2023 62nd IEEE Conference on Decision and Control (CDC)
December 13-15, 2023. Marina Bay Sands, Singapore

Reinforcement-Learning-Based Risk-Sensitive Optimal Feedback Mechanisms of Biological Motor Control

Leilei Cui, Bo Pang, and Zhong-Ping Jiang

Abstract— **Risk sensitivity is a fundamental aspect of biological motor control that accounts for both the expectation and variability of movement cost in the face of uncertainty. However, most computational models of biological motor control rely on model-based risk-sensitive optimal control, which requires an accurate internal representation in the central neural system to predict the outcomes of motor commands. In reality, the dynamics of human-environment interaction is too complex to be accurately modeled, and noise further complicates system identification. To address this issue, this paper proposes a novel risk-sensitive computational mechanism for biological motor control based on reinforcement learning (RL) and adaptive dynamic programming (ADP). The proposed ADP-based mechanism suggests that humans can directly learn an approximation of the risk-sensitive optimal feedback controller from noisy sensory data without the need for system identification. Numerical validation of the proposed mechanism is conducted on the arm-reaching task under divergent force field. The preliminary computational results align with the experimental observations from the past literature of computational neuroscience.**

I. INTRODUCTION

The computational mechanism underlying goal-directed movements has been the subject of extensive study over the past decades, seeking to explain how and why the brain selects a particular movement to complete a reaching task from a large set of possibilities. For instance, reaching trajectories typically exhibit a roughly straight path with a bell-shaped velocity profile [1]. Optimal feedback control is a widely accepted computational model that explains the behavior of human motor control [2], [3]. It posits that the central nervous system (CNS) exploits a feedback scheme to correct task-relevant errors, and that the optimal feedback control law is computed by minimizing a mixed cost function dependent on the task error and control command. However, these proposed optimal control models are risk-neutral and do not account for the variability of the movement cost.

Risk attitude is a fundamental aspect of human decision-making, with individuals exhibiting preferences for different levels of risk. One of the first mathematical models to quantify the magnitude of risk sensitivity was developed by Daniel Bernoulli, who hypothesized that humans maximize the logarithm of the monetary gain or minimize the exponential of the cost when making decisions [4]. Building on Bernoulli's work, in [5], Jacobson proposed the risk-sensitive optimal control framework, which generalizes the risk-neutral optimal control model by incorporating risk attitude. Interestingly, the results of the risk-averse optimal control coincide with that of the mixed $\mathscr{H}_2/\mathscr{H}_\infty$ control and zero-sum differential game [6], [7]. This framework has been applied to the study of human motor control, and many studies have shown that human subjects exhibit risk sensitivity in their motor behaviors [8], [9], [10], [11], [12], [13].

In the field of human motor control and learning, it is widely believed that humans first learn an internal model of the dynamics that simulates the behavior of the motor system [14]. However, there is currently no experimental or theoretical evidence to support this view of constructing an internal model for motor control. Moreover, the complexity of the environment and the sensory noise make internal model construction even more challenging. Reinforcement learning (RL) and adaptive dynamic programming (ADP) are biologically-inspired learning-based control approaches that directly minimize the cumulative cost or maximize the cumulative reward by interacting continuously with the environment with no need to identify the environment [15], [16]. Therefore, various computational models have been developed based on RL and ADP to account for the observed phenomena of human motor control and learning; see [17], [18], [19], [20] and related works therein. These models have demonstrated promising results in explaining the robustness, adaptivity, and flexibility of human motor control and learning, and they may provide a viable alternative to the traditional model-based approach. However, the risk sensitivity of human subjects is neglected in these RL and ADP based mechanisms.

In this paper, we propose a new approach to explain how the CNS learns a risk-sensitive optimal controller in a model-free manner. We argue that robust RL, specifically robust value iteration, offers a powerful adaptive optimal control method that does not require a model of the environment. First, we demonstrate that value iteration is robust to errors that may occur during the learning process of solving the risk-sensitive optimal control problem. We prove that value iteration can still find a near-optimal solution of the risk-sensitive optimal control problem as long as the noise at each learning step is sufficiently small. Building on this robustness property, we propose a novel learning-based control algorithm to explain the CNS's learning process for solving the risk-sensitive optimal control problem in the absence of the exact model knowledge of human-environment interaction.

*This work has been supported in part by the NSF grants EPCN-1903781, CNS-2227153, and ECCS-2210320.

L. Cui, B. Pang, and Z. P. Jiang are with the Control and Networks Lab, Department of Electrical and Computer Engineering, Tandon School of Engineering, New York University, Brooklyn, NY 11201, USA (l.cui@nyu.edu; bo.pang@nyu.edu; zjiang@nyu.edu).

979-8-3503-0123-6/23/$31.00 ©2023 IEEE

Our numerical simulations show that the proposed learning-based value iteration algorithm can find the near-optimal risk-sensitive controller even in the presence of unmeasurable noises. Furthermore, the numerical simulation results match the experimental results reported in the past literature of computational neuroscience [21], [22].

The rest of this paper is organized as follows: Section II introduces the preliminaries of the risk-sensitive optimal control, theoretically demonstrates the robustness of value iteration, and proposes a learning-based value iteration algorithm. In Section III, the numerical simulation of the proposed learning-based value iteration algorithm is conducted for the arm-reaching task. Finally, some concluding remarks are given in Section IV.

