Applied Intelligence (2019) 49:1489–1501
https://doi.org/10.1007/s10489-018-1331-y

Inverse discounted-based LQR algorithm for learning human movement behaviors

Haitham El-Hussieny[1] · Jee-Hwan Ryu[2]

Published online: 14 November 2018

Abstract
Recently, there has been an increasing interest towards understanding human movement behaviors. In this regard, one of the approaches is to retrieve the unknown underlying objective function that the human has to optimize while achieving a certain movement behavior. Existing research of behavioral understanding merely depends on predefined optimality criteria, where the minimum time, minimum variance or/and minimum effort are mainly adopted. These criteria are assumed to be constant, where the human is assumed to have the same preferences during the movement duration. However, in this paper, the optimality criteria underlying the kinematic characteristics of a certain human behavior are assumed to be exponentially discounted to account for the change in the human preferences that could happen while achieving this behavior. A new Inverse Discounted-based Linear Quadratic Regulator (ID-LQR) algorithm is developed in the light of Inverse Optimal Control (IOC) framework to find out the discounted cost function that could reproduce the measured human behavior perfectly. Meanwhile, an Incremental version of the ID-LQR algorithm is proposed to continuously refine the so far learned cost function in the case of sequentially presented demonstrations. The saccadic eye gaze movement is studied as an example to quantify both the proposed ID-LQR and Inverse ID-LQR approaches. Simulation results are encouraging and show that the saccadic trajectories generated by ID-LQR approach match the experimental data in many aspects, including position and velocity profiles of saccades. Moreover, when it is assessed by a subsequent set of scenarios, the incremental ID-LQR algorithm confirms its capability to generalize the so far retrieved cost function for the unseen saccadic demonstrations.

Keywords Inverse optimal control · Learning by demonstrations · Imitation learning · Behavior modeling

1 Introduction

A growing body of neuroscience literature realizes the value of modeling human motor behaviors in more than one aspect. Particularly, in Learning from Demonstration (LfD) [5], learning of human motions from a set of given demonstrations could allow an artificial agent to own biologically-inspired skills in unstructured and challenging environments [41]. Additionally, in biomedical research, a mathematical formulation of a certain biological movement provides new insights into the assessment of action quality for healthy and disordered subjects [1, 30]. Furthermore, understanding the connection between human perception and action will help in predicting the human motions beforehand (i.e. intent prediction) to facilitate a human-guided interaction [3, 22].

From an engineering perspective, biological movements could be interpreted as a feedback control system in which the Central Nervous System (CNS) processes sensory inputs and provides commands to the human muscles in a closed loop fashion [20]. It is logically assumed that this human sensory-motor coupling is achieved in an optimal way, complying with the well-known principle of optimality [43]. This assumption facilitates neuro-science researchers, with the help of optimal control theory, in obtaining the law that governs complex human behaviors. Since the cost/reward function is considered as the most concise representation of human behaviors [29], nowadays, understanding human

✉ Haitham El-Hussieny
haitham.elhussieny@feng.bu.edu.eg

Jee-Hwan Ryu
jhryu@kut.ac.kr

[1] Electrical Engineering Department, Faculty of Engineering (Shoubra), Benha University, Benha, Egypt

[2] School of Mechanical Engineering, Korea University of Technology and Education (KOREATECH), Cheonan City, South Korea

movements has been done in the light of Inverse Optimal Control (IOC) framework [33, 36, 45], which is also known as Inverse Reinforcement Learning (IRL). In this regard, the main question to be addressed is, given measurements of a certain human motor behavior, what is the underlying optimality criteria that best describes the human performance of that behavior. In the conventional Optimal Control (OC)/Reinforcement Learning (RL) approach [24], the cost/reward function is usually assumed and the goal is to find out the optimal policy or behavior that an artificial agent should follow to achieve the required task while satisfying the given optimality criteria. In contrary, by having some human demonstrations of a certain movement, the IOC/IRL aims to automate the task of retrieving the optimality criteria that best describe these given demonstrations.

