that the robustness issue studied in this paper is different from the robustness problem considered in Ueyama (2014), Jiang and Jiang (2015), Crevecoeur et al. (2019), Gravell et al. (2020) and Bian et al. (2020). The robustness of the closed-loop system consisting of the optimal controller (the optimal policy learned by the CNS) and the controlled plant (the sensorimotor system and the environment that the CNS is interacting with) to external disturbances is analyzed in Ueyama (2014), Jiang and Jiang (2015), Crevecoeur et al. (2019), Gravell et al. (2020) and Bian et al. (2020), while this paper is devoted more exclusively to investigating the robustness of the learning algorithm, Algorithm 3, against the errors in the learning process.

4.3 Exploration noise

For the convergence of our proposed computational algorithm, it is necessary to add the exploration noise y to the state-feedback term $-\hat{K}_i x$ in (10). Indeed, without adding the exploration noise, the control input u_i is a linear combination of the state x, i.e., $u_i = -\hat{K}_i x$, and thus, Assumption 1 cannot hold. As a result, there is no guarantee that the least-squares problem in (16) has reasonable solutions and Algorithm 3 will converge to a small neighborhood of the optimal solution. In other words, our computational mechanism suggests that the intrinsic control-dependent noises ξ_1 and ξ_2 in the human arm movement model (22) alone are not enough to guarantee that the CNS is able to successfully adapt to the force fields in the arm reaching experiments, and extrinsic exploration noise actively added to the control input is indispensable in the sensorimotor learning. This is consistent with the discoveries in the literature that the CNS actively regulates the motor variability through extrinsic noise to facilitate motor learning (Wu et al. 2014; Sternad 2018). In fact, it is reported that noise is even able to teach people to make specific movements (Thorp et al. 2017). Evidences can also be found in the study of songbird learning (Fiete et al. 2007; Tumer and Brainard 2007). The song-related avian brain area—robust nucleus of the arcopallium (RA)—is responsible for the song production (Hahnloser et al. 2002), i.e., the controller of the avian song control system. It is found in Fiete et al. (2007) that another song-related avian brain area lateral magnocellular nucleus of the nidopallium (LMAN) produces fluctuating glutamatergic input to area RA, to generate behavioral variability in the songbirds for trial-and-error learning. In particular, lesion of area LMAN has little immediate effect on song production in adults, but arrests song learning in juveniles (Fiete et al. 2007). Thus, the extrinsic signal provided to area RA by area LMAN in songbird learning plays

the same role to that of the extrinsic exploration noise in our Algorithm 3.

4.4 Infinite-horizon optimal control

In our proposed computational method, the cost function (2) penalizes the trajectory deviation and the energy consumption over an infinite time horizon, which yields a time-invariant control policy in each trial. This infinite-horizon optimal control framework has a main advantage that movement duration needs not to be prescribed by the CNS (Huh et al. 2010; Huh 2012; Qian et al. 2013; Li et al. 2015; Česonis and Franklin 2021). Intuitively this seems to be more realistic because the duration of each movement is difficult to be prescribed and is different from each other as a result of randomness in the trajectories caused by the signal-dependent noise. The finite-horizon optimal control framework is utilized in Liu and Todorov (2007), where the cost function is integrated over a prescribed finite time interval and the resultant control policy is time-varying. It is suggested in Liu and Todorov (2007) that if the target is not reached at the prescribed reaching time, the CNS can similarly plan an independent new trajectory between the actual position of the hand and the final target, and the final trajectory will be the superposition of all the trajectories. By contrast, our model matches the intuition that the motor system keeps moving the hand toward the target until it is reached.

Although both finite-horizon models and infinite-horizon models have been widely and successfully used to explain the goal-directed reaching tasks, it is still an unsettled problem which kind of models humans actually use (Li et al. 2015). Both of these models have unique merits and limitations (see Česonis and Franklin 2021 and the references therein for details). Recently, a novel model combining the strengths of the finite-horizon optimal control and the infinite-horizon optimal control is proposed in Česonis and Franklin (2021) to address their individual limitations.