Notations: $\mathbb{S}^n$ denotes the set of n-dimensional, real symmetric matrices. I_n denotes the n-dimensional identity matrix. $\|\cdot\|$ denotes the spectral norm of a matrix or Euclidean norm of a vector. $\|\cdot\|_F$ denotes the Frobenius norm of a matrix. $\|\cdot\|_\infty$ denotes the supremum norm of a matrix-valued function, i.e. $\|\Delta\|_\infty = \sup\{\|\Delta(s)\|_F : s \in \mathscr{I}\}$, where $\mathscr{I}$ is the domain of $\Delta(\cdot)$. For any $P \in \mathbb{S}^n$, $\mathrm{vecs}(P) = [p_{1,1}, \sqrt{2}p_{1,2}, \cdots, p_{2,2}, \sqrt{2}p_{2,3}, \cdots, p_{n,n}]^T$.

II. Robust Reinforcement Learning

A. Risk-Sensitive Optimal Control

Consider the following linear stochastic system characterizing human movement

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

$$y(t) = Ex(t) + Fu(t), \tag{1b}$$

where $x \in \mathbb{R}^n$ is the state; $u \in \mathbb{R}^m$ is the control input; $y \in \mathbb{R}^r$ is the controlled output; $w_k \in \mathbb{R}$ $(k = 1, \cdots, q)$ and $\zeta \in \mathbb{R}^p$ are unmeasurable noises that are independent Brownian motion; $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times m}$ are constant matrices associated with the system dynamics; $C_k \in \mathbb{R}^{n\times m}$ is the gain matrix of the control-dependent noise; $D \in \mathbb{R}^{n\times p}$ is the gain matrix of the process noise; $E \in \mathbb{R}^{r\times n}$ and $F \in \mathbb{R}^{r\times m}$ are constant matrices satisfying $E^T F = 0$, $E^T E = Q \succ 0$, and $F^T F = R \succ 0$. The control-dependent noise captures the physiological observation that the variation of muscle force grows linearly with its mean [2], [23]. The process noise ζ characterizes the Gaussian-type noise within the sensory motor system.

Following [5], [8], [9], the risk-sensitive optimal control problem entails finding a controller $u(t) = u(x(t))$ for system (1) without control-dependent noise, that minimizes the following exponential quadratic cost:

$$\mathscr{J}(x(0), u) = \lim_{\tau\to\infty} \frac{1}{\tau}\frac{2}{\alpha} \log \mathbb{E} \exp\left(\frac{\alpha}{2}\int_0^\tau y^T y\,\mathrm{d}t\right). \tag{2}$$

Here, α is a constant that describes the intensity of risk-sensitivity. By taking the Taylor expansion of $\mathscr{J}(x(0), u)$ around $\alpha = 0$, we have

$$\mathscr{J}(x(0), u) = \lim_{\tau\to\infty} \frac{1}{\tau}\left[\mathbb{E}\left(\int_0^\tau y^T y\,\mathrm{d}t\right) + \frac{\alpha}{4}\mathbb{V}\left(\int_0^\tau y^T y\,\mathrm{d}t\right)\right] + O(\alpha^2), \tag{3}$$

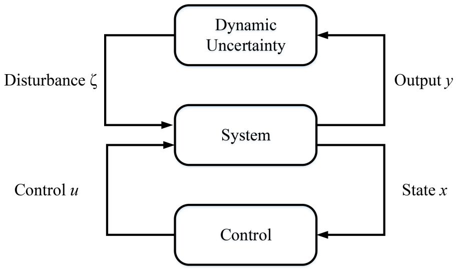

Fig. 1. The risk-sensitive optimal controller is robust to the dynamic uncertainty with the gain less than or equal to $\sqrt{\alpha}$.

where $\mathbb{V}$ stands for the variance. Hence, if $\alpha > 0$, human subjects are risk-averse since they minimize the expectation and variance of the cost. In contrast, $\alpha < 0$ implies that human subjects are risk-seeking since they prefer a large variance of the cost. When $\alpha \to 0$, (2) is equivalent to the risk-neutral cost studied in [2]. To ensure that the risk-sensitive optimal control problem is solvable, it is assumed that $\alpha < \alpha_\infty$, where α_∞ is the maximum of α such that there exists a controller under which the cost (2) is finite for all $\alpha < \alpha_\infty$.

According to the risk-sensitive optimal control theory, if the system matrices are known, as studied in [5], the optimal controller is

$$u^*(t) = -K^* x(t), \tag{4}$$

where $K^* = R^{-1}B^T P^*$ and $P^* \in \mathbb{S}^n$ is the unique positive definite solution of the generalized algebraic Riccati equation

$$A^T P^* + P^* A + Q - P^*(BR^{-1}B^T - \alpha DD^T)P^* = 0. \tag{5}$$

When $\alpha \geq 0$, one important feature of the risk-averse optimal controller is that K^* is stabilizing in the sense that all the eigenvalues of $A - BK^*$ have negative real parts. Furthermore, as studied in [6] and illustrated in Fig. 1, when system (1a) is connected with the dynamic uncertainty whose gain is less than or equal to $\sqrt{\alpha}$ $(\alpha \geq 0)$, the stability of the closed-loop system is still ensured. Here, $\frac{\mathrm{d}\zeta}{\mathrm{d}t}$ is no longer a Gaussian-type noise as in (1a), but a deterministic disturbance as the output of the dynamic uncertainty. We conjecture that one of the reasons why human adopts risk-sensitive optimal control is to handle the dynamic uncertainty. Dynamic uncertainty may be caused by the mismatch between the human internal model and the exact dynamics of the motor system.