In recent IOC/IRL studies, the unknown cost/reward function underlying a certain behavior has assumed to be constant while achieving that behavior. However, unsatisfactory results could be obtained under this assumption if the motion under discussion is short, fast and jerky such as in saccadic eye-gaze behavior. In such movements, the estimated cost function has to show a significant penalization toward the motion features which in turn increases the magnitude of muscles activation beyond the logical limits. In this paper, the human preferences are assumed to be time-varying during the achievement of a certain behavior. In this regard, with the mentioned fast movements, the required cost function could be varying with time with no need to afford a significant penalization during the whole motion sequence. This key assumption is motivated by the well-known Q-learning approach [44], where long-term rewards are discounted based on their time of occurrence. Moreover, it has often been assumed that the human demonstrations are available at prior and learning is done as a one-shot process. Subsequently, the learned cost function is exclusively matching the available measurements with no guarantee that it can be adapted with new demonstrations for the same behavior. Meanwhile, having a large number of training measurements is not applicable due to memory and time constraints.

The main contributions of this paper are twofold:

- First, the optimality criteria behind a certain human movement behavior are recovered with the help of the IOC framework while the cost function holds an exponential discount factor. An Inverse Discount-based Linear Quadratic Regulator (ID-LQR) is proposed to retrieve both the weighting matrices and the discount factors that best describe the given behavior in terms of system states and inputs. The newly proposed ID-LQR is believed to aid the retrieving of the cost function for not only normal movements but also for ballistic, jerky and fast movement behaviors without affecting the precision compared to existing approaches found in [10] and [34]. Without loss of generality, the eye gaze saccadic movement behavior is taken as an example of a ballistic and jerky movement to be modeled with the proposed ID-LQR approach.
- Second, the problem of incremental learning of the cost function is addressed by mimicking the natural way of human beings in acquiring knowledge over time. Once new information becomes available, the learned knowledge is revised (i.e. evolved) to generalize for the recently available information. To date, the incremental IOC has received a scant attention in research in spite of its importance in overcoming the problem of overfitting [26].

The rest of this paper is organized as follows, the related work is discussed in Section 2. In Section 3, the discount-based LQR problem is briefly reviewed while the proposed ID-LQR algorithm is described in details in Section 4. The model of the saccadic gaze shift is presented and discussed in Section 5. Results are highlighted and discussed in Section 6. Finally, the conclusion is given in Section 7 with remarks on the future implications.

2 Related work

Recently, there has been an extensive research on inverse optimal control. In this section, a general overview of the available IOC algorithms is highlighted while the approaches that are most related to the modeling of human movement behavior are briefly discussed. The reader is suggested to follow the survey paper of Zhifei and Joo [46] for a detailed overview of the IOC framework.

In literature, several studies have addressed the problem of retrieving the objective function behind a particular human motor movement behavior. For instance, in [42], this problem is tackled from the imitation prospective, i.e. the reproduction of the measured demonstrations over a robotic agent with specific kinematic and dynamic ranges. However, unsatisfactory results were obtained in reproducing the same behavior within a different unstructured or dynamic environments [4]. Other studies, such as [9] assumed a heuristic cost function that human aims to minimize while performing a certain behavior. This cost function was selected by intuition as a single feature that covers the human behavior.

Recently, IOC has been successfully used in literature to retrieve the *cost* function that best describes the given demonstrated sequence(s) of human actions. The original IOC algorithms [2, 29] were introduced assuming that the unknown cost is a parametric function that is a combination of a set of selected features. Although the

effectiveness of this assumption has been proved by some applications, it could be wrong when the human behaves according to another form of reward function. Thus, nonlinear cost functions were introduced in LEArning and seaRCH (LEARCH) algorithm [37]. In Bayesian IOC [35], a probability distribution was introduced to model the uncertainty that could exist in the obtained cost function. It considers the human demonstration as evidence that a prior on the cost function has to be found out. Recent work suggests using deep neural networks to learn the cost function in inverse reinforcement learning to address the need for representative features that compromise the unknown cost function as in [14, 31].

The IOC framework has been used so far in the context of understanding human behaviors. In [27], the human locomotion trajectory has been studied to figure out how humans move from one point to another endpoint. The inverse problem of the optimal control theory was adapted and solved by Bound Optimization BY Quadratic Approximation (BOBYQA) technique [32]. In [13], the cost functions of human manipulation task were modeled from given demonstrations while considering contact with the environment. The inverse Karush-Kuhn-Tucker (KKT) inverse optimal control algorithm was introduced to learn the cost function of manipulation task with the contact constraints. The nonlinear dynamics of the behavior under discussion were approximated around the nominal points and the IOC problem was solved recursively. The computational demand of those approaches motivates the use of Linear Quadratic Regulation (LQR) problem to model the human movement behavior. The Inverse LQR algorithm [34] has been used for modeling a range of human behaviors assuming a quadratic cost function in terms of the motion states and effort with linear movement dynamics. For instance, the reach-to-grasp behavior is modeled in [10] while the sit-to-stand cost function is learned in [11].