5 Conclusion

In this paper, we have proposed a new computational mechanism based on robust reinforcement learning, named optimistic least-squares policy iteration (O-LSPI), to model the robustness phenomenon in motor learning, control and adaptation. The O-LSPI model suggests that in spite of the estimation errors caused by the unmeasurable control-dependent noise, the CNS could still find near-optimal control policies directly from the noisy sensory data. Simulated movement trajectories, velocity and endpoint force curves and stiffness geometries consistent with the experimental observations in Burdet et al. (2001, 2006) and Franklin et al. (2003) are reproduced using the O-LSPI model, with estimation errors as large as 90% of the corresponding true values. Therefore, we argue that human sensorimotor systems may use a mechanism in the spirit of robust reinforcement learning to achieve motor learning, control and adaptation that are robust to the estimation errors present in the sensorimotor interactions.

Appendix A Proof of Theorem 2

Let $\epsilon = \epsilon_1/2$ in Theorem 1. We firstly show that for each stabilizing $\hat{K}_i \in \mathbb{R}^{m\times n}$, there exist an integer N_i and a constant $s_i > 0$, such that if Assumption 1 is satisfied, then for any $N > N_i$ and $s > s_i$, almost surely

$$\|\Delta G_i\|_F < \delta. \tag{27}$$

By definition,

$$\begin{aligned}\|\Delta G_i\|_F &= \|G(\hat{P}_i) - \hat{G}_i\|_F \le 2\|\theta(\hat{P}_i) - \hat{\theta}(\check{P}_i(s))\|_F \\ &= 2\|\theta(\hat{P}_i) - \theta(\check{P}_i(s)) + \theta(\check{P}_i(s)) - \hat{\theta}(\check{P}_i(s))\|_F \\ &\le 2C_0\|\hat{P}_i - \check{P}_i(s)\|_F + 2\|\theta(\check{P}_i(s)) - \hat{\theta}(\check{P}_i(s))\|_F,\end{aligned}$$

where $C_0 > 0$ is a constant. Then, we only need to show that for any $N > N_i$ and $s > s_i$,

$$\|\hat{P}_i - \check{P}_i(s)\|_F < \frac{\delta}{4C_0}, \quad \|\theta(\check{P}_i(s)) - \hat{\theta}(\check{P}_i(s))\|_F < \frac{\delta}{4}.$$

Vectorizing (15) yields

$$\begin{aligned}\dot{\bar{p}}_i &= \mathcal{T}(\Phi_{i,M,N}, \Psi_{i,M,N}, \hat{K}_i)\bar{p}_i \\ &\quad + \left[I_n, -\hat{K}_i^{\mathrm{T}}\right] \otimes \left[I_n, -\hat{K}_i^{\mathrm{T}}\right] \mathrm{vec}(Q \oplus R),\end{aligned} \tag{28}$$

where $\bar{p}_i = \mathrm{vec}(\bar{P}_i)$, i.e., the vectorization of matrix $\bar{P}_i$,

$$\begin{aligned}&\mathcal{T}(\Phi_{i,M,N}, \Psi_{i,M,N}, \hat{K}_i) \\ &= \left[I_n, -\hat{K}_i^{\mathrm{T}}\right] \otimes \left[I_n, -\hat{K}_i^{\mathrm{T}}\right] \left(I_{(m+n)^2} - (0_n \oplus I_m)\right. \\ &\quad \left.\otimes(0_n \oplus I_m)\right) D_{m+n}\Phi_{i,M,N}^{\dagger}\Psi_{i,M,N} D_n^{\dagger},\end{aligned}$$

and for $Y \in \mathbb{S}^n$, $D_n \in \mathbb{R}^{n^2\times \frac{1}{2}n(n+1)}$ is the unique matrix with full column rank (Magnus and Neudecker 2007, Page 57) such that

$$\mathrm{vec}(Y) = D_n\,\mathrm{svec}(Y), \quad \mathrm{svec}(Y) = D_n^{\dagger}\,\mathrm{vec}(Y).$$

Vectorizing (18) yields

$$\begin{aligned}\dot{\check{p}}_i &= \mathcal{T}(\hat{\Phi}_{i,M,N}, \hat{\Psi}_{i,M,N}, \hat{K}_i)\check{p}_i \\ &\quad + \left[I_n, -\hat{K}_i^{\mathrm{T}}\right] \otimes \left[I_n, -\hat{K}_i^{\mathrm{T}}\right] \mathrm{vec}(Q \oplus R),\end{aligned} \tag{29}$$

where $\check{p}_i = \mathrm{vec}(\check{P}_i)$. Since (8), (9), (28) and (15) are mutually equivalent with Assumption 1, by (17)

$$\begin{aligned}&\lim_{N\to\infty} \mathcal{T}(\hat{\Phi}_{i,M,N}, \hat{\Psi}_{i,M,N}, \hat{K}_i)\\ &= \mathcal{T}(\Phi_{i,M,N}, \Psi_{i,M,N}, \hat{K}_i)\\ &= \left(I_n \otimes (A - B\hat{K}_i)^{\mathrm{T}} + (A - B\hat{K}_i)^{\mathrm{T}} \otimes I_n\right).\end{aligned} \tag{30}$$