B. Robust Value Iteration

Value iteration is fundamental for the development of learning-based control algorithms. It finds the optimal solution P^* by solving the following differential Riccati equation

$$\frac{\mathrm{d}P(s)}{\mathrm{d}s} = A^T P(s) + P(s)A + Q - P(s)(BR^{-1}B^T - \alpha DD^T)P(s)$$
$$P(0) = 0. \tag{6}$$

The following lemma shows that the solution of (6) converges to the optimal solution P^* as $s \to \infty$.

Lemma 1: The solution of (6) converges to P^* as $s \to \infty$.

Proof: See [6, Theorem 9.7] for $0 \leq \alpha < \alpha_\infty$ and [24, Remark 21] for $\alpha < 0$, by considering $\bar{B} = [B, D]$ and $\bar{R} = \mathrm{diag}(R, \alpha^{-1} I_p)$. ■

Equation (6) can be viewed as the exact value iteration since it assumes that the precise system matrices are attainable for solving the differential equation (6). In reality, it is hard to get such an accurate model, and the update of $P(s)$ is subject to the errors that may come from the unmeasurable noises of the system. Considering the influence of the error, the inexact value iteration is

$$\begin{aligned}\frac{\mathrm{d} P_\Delta(s)}{\mathrm{d} s} &= A^T P_\Delta(s) + P_\Delta(s) A + Q \\ &- P_\Delta(s)(BR^{-1}B^T - \alpha DD^T) P_\Delta(s) + \Delta(s), \quad P_\Delta(0) = 0.\end{aligned} \tag{7}$$

where $\Delta(s) \in \mathbb{S}^n$ denotes the error during the learning process. Under the influence of the error $\Delta(s)$, it is important to investigate whether value iteration is robust to the error during the learning process, or in other words, whether (7) can still converge to a neighbourhood of P^*. The following theorem provides the solution to this problem.

Theorem 1: For any $\varepsilon > 0$, there exists $d(\varepsilon) > 0$, such that if $\|\Delta\|_\infty \leq d(\varepsilon)$, it holds

$$\limsup_{s \to \infty} \|P_\Delta(s) - P^*\|_F \leq \varepsilon. \tag{8}$$

Proof: Let P_i^* $(i = 1, 2)$ be the solutions of

$$A^T P_i^* + P_i^* A - P_i^*(BR^{-1}B^T - \alpha DD^T) P_i^* + d_i I_n + Q = 0, \tag{9}$$

where $d_1 = d$ and $d_2 = -d$. In addition, let $P_1(0) = P_2(0) = 0$, and $P_1(s)$ and $P_2(s)$ be the solutions of

$$\begin{aligned}\frac{\mathrm{d} P_i(s)}{\mathrm{d} s} &= A^T P_i(s) + P_i(s) A \\ &- P_i(s)(BR^{-1}B^T - \alpha DD^T) P_i(s) + d_i I_n + Q, \quad i = 1, 2\end{aligned} \tag{10}$$

Since $P_1^* = P_2^* = P^*$ when $d = 0$, and P_1^* and P_2^* are continuous with respect to d (can be demonstrated by the celebrated implicit function theorem), there exists $d(\varepsilon) > 0$,

$$\|P_1^* - P^*\|_F \leq \varepsilon, \quad \|P_2^* - P^*\|_F \leq \varepsilon. \tag{11}$$

According to the monotonicity of $P_\Delta(s)$ with respect to $\Delta(s)$ [25, Theorem 3.1], if $-d(\varepsilon) I_n \preceq \Delta(s) \preceq d(\varepsilon) I_n$ for all $s \geq 0$, we have $P_2(s) \preceq P_\Delta(s) \preceq P_1(s)$. In addition, by Lemma 1, $\lim_{s \to \infty} P_1(s) = P_1^*$ and $\lim_{s \to \infty} P_2(s) = P_2^*$. Hence, if $\|\Delta\|_\infty \leq d(\varepsilon)$, we have

$$0 \preceq \limsup_{s \to \infty} (P_\Delta(s) - P_2^*) \preceq P_1^* - P_2^*. \tag{12}$$

Taking the trace of (12) and considering $\mathrm{Tr}(P) \leq \sqrt{n} \|P\|_F$ yield

$$\limsup_{s \to \infty} \mathrm{Tr}(P_\Delta(s) - P_2^*) \leq \sqrt{n} \|P_1^* - P_2^*\|_F. \tag{13}$$

Plugging (11) into (13), and considering the trace bound in [26, Lemma 1] and $\|P\|_F \leq \mathrm{Tr}(P)$, we have

$$\limsup_{s \to \infty} \|P_\Delta(s) - P_2^*\|_F \leq 2\sqrt{n}\varepsilon. \tag{14}$$

Again, using the triangle inequality, it follows from (14) that

$$\limsup_{s \to \infty} \|P_\Delta(s) - P^*\|_F \leq (2\sqrt{n} + 1)\varepsilon. \tag{15}$$

Hence, the theorem is proved by resetting ε as $\frac{\varepsilon}{2\sqrt{n}+1}$. ■

C. Learning-Based Value Iteration

The value iteration algorithm in the previous subsection relies on the system matrices. When system matrices (A, B) are not attainable, and the noises ($\sum_{k=1}^{q} C_k u \,\mathrm{d} w_k$ and $D \,\mathrm{d}\zeta$ in (1)) are not measurable, it is hypothesized that the CNS utilizes the input-state trajectory data of (1a) to learn the risk-sensitive optimal controller.