In the previously mentioned IOC algorithms, the unknown cost functions were assumed to be time-invariant while achieving the whole movement. However, in modeling of fast movements, such as the saccadic eye-gaze behavior, this assumption will lead to significant penalization factors in the learned cost function which opposes the limitations of muscle activation inputs. Thus, in this paper we focused on solving this issue taking the saccadic eye movement as a test case to learn the optimality criteria behind this kind of fast movement. In early studies of the saccadic eye behavior, the optimality criteria underlying the kinematics of saccades have often been expressed as a linear combination of hand-coded features. For instance, the minimum-time principle is proposed in [12] where the cost function of saccades is assumed to be the minimization of the time required to foveate a target. However, this assumption has failed in generating velocity profiles that comply with the actual biological data [18]. Thus, in [19], in addition to the minimum time, the minimum-variance principle is proposed to obey the existence of the additive white noise in the oculomotor neurons. By combining these two optimality principles, reasonable biological gaze profiles are generated based on the trade-off between reaching the target accurately vs. reaching it quickly. Recently, the minimum-effort principle has been discussed taking into account the minimization of oculomotor torques during the saccadic movements [21]. In this paper, the general optimality principle of the saccadic eye-gaze movement has been studied taking into account the jerky nature of that motion with the proposed inverse discounted-based LQR approach as will be discussed.

3 Discounted-based linear quadratic regulator

The main concern of optimal control theory is to find out the control signals that derive a given system to the desired state while satisfying certain optimality criteria [23]. Linear Quadratic Regulator (LQR) is a variant of optimal control, where the system dynamics are modeled by a set of linear differential equations. In addition, the optimality criteria are described by a quadratic function that comprises the system states and inputs.

Given a continuous-time linear system,

$$\begin{aligned} \dot{\boldsymbol{x}}(t) &= Ax(t) + Bu(t), \qquad x(0) = x_0 \\ y(t) &= Cx(t) + Du(t), \end{aligned} \tag{1}$$

where A, B, C and D are the system matrices, and $\boldsymbol{x} \in \mathbb{R}^n$, $\boldsymbol{u} \in \mathbb{R}^m$ and $\boldsymbol{y} \in \mathbb{R}^p$ are the state vector, control inputs and the measured outputs respectively. The discounted-based LQR aims to control the system in (1) such as to minimize the quadratic cost function

$$J = \left\{ \int_0^\infty e^{2\alpha t} \left(x(t)^T Q x(t) + u(t)^T R u(t) \right) dt \right\} \tag{2}$$

where $Q \in \mathbb{R}^{n \times n}$ and $R \in \mathbb{R}^{p \times p}$ are the state and input weighting matrices with $Q = Q^T \succeq 0$ and $R = R^T \succ 0$, where $\succeq$ and $\succ$ denote semi-positive definite and positive definite matrices.

The exponential term $e^{2\alpha t}$ in the cost function (2) is called the discount term or the prescribed degree of stability which generalizes the standard form of the original, i.e. $\alpha = 0$, LQR problem [25]. The scalar value α determines how much the cost function in (2) is discounted over the time horizon. That is, for some discount factor $\alpha \neq 0$, the states $\boldsymbol{x}(t)$ approach zero at least as fast as $e^{2\alpha t}$. By introducing the discount term in (2), it seems to be a time-varying LQR problem. However, the same procedures of solving the ordinary time-invariant LQR problem could be

invoked by replacing A with $A_\alpha = A + \alpha I$ [6]. This, in fact, ensures that the eigenvalues of the closed-loop system are smaller than $-\alpha$. Particularly, positive values of α result in more faster controllers with higher gains to drive the system to the desired states. In opposite, negative values of α result in more slower controllers with lower gains.