Since $\hat{K}_i$ is stabilizing, by continuity, there exists an integer $N_{i,1}$, such that for any $N > N_{i,1}$, $\mathcal{T}(\hat{\Phi}_{i,M,N}, \hat{\Psi}_{i,M,N}, \hat{K}_i)$ is Hurwitz. Then, we have

$$\lim_{t\to\infty} \bar{P}_i(t) = \hat{P}_i, \qquad \lim_{t\to\infty} \check{P}_i(t) = \mathring{P}_i, \tag{31}$$

where

$$\begin{aligned}\mathrm{vec}(\hat{P}_i) &= -\mathcal{T}^{\dagger}(\Phi_{i,M,N}, \Psi_{i,M,N}, \hat{K}_i)\\ &\quad \left[I_n, -\hat{K}_i^{\mathrm{T}}\right] \otimes \left[I_n, -\hat{K}_i^{\mathrm{T}}\right] \mathrm{vec}(Q \oplus R),\\ \mathrm{vec}(\mathring{P}_i) &= -\mathcal{T}^{\dagger}(\hat{\Phi}_{i,M,N}, \hat{\Psi}_{i,M,N}, \hat{K}_i)\\ &\quad \left[I_n, -\hat{K}_i^{\mathrm{T}}\right] \otimes \left[I_n, -\hat{K}_i^{\mathrm{T}}\right] \mathrm{vec}(Q \oplus R).\end{aligned}$$

By continuity of matrix inversion, (17) and (31), there exist an integer $N_{i,2} \geq N_{i,1}$ and $s_i > 0$, such that for any $N > N_{i,2}$ and $s > s_i$

$$\|\hat{P}_i - \check{P}_i(s)\|_F \leq \|\hat{P}_i - \mathring{P}_i\|_F + \|\mathring{P}_i - \check{P}_i(s)\|_F < \frac{\delta}{4C_0},$$

and

$$\begin{aligned}&\|\theta(\check{P}_i(s)) - \hat{\theta}(\check{P}_i(s))\|_F\\ &\leq \|\hat{\Phi}_{i,M,N}^{\dagger}\hat{\Psi}_{i,M,N} - \Phi_{i,M,N}^{\dagger}\Psi_{i,M,N}\|_F \|\check{P}_i(s)\|_F < \frac{\delta}{4}.\end{aligned}$$

Setting $N_i = N_{i,2}$ completes the proof of (27). By Theorem 1, there exists an integer $\bar{I}$, such that if

$$\|\Delta G_i\|_F < \delta, \qquad i = 1, \ldots, \bar{I}, \tag{32}$$

then $\|\hat{K}_{\bar{I}} - K^*\|_F < \epsilon_1$. Condition (32) can be satisfied by setting

$$N_0 = \max(N_1, \ldots, N_{\bar{I}}), \quad s_0 = \max(s_1, \ldots, s_{\bar{I}}).$$

This completes the proof.

References

Acerbi L, Vijayakumar S, Wolpert DM (2017) Target uncertainty mediates sensorimotor error correction. PLoS ONE 12(1):1–21

Åström KJ, Wittenmark B (1995) Adaptive control, 2nd edn. Addison-Wesley, Reading

Bach DR, Dolan RJ (2012) Knowing how much you don't know: a neural organization of uncertainty estimates. Nat Rev Neurosci 13(8):572–586

Balas G, Chiang R, Packard A, Safonov M (2007) Robust control toolbox user's guide. The Math Works Inc, Tech Rep

Bertsekas DP (2011) Approximate policy iteration: a survey and some new methods. J Control Theory Appl 9(3):310–335

Bertsekas DP (2019) Reinforcement learning and optimal control. Athena Scientific, Belmont

Bian T, Jiang ZP (2019) Continuous-time robust dynamic programming. SIAM J Control Optim 57(6):4150–4174

Bian T, Jiang Y, Jiang ZP (2016) Adaptive dynamic programming for stochastic systems with state and control dependent noise. IEEE Trans Autom Control 61(12):4170–4175