To begin with, along the trajectories of system (1a) under the exploratory control input u and by Itô's lemma [27, Lemma 3.2], for any $X \in \mathbb{S}^n$, we have

$$\begin{aligned}\mathrm{d}(x^T X x) &= x^T (A^T X + XA) x \,\mathrm{d}t + 2u^T B^T X x \,\mathrm{d}t + 2x^T X D \,\mathrm{d}\zeta \\ &+ 2x^T X \sum_{k=1}^{q} C_k u \,\mathrm{d} w_k + \sum_{k=1}^{q} u^T C_k^T X C_k u \,\mathrm{d}t + \mathrm{Tr}(D^T X D) \,\mathrm{d}t.\end{aligned} \tag{16}$$

Integrating (16) from t_j to t_{j+1} with $t_{j+1} > t_j$ yields

$$\begin{aligned}x^T(t_{j+1}) X x(t_{j+1}) - x^T(t_j) X x(t_j) &= \int_{t_j}^{t_{j+1}} z^T(t) \theta(X) z(t) \,\mathrm{d}t \\ + 2\int_{t_j}^{t_{j+1}} x^T X \sum_{k=1}^{q} C_k u \,\mathrm{d} w_k &+ 2\int_{t_j}^{t_{j+1}} x^T X D \,\mathrm{d}\zeta,\end{aligned} \tag{17}$$

where

$$z = [x^T, u^T, 1]^T, \tag{18}$$

and

$$\begin{aligned}\theta(X) &= \begin{bmatrix} A^T X + XA & XB & 0_{n\times 1} \\ B^T X & \sum_{k=1}^{p} C_k^T X C_k & 0_{m\times 1} \\ 0_{1\times n} & 0_{1\times m} & \mathrm{Tr}(D^T X D) \end{bmatrix} \\ &= \begin{bmatrix} \theta_{xx}(X) & \theta_{xu}(X) & 0_{n\times 1} \\ \theta_{xu}^T(X) & \theta_{uu}(X) & 0_{m\times 1} \\ 0_{1\times n} & 0_{1\times m} & \theta_{\zeta\zeta}(X) \end{bmatrix}.\end{aligned} \tag{19}$$

Taking the expectation of (17) and considering $x^T X x = \mathrm{vecs}(xx^T)^T \mathrm{vecs}(X)$, we have

$$(\bar{x}_{j+1} - \bar{x}_j)^T \mathrm{vecs}(X) = \bar{z}_j^T \mathrm{vecs}(\theta(X)), \tag{20}$$

where

$$\begin{aligned}\bar{x}_j &= \mathbb{E}\left[\mathrm{vecs}(x(t_j) x^T(t_j))\right], \\ \bar{z}_j &= \mathbb{E}\left[\int_{t_j}^{t_{j+1}} \mathrm{vecs}(z(t) z^T(t)) \,\mathrm{d}t\right].\end{aligned} \tag{21}$$

Repeating (20) for $t_1 < t_2 < \cdots < t_M$ and stacking them into a vector form, we have

$$\Phi_M \,\mathrm{vecs}(\theta(X)) = \Psi_M \,\mathrm{vecs}(X), \tag{22}$$

where

$$\begin{aligned}\Phi_M &= [\bar{z}_1, \bar{z}_2, \cdots, \bar{z}_{M-1}]^T, \\ \Psi_M &= [\bar{x}_2 - \bar{x}_1, \bar{x}_3 - \bar{x}_2, \cdots, \bar{x}_M - \bar{x}_{M-1}]^T.\end{aligned} \tag{23}$$

The following assumption is made on the data-dependent matrix Φ_M.

Assumption 1: Φ_M is full column rank.

Remark 1: Assumption 1 is reminiscent of the classical persistent excitation condition in adaptive control to guarantee the uniqueness of the linear regression solution of (22).

Similar assumptions can be found in the literature of RL and ADP [15], [28], [29]. One can fulfill the assumption by means of adding exploratory noise to the control input u.

Under Assumption 1, $\theta(X)$ can be computed by the data-dependent matrices Φ_M and Ψ_M in the absence of the system matrices (A,B), that is

$$\mathrm{vecs}(\theta(X)) = \Phi_M^{\dagger}\Psi_M \,\mathrm{vecs}(X). \tag{24}$$

Plugging (24) into (6) with X replaced by $P(s)$ yields

$$\begin{aligned} \frac{\mathrm{d}P(s)}{\mathrm{d}s} &= \theta_{xx}(P(s)) + Q \\ &- \theta_{xu}(P(s))R^{-1}\theta_{xu}^T(P(s)) + \alpha P(s)DD^T P(s) \end{aligned} \tag{25}$$

It is noticed that in (25), $\theta(P(s))$ is computed by (24) in a model-free manner and the system matrices (A,B) are no longer required.