Given (A_α, B) is controllable and (A_α, C) is detectable, the optimal control input which minimizes the deterministic version of the cost function (2) is given by:

$$u(t) = -K_\alpha x(t) \tag{3}$$

where,

$$K_\alpha = R^{-1} B^T P_\alpha \tag{4}$$

in which P_α is obtained by solving the following Continuous-time Algebraic Riccati Equation (CARE):

$$A_\alpha^T P_\alpha + P_\alpha A_\alpha - P_\alpha B R^{-1} B^T P_\alpha + Q = 0 \tag{5}$$

In summary, the discounted-base LQR problem aims to satisfy the quadratic cost function given in (2), given the system dynamics, the set of weighting matrices Q and R besides the value of the discount factor α.

4 Inverse discounted-based LQR approach

In this research, the proposed Inverse Discounted-based LQR (ID-LQR) aims to find the cost function for which a certain demonstrated human behavior is optimal. Definitely, we are interested in retrieving the objective function defined in (2) that best describes that behavior. Assume that the system matrices in (1) are known, finding the set of weighting matrices Q and R alongside the discount factor α is our key objective to model the human behavior under discussion.

In (2), the weighting matrix Q penalizes the deviation of each state from its desired value, while R penalizes the control efforts $\boldsymbol{u}(t)$ toward minimum values. Thus, in this context, the optimality criteria underlying a certain biological behavior could be relatively explained in the light of these two weighting matrices. In addition, finding the exponent factor α in (2), could perhaps explain the degree of aggressiveness the controller has toward driving states to their desired values.

In this work, we have replaced the scalar discount exponent α by a discount diagonal matrix $\boldsymbol{\alpha} \in \mathbb{R}^{n\times n}$, where each element α_i on the main diagonal is an individual discount factor for its associated state. Thus, additional tuning parameters could be facilitated in understanding the given human behavior. In fact, the retrieved discount matrix affords knowledge about the degree of aggressiveness the controller has for each state as mentioned. The same procedures are hold for the scalar discount exponent where $A_{\boldsymbol{\alpha}} = A + \boldsymbol{\alpha}$. To mention, the term "weighting matrices" refers to Q, R and $\boldsymbol{\alpha}$ in this research.

Figure 1 illustrates the key idea behind how ID-LQR works to model human movements in general. At first, the human behavior under discussion is measured and assigned as the given optimal trajectory. Subsequently, by generating candidate weighting matrices, the ID-LQR induces a simulated response in favor to match that given optimal trajectory. This process is continuously repeated until finding the response that best matches the given demonstrated trajectory. At that time, the candidate weighting matrices are identified as the final weights that comprise the optimality criteria underlying the given human behavior. Mathematically, the ID-LQR problem is formulated as follows, given k measurements of a demonstrated human behavior $\boldsymbol{y}_h(k) \in \mathbb{R}^{p\times k}$ alongside the system defined in (1), the goal is to estimate the set of weighting matrices $\hat{Q}$, $\hat{R}$ and $\hat{\boldsymbol{\alpha}}$ that could reproduce the given behavior perfectly. The ID-LQR problem is then formulated as

$$\begin{aligned} \{\hat{Q}, \hat{R}, \hat{\boldsymbol{\alpha}}\} &= \min_{Q,R,\boldsymbol{\alpha}} \|y_h - \hat{y}(Q, R, \boldsymbol{\alpha})\|_F^2, \quad \text{such that} \\ &\hat{Q} \succeq 0, \\ &\hat{R} \succ 0, \\ \lambda_{max} &\left(\begin{bmatrix} \hat{Q} & \mathbf{0} \\ \mathbf{0} & \hat{R} \end{bmatrix} \right) \leq \gamma \\ \lambda_{min} &\left(\begin{bmatrix} \hat{Q} & \mathbf{0} \\ \mathbf{0} & \hat{R} \end{bmatrix} \right) \geq 1 \end{aligned} \tag{6}$$

where $\| \bullet \|_F$ denotes the Frobenius norm, i.e., $\|A\|_F := \sqrt{trace(A^T A)}$ (for real A) and $\hat{y}(Q, R, \boldsymbol{\alpha})$ is the simulated response for a certain guess of the weighting matrices. The first two constraints are to ensure that the generated $\hat{Q}$ and $\hat{R}$ are complying with the LQR problem mentioned