Bian T, Wolpert DM, Jiang ZP (2020) Model-free robust optimal feedback mechanisms of biological motor control. Neural Comput 32(3):562–595

Braun DA, Aertsen A, Wolpert DM, Mehring C (2009) Learning optimal adaptation strategies in unpredictable motor tasks. J Neurosci 29(20):6472–6478

Burdet E, Osu R, Franklin DW, Milner TE, Kawato M (2001) The central nervous system stabilizes unstable dynamics by learning optimal impedance. Nature 414(6862):446–449

Burdet E, Tee KP, Mareels I, Milner TE, Chew CM, Franklin DW, Osu R, Kawato M (2006) Stability and motor adaptation in human arm movements. Biol Cybern 94(1):20–32

Česonis J, Franklin DW (2020) Time-to-target simplifies optimal control of visuomotor feedback responses. eNeuro 7(2):ENEURO.0514–19.2020

Česonis J, Franklin DW (2021) Mixed-horizon optimal feedback control as a model of human movement. arXiv preprint arXiv:2104.06275

Cluff T, Scott SH (2015) Apparent and actual trajectory control depend on the behavioral context in upper limb motor tasks. J Neurosci 35(36):12465–12476

Crevecoeur F, Scott SH, Cluff T (2019) Robust control in human reaching movements: a model-free strategy to compensate for unpredictable disturbances. J Neurosci 39(41):8135–8148

Crevecoeur F, Thonnard JL, Lefèvre P (2020) A very fast time scale of human motor adaptation: within movement adjustments of internal representations during reaching. eNeuro 7(1):1–16

d'Acremont M, Lu ZL, Li X, Van der Linden M, Bechara A (2009) Neural correlates of risk prediction error during reinforcement learning in humans. NeuroImage 47(4):1929–1939

Fiete IR, Fee MS, Seung HS (2007) Model of birdsong learning based on gradient estimation by dynamic perturbation of neural conductances. J Neurophysiol 98(4):2038–2057

Fitts PM (1954) The information capacity of the human motor system in controlling the amplitude of movement. J Exp Psychol 47(6):381

Flash T, Hogan N (1985) The coordination of arm movements: an experimentally confirmed mathematical model. J Neurosci 5(7):1688–1703

Franklin DW, Wolpert DM (2011) Computational mechanisms of sensorimotor control. Neuron 72(3):425–442

Franklin DW, Burdet E, Osu R, Kawato M, Milner TE (2003) Functional significance of stiffness in adaptation of multijoint arm movements to stable and unstable dynamics. Exp Brain Res 151(2):145–157

Franklin DW, Burdet E, Peng Tee K, Osu R, Chew CM, Milner TE, Kawato M (2008) CNS learns stable, accurate, and efficient movements using a simple algorithm. J Neurosci 28(44):11165–11173

Gaveau J, Berret B, Demougeot L, Fadiga L, Pozzo T, Papaxanthis C (2014) Energy-related optimal control accounts for gravitational load: comparing shoulder, elbow, and wrist rotations. J Neurophysiol 111(1):4–16