In (21) and (23), it is hard to compute the data-dependent matrices Φ_M and Ψ_M directly due to the computation of the expectation. Next, we use in total L trajectories to approximate the expectation. Particularly, L trajectories within the interval $[t_1,t_M]$ are collected from system (1) with the same initial condition and control input. Let the superscript l denote the lth input-state trajectory. Then, the approximations of $\bar{x}_j$ and $\bar{z}_j$ are

$$\begin{aligned} \hat{\bar{x}}_j &= \frac{1}{L}\sum_{l=1}^{L} \mathrm{vecs}(x^l(t_j)x^{l,T}(t_j)), \\ \hat{\bar{z}}_j &= \frac{1}{L}\sum_{l=1}^{L}\int_{t_j}^{t_{j+1}} \mathrm{vecs}(z^l(t)z^{l,T}(t))\,\mathrm{d}t. \end{aligned} \tag{26}$$

Consequently, the approximations of Φ_M and Ψ_M in (23) are

$$\begin{aligned} \hat{\Phi}_{M,L} &= [\hat{\bar{z}}_1, \hat{\bar{z}}_2, \cdots, \hat{\bar{z}}_{M-1}]^T, \\ \hat{\Psi}_{M,L} &= [\hat{\bar{x}}_2 - \hat{\bar{x}}_1, \hat{\bar{x}}_3 - \hat{\bar{x}}_2, \cdots, \hat{\bar{x}}_M - \hat{\bar{x}}_{M-1}]^T. \end{aligned} \tag{27}$$

By the Strong Law of Large Numbers, Φ_M and Ψ_M are well approximated by $\hat{\Phi}_{M,L}$ and $\hat{\Psi}_{M,L}$ when L is sufficiently large. In other words, the following relation holds *almost surely*

$$\lim_{L\to\infty}\hat{\Phi}_{M,L} = \Phi_M, \ \lim_{L\to\infty}\hat{\Psi}_{M,L} = \Psi_M. \tag{28}$$

For any $X \in \mathbb{S}^n$, the approximation of $\theta(X)$ in (24) is

$$\mathrm{vecs}(\hat{\theta}(X)) = \hat{\Phi}_{M,L}^{\dagger}\hat{\Psi}_{M,L}\,\mathrm{vecs}(X). \tag{29}$$

The learning-based value iteration is finally represented as

$$\begin{aligned} \frac{\mathrm{d}\hat{P}(s)}{\mathrm{d}s} &= \hat{\theta}_{xx}(\hat{P}(s)) + Q \\ &- \hat{\theta}_{xu}(\hat{P}(s))R^{-1}\hat{\theta}_{xu}^T(\hat{P}(s)) + \alpha\hat{P}(s)DD^T\hat{P}(s). \end{aligned} \tag{30}$$

The detailed algorithm is shown in Algorithm 1. The following main theorem guarantees the convergence of the algorithm to the near-optimal risk-sensitive optimal control.

Theorem 2: For any $\varepsilon > 0$, there exist $s_f > 0$ and $L^* > 0$, such that for all $L > L^*$

$$\left\|\hat{K}(s_f) - K^*\right\|_F \le \varepsilon, \tag{31}$$

where $\hat{K}(s_f) = R^{-1}\hat{\theta}_{xu}^T(\hat{P}(s_f))$.

Algorithm 1 Learning-based Value Iteration

```
Select the parameters $M$ and $L$.
Select the driving input $u(t)$ to explore system (1a) and collect the input-state data $u(t), x(t), t \in [t_1, t_M]$.
Select the terminal time $s_f$ of (30).
Construct data matrices $\hat{\Phi}_{M,L}$ and $\hat{\Psi}_{M,L}$ by (27).
Solve (30) on the interval $[0, s_f]$ and get $\hat{\theta}(\hat{P}(s_f))$.
Get $\hat{K}(s_f) = R^{-1}\hat{\theta}_{xu}^T(\hat{P}(s_f))$.
```

III. Numerical Studies for Arm Reaching Movement

A. Model of Arm Reaching

The arm-reaching tasks studied in [21], [22] are used to validate Algorithm 1, where human subjects are asked to finish the point-to-point movement on a horizontal plane. The aim is to reproduce the similar numerical results as the experimental results reported in [21], [22]. We consider the two-joint arm movement [30], of which the dynamics is

$$\mathrm{d}p = v\,\mathrm{d}t, \tag{32a}$$

$$m\,\mathrm{d}v = (a - bv + f)\,\mathrm{d}t + D_1\,\mathrm{d}\zeta_1, \tag{32b}$$

$$\tau\,\mathrm{d}a = (u - a)\,\mathrm{d}t + C_1 u\,\mathrm{d}w_1 + C_2 u\,\mathrm{d}w_2 + D_2\,\mathrm{d}\zeta_2, \tag{32c}$$

where $p = [p_x, p_y]^T$, $v = [v_x, v_y]^T$, and $a = [a_x, a_y]^T$ are the two-dimensional position, velocity, and actuator state of the hand. $f = [f_x, f_y]^T$ is the external force generated from the given force fields. $u = [u_x, u_y]^T$ is the motor command. $w_1, w_2 \in \mathbb{R}$ and $\zeta_1, \zeta_2 \in \mathbb{R}^2$ are standard Brownian motions. The parameters m, τ and b are the same as [19].