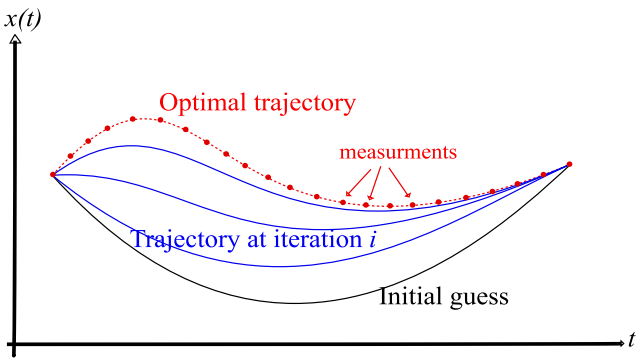


Fig. 1 Generation of the candidate responses in solid line by the ID-LQR to minimize the discrepancy to the given optimal behavior in dotted line

before. On the other hand, the last two constraints are to keep maximum and the minimum eigenvalues of the block matrix $\begin{bmatrix} \hat{Q} & \mathbf{0} \\ \mathbf{0} & \hat{R} \end{bmatrix}$ within the range $[1, \gamma]$. This, in turn, ensures having a minimum condition number of the generated weighting matrices and subsequently increases the precision of arithmetic operations involving $\hat{Q}$ and $\hat{R}$. An off-the-shelf Genetic Algorithm (GA) technique [8] is used to solve the optimization problem of the proposed ID-LQR.

The required matrix $\hat{Q}$ is formed as $\hat{Q} = G + G_u^T$ to obtain a symmetric matrix, where $G = (g_{ij})\ \forall i \geq j$ is an upper triangular matrix with $n(n+1)/2$ elements, while G_u is a matrix formed with the elements of G above the main diagonal. Moreover, the matrix $\hat{R}$ is formulated as a diagonal matrix, i.e. $R = (r_{ij})\ \forall i = j$. Finally, the discount matrix $\hat{\boldsymbol{\alpha}}$ is also formulated as a diagonal matrix where each diagonal element α_i describes the degree of aggressiveness the controller affords to the state x_i. All these elements of the weighing matrices are appended as a one-dimensional vector $\boldsymbol{\theta}$ which serves as the population for the ID-LQR optimization problem,

$$\boldsymbol{\theta} := [g_{11}, g_{12}, \ldots g_{nn}, r_1, r_2, \ldots, r_m, \alpha_1, \ldots, \alpha_n] \tag{7}$$

Algorithm 1 Inverse Discounted-based LQR algorithm

Input:

System dynamics: A, B, C, D and $\boldsymbol{x}_0$.
Human behavior demonstration: $\boldsymbol{y}_h(k) \in \mathbb{R}^{p \times k}$.
Threshold of convergence: ϵ.

Output:

The weighting matrices, $\hat{Q}$, $\hat{R}$ and $\hat{\boldsymbol{\alpha}}$.

1: $\boldsymbol{\theta}^{(0)} \leftarrow random()$ ▷ form an initial random guess
2: $i \leftarrow 0$ ▷ initialize iterations counter
3: **do**
4: $\quad i \leftarrow i + 1$ ▷ increment
5: $\quad \boldsymbol{\theta}^{(i)} \leftarrow \mathbf{ga}(\boldsymbol{\theta}^{(i-1)})$ ▷ hypothesized solution at iteration i based on GA
6: $\quad \{\hat{Q}, \hat{R}, \hat{\boldsymbol{\alpha}}\} \leftarrow$ formulateMatrices$(\boldsymbol{\theta}^{(i)})$ ▷ formulate the weighting matrices
7: $\quad \hat{K} \leftarrow$ lqr$(A_{\boldsymbol{\alpha}}, B, \hat{Q}, \hat{R})$ ▷ solve the LQR problem at iteration i
8: $\quad \hat{y}(i) \leftarrow$ simulate$(\hat{K}, A, B, C, D, x_0)$ ▷ simulated response at iteration i
9: $\quad F_i \leftarrow \|\boldsymbol{y}_h - \hat{\boldsymbol{y}}(i)\|_F^2$ ▷ Frobineus norm value as a fitness function
10: $\quad \Delta F_i = F_i - F_{i-1}$ ▷ find change in Frobineus norms
11: **while** $\Delta F_i > \epsilon$ ▷ no convergence achieved
12: $\{Q, R, \boldsymbol{\alpha}\} \leftarrow$ formulateMatrices$(\boldsymbol{\theta}^{(i)})$ ▷ construct the final weighting matrices
13: **return** $Q, R, \boldsymbol{\alpha}$ ▷ return the final weighting matrices