Gomi H, Kawato M (1996) Equilibrium-point control hypothesis examined by measured arm stiffness during multijoint movement. Science 272(5258):117–120
Gravell BJ, Esfahani PM, Summers TH (2020) Robust control design for linear systems via multiplicative noise. IFAC-PapersOnLine 53(2):7392–7399
Hadjiosif AM, Krakauer JW, Haith AM (2021) Did we get sensorimotor adaptation wrong? Implicit adaptation as direct policy updating rather than forward-model-based learning. J Neurosci 41(12):2747–2761
Hahnloser RHR, Kozhevnikov AA, Fee MS (2002) An ultra-sparse code underliesthe generation of neural sequences in a songbird. Nature 419(6902):65–70
Haith AM, Krakauer JW (2013) Model-based and model-free mechanisms of human motor learning. In: Richardson MJ, Riley MA, Shockley K (eds) Progress in motor control. Springer, New York, pp 1–21
Harris CM, Wolpert DM (1998) Signal-dependent noise determines motor planning. Nature 394(6695):780–784
Huang VS, Haith A, Mazzoni P, Krakauer JW (2011) Rethinking motor learning and savings in adaptation paradigms: model-free memory for successful actions combines with internal models. Neuron 70(4):787–801
Huh D (2012) Rethinking optimal control of human movements. PhD thesis, UC San Diego
Huh D, Todorov E, Sejnowski T et al (2010) Infinite horizon optimal control framework for goal directed movements. In: Proceedings of the 9th annual symposium on advances in computational motor control, vol 12
Izawa J, Shadmehr R (2011) Learning from sensory and reward prediction errors during motor adaptation. PLoS Comput Biol 7(3):e1002012
Jiang Y, Jiang ZP (2014) Adaptive dynamic programming as a theory of sensorimotor control. Biol Cybern 108(4):459–473
Jiang Y, Jiang ZP (2015) A robust adaptive dynamic programming principle for sensorimotor control with signal-dependent noise. J Syst Sci Complex 28(2):261–288
Jiang Y, Jiang ZP (2017) Robust adaptive dynamic programming. Wiley-IEEE Press, Hoboken
Jiang Z, Bian T, Gao W (2020) Learning-based control: a tutorial and some recent results. Found Trends Syst Control 8:176–284
Kadiallah A, Liaw G, Kawato M, Franklin DW, Burdet E (2011) Impedance control is selectively tuned to multiple directions of movement. J Neurophysiol 106(5):2737–2748
Kamalapurkar R, Walters P, Rosenfeld J, Dixon W (2018) Reinforcement learning for optimal feedback control: a Lyapunov-based approach. Springer, Berlin
Khalil HK (2002) Nonlinear systems, 3rd edn. Prentice-Hall, Upper Saddle River
Kiumarsi B, Vamvoudakis KG, Modares H, Lewis FL (2018) Optimal and autonomous control using reinforcement learning: a survey. IEEE Trans Neural Netw Learn Syst 29(6):2042–2062
Kleinman D (1968) On an iterative technique for Riccati equation computations. IEEE Trans Autom Control 13(1):114–115
Kleinman D (1969) On the stability of linear stochastic systems. IEEE Trans Autom Control 14(4):429–430
Körding KP, Wolpert DM (2004) Bayesian integration in sensorimotor learning. Nature 427(6971):244–247
Körding KP, Wolpert DM (2006) Bayesian decision theory in sensorimotor control. Trends Cognit Sci 10(7):319–326
Krakauer JW, Hadjiosif AM, Xu J, Wong AL, Haith AM (2019) Motor learning. American Cancer Society, Atlanta, pp 613–663
Li L, Imamizu H, Tanaka H (2015) Is movement duration predetermined in visually guided reaching? A comparison of finite-and infinite-horizon optimal feedback control. In: The abstracts of the international conference on advanced mechatronics: toward evolutionary fusion of IT and mechatronics: ICAM 2015.6. The Japan Society of Mechanical Engineers, pp 247–248
Liberzon D (2012) Calculus of variations and optimal control theory: a concise introduction. Princeton University Press, Princeton
Liu D, Todorov E (2007) Evidence for the flexible sensorimotor strategies predicted by optimal feedback control. J Neurosci 27(35):9354–9368
Magnus JR, Neudecker H (2007) Matrix differential calculus with applications in statistics and econmerices. Wiley, New York
Mistry M, Theodorou E, Schaal S, Kawato M (2013) Optimal control of reaching includes kinematic constraints. J Neurophysiol 110(1):1–11
Morasso P (1981) Spatial control of arm movements. Exp Brain Res 42(2):223–227
Mori T, Fukuma N, Kuwahara M (1986) On the Lyapunov matrix differential equation. IEEE Trans Autom Control 31(9):868–869
Mussa-Ivaldi F, Hogan N, Bizzi E (1985) Neural, mechanical, and geometric factors subserving arm posture in humans. J Neurosci 5(10):2732–2743
Orbán G, Wolpert DM (2011) Representations of uncertainty in sensorimotor control. Curr Opin Neurobiol 21(4):629–635
Pang B, Jiang ZP (2020) Adaptive optimal control of linear periodic systems: an off-policy value iteration approach. IEEE Trans Autom Control 66(2):888–894
Pang B, Jiang ZP (2021) Robust reinforcement learning: a case study in linear quadratic regulation. In: The 35th AAAI conference on artificial intelligence (AAAI). pp 9303–9311
Pang B, Bian T, Jiang ZP (2019) Adaptive dynamic programming for finite-horizon optimal control of linear time-varying discrete-time systems. Control Theory Technol 17(1):18–29
Pang B, Jiang ZP, Mareels I (2020) Reinforcement learning for adaptive optimal control of continuous-time linear periodic systems. Automatica 118:109035
Pang B, Bian T, Jiang ZP (2021) Robust policy iteration for continuous-time linear quadratic regulation. IEEE Trans Autom Control. https://doi.org/10.1109/TAC.2021.3085510
Parker A, Derrington A, Blakemore C, van Beers RJ, Baraduc P, Wolpert DM (2002) Role of uncertainty in sensorimotor control. Philos Trans R So Lond Ser B Biol Sci 357(1424):1137–1145
Pavliotis GA (2014) Stochastic processes and applications. Springer, New York
Qian N, Jiang Y, Jiang ZP, Mazzoni P (2013) Movement duration, Fitts's law, and an infinite-horizon optimal feedback control model for biological motor systems. Neural Comput 25(3):697–724
Schmidt RA, Lee TD, Winstein C, Wulf G, Zelaznik HN (2018) Motor control and learning: a behavioral emphasis. In: Human kinetics
Selen LPJ, Franklin DW, Wolpert DM (2009) Impedance control reduces instability that arises from motor noise. J Neurosci 29(40):12606–12616
Shadmehr R, Mussa-Ivaldi S (2012) Biological learning and control: how the brain builds representations, predicts events, and makes decisions. MIT Press, Cambridge
Shmuelof L, Huang VS, Haith AM, Delnicki RJ, Mazzoni P, Krakauer JW (2012) Overcoming motor forgetting through reinforcement of learned actions. J Neurosci 32(42):14617–14621a
Sternad D (2018) It's not (only) the mean that matters: variability, noise and exploration in skill learning. Curr Opin Behav Sci 20:183–195
Sternad D, Abe MO, Hu X, Müller H (2011) Neuromotor noise, error tolerance and velocity-dependent costs in skilled performance. PLOS Comput Biol 7(9):1–15
Sutton RS, Barto AG (2018) Reinforcement learning: an introduction, 2nd edn. MIT Press, Cambridge
Thorp EB, Kording KP, Mussa-Ivaldi FA (2017) Using noise to shape motor learning. J Neurophysiol 117(2):728–737