$$C_1 = \begin{bmatrix} c_1 & 0 \\ c_2 & 0 \end{bmatrix}, \ C_2 = \begin{bmatrix} 0 & c_2 \\ 0 & c_1 \end{bmatrix}, \tag{33}$$

where $c_1 = 0.0075$ and $c_2 = 0.0025$. $D_1 = 0.13I_2$ and $D_2 = 0.005I_2$. The state-space representation of system (32) is

$$\mathrm{d}x = (Ax + Gf + Bu)\,\mathrm{d}t + B(C_1 u\,\mathrm{d}w_1 + C_2 u\,\mathrm{d}w_2) + D\,\mathrm{d}\zeta, \tag{34}$$

where A, B, D and G can be computed from (32). $\alpha = 0.25$. The weighting matrices of the cost (2) are the same as [19]. The external force f in (32) is generated from two force fields, including null field (NF) and divergent force field (DF) [22], [31]. For NF, $f = 0$. The DF is

$$f = \left[\begin{array}{cc} \beta & 0 \\ 0 & 0 \end{array}\right]\left[\begin{array}{c} p_x \\ 0 \end{array}\right],$$

where $\beta = 230.77$ is sufficiently large such that the system with the external force is unstable.

B. Sensorimotor Control in Divergent Force Field

In the experiments of [22], [31], the human subject preforms a series of point-to-point arm-reaching movements in a horizontal plane. The starting point of the hand is 0.25m away from the target point which is marked as a red dot in Fig. 2. For each trial, it is considered successful if the human subject reaches the target within 0.6 ± 0.1s. The human subject first practices in NF until she can successfully complete the task several times. Then, DF is activated without notifying the human subject. The subject practices in DF until successfully reaching the target. After a short break, the subject is asked to perform several arm-reaching tasks in NF. These trials after the break are called after effects to confirm that the human subject has adapted to the force field. In the experiments of [22], [31], when initially exposed to DF, the hand trajectory is drastically distorted. After practicing several trials in DF, the human subject can reach the target along a relatively straight line.

First, Algorithm 1 is implemented for the dynamic model (34) in NF. After practicing enough trials and collecting the input-state data, Algorithm 1 starts to find the approximate risk-sensitive optimal control gain. Next, the scenario where the human subject is initially exposed to DF is simulated. Since the human subject is not notified when DF is activated, the human subject still uses the same control gain as in NF. After several trials, the human subject realizes that the force field is changed and starts to learn a new control gain in DF.

The trajectories, velocity and endpoint force profiles are shown in Figures 2 and 3. After practicing in NF, the subject can successfully reach the target along a relatively straight line. However, when the force field is changed to DF without notifying the subject, the hand of the subject touches the border of the horizontal plane, and unstable behavior happens. After the learning process by Algorithm 1 in DF, the human subject can successfully reach the target along a relatively straight line again. Finally, the after-effect trials are conducted. Compared with the initial trials in NF, the after-effect trajectories are straighter. This is because the human subject still uses the control gain learned in DF with a higher stiffness along the x-direction.

In Fig. 4, we see that the learned control gain $\hat{K}(s)$ approaches the optimal control gain as the iteration proceeds. Finally, the relative error of $\left\|\hat{K}_{DF}(s_f) - K^*_{DF}\right\|_F / \|K^*_{DF}\|_F$ is 0.8%, which indicates that the human subject can still find a near-optimal control gain even using noisy sensory data. By comparing the numerical results in Figures 2 and 3 with the experimental results in [22], [31], one can find that Algorithm 1 can reproduce the human motor control and learning in reality. Consequently, the risk-sensitive computational mechanism in Algorithm 1 provides a new perspective to explain human motor control and learning.

IV. Conclusion

In this paper, based on robust RL, we have proposed a novel computational mechanism to model the risk-sensitive optimality of human behavior in motor learning, control and adaptation. The proposed computational mechanism suggests that the CNS can still find an approximation of the risk-sensitive optimal controller, despite the errors in the learning process caused by unmeasurable sensory noise. We have conducted the simulation of the point-to-point reaching movements using the proposed computational mechanism. The simulated results of the hand trajectory, velocity, and acceleration profiles are compatible with the experimental results reported in [22], [31]. Hence, we argue that the CNS may apply the same mechanism as the robust RL to find the risk-sensitive optimal controller directly from data.

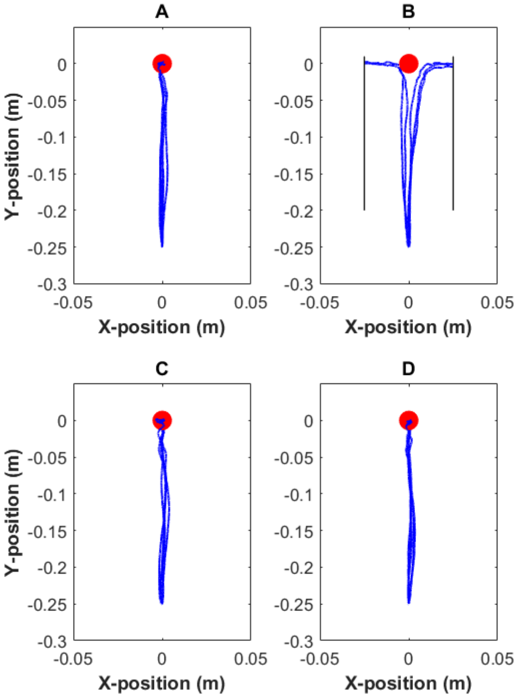

Fig. 2. Simulated movement trajectory generated by Algorithm 1. **A.** Five movement trajectories of the subject after learning in the NF. **B.** Five movement trajectories of the subject when it is initially exposed to the DF without notification. **C.** Five movement trajectories of the subject after learning in the DF. **D.** Five after-effect trials in the DF.