Algorithm 1 explains the procedures of the proposed ID-LQR approach to retrieve the weighing matrices Q, R and $\boldsymbol{\alpha}$ that explain a given human behavior measurements, $\boldsymbol{y}_h(k)$. In line 1, an initial guess $\boldsymbol{\theta}^{(0)}$ is randomly generated that serves as the first GA population. Afterwards, at iteration i, the candidate weighing matrices $\hat{Q}$, $\hat{R}$ and $\hat{\alpha}$ are back formulated from the population vector $\boldsymbol{\theta}^{(i)}$ using the function ***formulateMatrices()*** as depicted in lines 6. Henceforth, in line 7 the discounted-based LQR problem is solved and the state feedback gain $\hat{K}_i$ is found by solving the CARE mentioned before in (4). In lines 8, the estimated response $\hat{y}$ is obtained by simulating the closed loop system dynamics with $\hat{K}_i$ and the initial condition $\boldsymbol{x}_0$ using the function ***simulate()***. The Frobenius norm F_i of the error between the estimated and the demonstrated behavior is calculated in line 9. The change in Frobenius norms, ΔF_i is measured as an indication for the convergence of the solution. The algorithm is repeated again if this change is larger than the convergence threshold ϵ. At that point, another candidate vector $\boldsymbol{\theta}$ is generated using the Genetic Algorithm optimization based on its previous estimate starting from line 5. If ΔF_i is less than ϵ, the algorithm stops and the final candidates of the weighting matrices are formulated and returned as highlighted in lines 12-13.

4.1 Incremental ID-LQR algorithm

In machine learning, building a model from few training examples could perhaps lead to the well-known over-fitting problem. In such cases, the learned parameters are perfectly fitting with the few given examples, however, unsatisfactory results could be obtained with the remaining unseen or test

examples. The same problem could occur in our case, where the cost function obtained from few demonstrations are most likely be fitting with these demonstrations only. In the other side, learning from multiple demonstrations has two issues. First, how to mathematically combine these multiple demonstrations that could differ in lengths, sampling rates or even the initial conditions. Second, even though the first issue is tackled by wrapping the demonstrations non-linearly in time-space [17], how to respond if a new demonstration has appeared after learning the cost function from the past demonstrations. One solution is to always keep the previous measurements and append the new ones with them. Diffidently such a solution is not a practical one since it requires more storage and an exhaustive computations.

In this research, such issues have been tackled by introducing the incremental ID-LQR algorithm that could adapt to the newly coming scenarios. The weighing matrices, $Q^{(s-1)}$, $R^{(s-1)}$ and $\boldsymbol{\alpha}^{(s-1)}$ that are obtained from the previous scenario $s-1$ are refined once a new demonstration $\boldsymbol{y}_h^{(s)}$ exists at s scenario. This kind of refinement is done without the need to either keep previous demonstrations or even having demonstrations with similar characteristics.

Suppose that at a previous scenario $s-1$, the learned weighing matrices $Q^{(s-1)}$, $R^{(s-1)}$ and $\boldsymbol{\alpha}^{(s-1)}$ are retrieved from the given demonstration $\boldsymbol{y}^{(s-1)_h}$ with an initial condition $\boldsymbol{x}_0^{(s-1)}$ using the ID-LQR mentioned in the previous section. If another demonstration $\boldsymbol{y}_h^{(s)}$ is available at the current scenario s, the so far learned weighing matrices are altered in a way to generalize with the new demonstration that could have a different initial condition $\boldsymbol{x}_0^{(s)}$ and number of samples. Since the biological motor movements exhibit invariant features [40], the weighting matrices should be further generalized with the human optimality bounds after being refined with the newly available demonstrations. The key idea to achieve this adaptation is to mimic the same mechanism of applying the moving average technique over subsequent data as follows,

$$\left.\begin{aligned} Q^{(s)} &= Q^{(s-1)} + \frac{Q_s - Q^{(s-1)}}{s+1} \\ R^{(s)} &= R^{(s-1)} + \frac{R_s - R^{(s-1)}}{s+1} \\ \boldsymbol{\alpha}^{(s)} &= \boldsymbol{\alpha}^{(s-1)} + \frac{\boldsymbol{\alpha}_s - \boldsymbol{\alpha}^{(s-1)}}{s+1} \end{aligned}\right\} \forall\, s > 0 \tag{8}$$