Todorov E (2005) Stochastic optimal control and estimation methods adapted to the noise characteristics of the sensorimotor system. Neural Comput 17(5):1084–1108

Todorov E, Jordan MI (2002) Optimal feedback control as a theory of motor coordination. Nat Neurosci 5(11):1226–1235

Tsitsiklis JN (2002) On the convergence of optimistic policy iteration. J Mach Learn Res 3:59–72

Tumer EC, Brainard MS (2007) Performance variability enables adaptive plasticity of 'crystallized' adult birdsong. Nature 450(7173):1240–1244

Ueyama Y (2014) Mini-max feedback control as a computational theory of sensorimotor control in the presence of structural uncertainty. Front Comput Neurosci 8:1–14

Uno Y, Kawato M, Suzuki R (1989) Formation and control of optimal trajectory in human multijoint arm movement. Biol Cybern 61(2):89–101

Vaswani PA, Shmuelof L, Haith AM, Delnicki RJ, Huang VS, Mazzoni P, Shadmehr R, Krakauer JW (2015) Persistent residual errors in motor adaptation tasks: reversion to baseline and exploratory escape. J Neurosci 35(17):6969–6977

Willems JL, Willems JC (1976) Feedback stabilizability for stochastic systems with state and control dependent noise. Automatica 12(3):277–283

Wolpert DM (2007) Probabilistic models in human sensorimotor control. Hum Mov Sci 26(4):511–524

Wolpert D, Ghahramani Z, Jordan M (1995) An internal model for sensorimotor integration. Science 269(5232):1880–1882

Wu HG, Miyamoto YR, Castro LNG, Ölveczky BP, Smith MA (2014) Temporal structure of motor variability is dynamically regulated and predicts motor learning ability. Nat Neurosci 17(2):312–321

Yeo SH, Franklin DW, Wolpert DM (2016) When optimal feedback control is not enough: feedforward strategies are required for optimal control with active sensing. PLOS Comput Biol 12:1–22

Zhou K, Doyle JC (1998) Essentials of robust control, vol 104. Prentice Hall, Upper Saddle River

Zhou SH, Oetomo D, Tan Y, Burdet E, Mareels I (2011) Human motor learning through iterative model reference adaptive control. IFAC Proc Vol 44(1):2883–2888

Zhou SH, Tan Y, Oetomo D, Freeman C, Burdet E, Mareels I (2017) Modeling of endpoint feedback learning implemented through point-to-point learning control. IEEE Trans Control Syst Technol 25(5):1576–1585