References

[1] P. Morasso, "Spatial control of arm movements," *Experimental Brain Research*, vol. 42, no. 2, pp. 223–227, 1981.

[2] E. Todorov and M. I. Jordan, "Optimal feedback control as a theory of motor coordination," *Nature Neuroscience*, vol. 5, no. 11, pp. 1226–1235, 2002.

[3] E. Todorov, "Optimality principles in sensorimotor control," *Nature Neuroscience*, vol. 7, p. pages907–915, 2004.

[4] D. Bernoulli, "Exposition of a new theory on the measurement of risk," *Econometrica*, vol. 22, no. 1, pp. 23–36, 1954.

[5] D. Jacobson, "Optimal stochastic linear systems with exponential performance criteria and their relation to deterministic differential games," *IEEE Trans. Autom. Control*, vol. 18, no. 2, pp. 124–131, 1973.

[6] T. Başar and P. Bernhard, *H_∞-Optimal Control and Related Minimax Design Problems: A Dynamic Game Approach*. Springer, 2008.

[7] L. Cui, T. Başar, and Z. P. Jiang, "A reinforcement learning look at risk-sensitive linear quadratic gaussian control," in *Proceedings of The 5th Annual Learning for Dynamics and Control Conference* (N. Matni, M. Morari, and G. J. Pappas, eds.), vol. 211 of *Proceedings of Machine Learning Research*, pp. 534–546, PMLR, 15–16 Jun 2023.

[8] A. J. Nagengast, D. A. Braun, and D. M. Wolpert, "Risk-sensitive optimal feedback control accounts for sensorimotor behavior under uncertainty," *PLOS Computational Biology*, vol. 6, pp. 1–15, 07 2010.

[9] D. A. Braun, A. J. Nagengast, and D. M. Wolpert, "Risk-sensitivity in sensorimotor control," *Frontiers in Human Neuroscience*, vol. 5, pp. 1–9, 2011.

[10] S.-W. Wu, J. Trommershäuser, L. T. Maloney, and M. S. Landy, "Limits to human movement planning in tasks with asymmetric gain landscapes," *Journal of Vision*, vol. 6, pp. 5–5, 01 2006.

[11] M. K. O'Brien and A. A. Ahmed, "Does risk-sensitivity transfer across movements?," *Journal of Neurophysiology*, vol. 109, no. 7, pp. 1866–1875, 2013. PMID: 23324319.

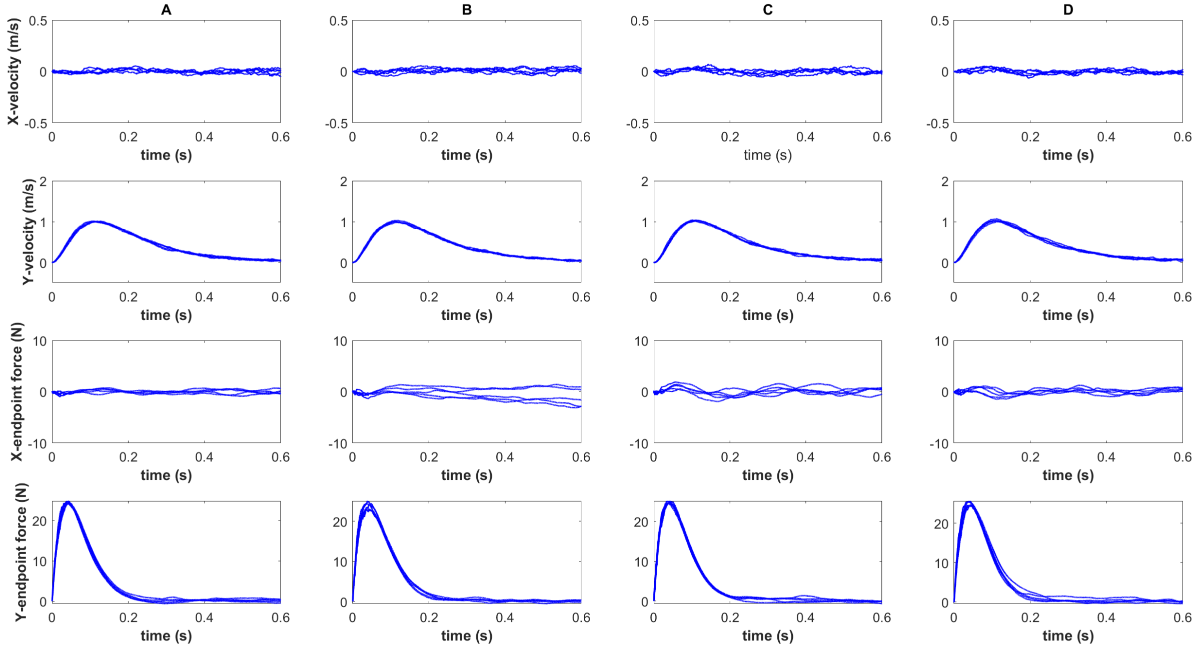

Fig. 3. Simulated velocity and force profiles generated by Algorithm 1. **A.** Simulated velocity and endpoint force profiles of the subject after learning in NF. **B.** Simulated velocity and force profiles when the subject is initially exposed to DF without notification. **C.** Simulated velocity and force profiles of the subject after learning in DF. **D.** After-effect trials in NF. The numerical results resembles the experimental results reported in [22].

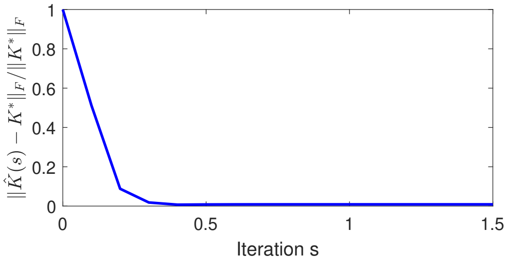

Fig. 4. Relative error between the learned control gain $\hat{K}(s)$ and the optimal control gain K^* in DF.

[12] Y. Ueyama, "Mini-max feedback control as a computational theory of sensorimotor control in the presence of structural uncertainty," *Frontiers in Computational Neuroscience*, vol. 8, pp. 1–14, 2014.

[13] F. Crevecoeur, S. H. Scott, and T. Cluff, "Robust control in human reaching movements: a model-free strategy to compensate for unpredictable disturbances," *Journal of Neuroscience*, vol. 39, no. 41, pp. 8135–8148, 2019.

[14] D. M. Wolpert, Z. Ghahramani, and M. I. Jordan, "An internal model for sensorimotor integration.," *Science*, vol. 269, pp. 1880–1882, 1995.

[15] Z. P. Jiang, T. Bian, and W. Gao, "Learning-based control: A tutorial and some recent results," *Foundations and Trends in Systems and Control*, pp. 176–284, 2020.

[16] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, Massachusetts: MIT Press, 2nd ed., 2018.

[17] Y. Jiang and Z. P. Jiang, "Adaptive dynamic programming as a theory of sensorimotor control," *Biological Cybernetics*, vol. 108, no. 4, pp. 459–473, 2014.

[18] T. Bian, D. M. Wolpert, and Z. P. Jiang, "Model-free robust optimal feedback mechanisms of biological motor control," *Neural Computation*, vol. 32, no. 3, pp. 562–595, 2020.

[19] B. Pang, L. Cui, and Z. P. Jiang, "Human motor learning is robust to control-dependent noise," *Biological Cybernetics*, vol. 116, no. 3, pp. 307–325, 2022.

[20] Y. Jiang and Z. P. Jiang, "A robust adaptive dynamic programming principle for sensorimotor control with signal-dependent noise," *Journal of Systems Science and Complexity*, vol. 28, p. 261–288, 2015.

[21] E. Burdet, R. Osu, D. W. Franklin, T. E. Milner, and M. Kawato, "The central nervous system stabilizes unstable dynamics by learning optimal impedance," *Nature*, vol. 414, no. 6862, pp. 446–449, 2001.

[22] D. W. Franklin, E. Burdet, R. Osu, M. Kawato, and T. E. Milner, "Functional significance of stiffness in adaptation of multijoint arm movements to stable and unstable dynamics.," *Experimental Brain Research*, vol. 151, no. 2, pp. 145 – 157, 2003.

[23] R. A. Schmidt, H. Zelaznik, B. Hawkins, J. S. Frank, and J. J. T. Quinn, "Motor-output variability: A theory for the accuracy of rapid motor acts," *Psychological Review*, vol. 86, no. 5, pp. 415–451, 1979.

[24] J. Willems, "Least squares stationary optimal control and the algebraic riccati equation," *IEEE Trans. Autom. Control*, vol. 16, no. 6, pp. 621–634, 1971.

[25] G. Freiling, G. Jank, and H. Abou-Kandil, "Generalized riccati difference and differential equations," *Linear Algebra and its Applications*, vol. 241-243, pp. 291–303, 1996. Proceedings of the Fourth Conference of the International Linear Algebra Society.

[26] S.-D. Wang, T.-S. Kuo, and C.-F. Hsu, "Trace bounds on the solution of the algebraic matrix riccati and lyapunov equation," *IEEE Transactions on Automatic Control*, vol. 31, no. 7, pp. 654–656, 1986.

[27] G. A. Pavliotis, *Stochastic Processes and Applications*. Springer, 2014.

[28] T. Liu, L. Cui, B. Pang, and Z. P. Jiang, "A unified framework for data-driven optimal control of connected vehicles in mixed traffic," *IEEE Transactions on Intelligent Vehicles*, pp. 1–15, 2023.

[29] L. Cui, B. Pang, and Z. P. Jiang, "Learning-based adaptive optimal control of linear time-delay systems: A policy iteration approach," *IEEE Transactions on Automatic Control*, pp. 1–8, 2023.

[30] D. Liu and E. Todorov, "Evidence for the flexible sensorimotor strategies predicted by optimal feedback control," *Journal of Neuroscience*, vol. 27, no. 35, pp. 9354–9368, 2007.

[31] E. Burdet, K. P. Tee, I. Mareels, T. E. Milner, C. M. Chew, D. W. Franklin, R. Osu, and M. Kawato, "Stability and motor adaptation in human arm movements," *Biological Cybernetics*, vol. 94, no. 1, pp. 20–32, 2006.