where, for instance, $Q^{(s)}$ is the adapted state weighing matrix after having the new demonstration $\boldsymbol{y}_h^{(s)}$ at scenario s, while Q_s is the weighting matrix that is currently obtained by applying the ID-LQR algorithm over the $\boldsymbol{y}_h^{(s)}$ measurement only. The same applies for both the input matrix $R^{(s)}$ and $\boldsymbol{\alpha}^{(s)}$. Thus, instead of applying the arithmetic mean over the subsequent demonstrations we get, the averaging is performed over the retrieved weighting matrices without caring about the dissimilarity that could exist between the demonstrations.

One thing to mention is, there exists an ambiguity in solving the inverse of the LQR problem. If we have certain matrices, Q and R, and then are multiplied by a scalar $\eta > 0$, we will get the same response, no matter what the value of η is. This, in turn, could complicate the calculation of the average matrices mention in (8). To deal with this issue, a prior normalization step is performed by the incremental ID-LQR algorithm over Q_s and R_s to have a consistent scale. This normalization perhaps will not affect the relationship between the states and the inputs, and the given behavior could still be explained in the light of them. In doing so, the min-max normalization is used with the aid of the minimum and the maximum eigenvalues of the block diagonal matrix QR as follows

$$\begin{aligned} Q_s &= \frac{Q_s - \lambda_{min}(QR)}{\lambda_{max}(QR) - \lambda_{min}(QR)} \\ R_s &= \frac{R_s - \lambda_{min}(QR)}{\lambda_{max}(QR) - \lambda_{min}(QR)} \end{aligned} \tag{9}$$

5 Application to saccadic gaze movements

During the daily active perception tasks, humans are continuously directing their eyes gaze towards the targets of interest. Generally, this intentional gaze shift is accomplished by a coordinated movement of several parts in the human body; including eyes, head, torso and/or the whole body, depending on the amplitude of the intended gaze [39]. For instance, a target that is located within an amplitude of less than 10° from the visual axis can be reached with eye-only, head-restrained saccades. On the other side, a coordinated movement of both eyes and head is required to foveate a target that is located at larger amplitudes [16]. Modeling of such human saccadic gaze behavior is promising in the context of understanding the oculomotor neural architecture [15, 39]. Meanwhile, such kind of modeling facilitates human-like robots to imitate and convey a more naturally gaze cueing in Human-Robot Interaction (HRI) scenarios [28, 38].

The proposed ID-LQR approach is applied to understand the kinematic characteristics of the saccadic eye movements. Particularly, we are interested in finding the optimality principles behind these jerky movements that precede the eye fixations.

5.1 The model

Instead of using the mathematical model of the oculomotor plant, the eye gaze positions projected on a front screen in

the unit of pixels were used. The instantaneous gaze state $\mathbf{X}_t$ at time step t is defined as,

$$\mathbf{X}_t = [x_t \ y_t \ \dot{x}_t \ \dot{y}_t \ \ddot{x}_t \ \ddot{y}_t]^T$$

where the position, velocity, and acceleration of the eye gaze towards a predefined target are retrieved. Velocities and accelerations are found according to difference equations up to a scale factor,

$$\begin{pmatrix} \dot{x}_t \\ \dot{y}_t \end{pmatrix} = \begin{pmatrix} x_t - x_{t-1} \\ y_t - y_{t-1} \end{pmatrix} \tag{10}$$

$$\begin{pmatrix} \ddot{x}_t \\ \ddot{y}_t \end{pmatrix} = \begin{pmatrix} \dot{x}_t - \dot{x}_{t-1} \\ \dot{y}_t - \dot{y}_{t-1} \end{pmatrix} \tag{11}$$

Eye movement is expressed as a third-order linear dynamic model with thinking of control vector $\mathbf{u_t} = [\dddot{x}_t \ \dddot{y}_t]^T$ representing the jerk; the time derivative of the gaze acceleration. We assume the full states of the system are measured to be able to use the LQR problem. Under this dynamic model, the eye gaze movements over a front screen follow a linear relationship

$$\dot{\mathbf{X}}_t = A\mathbf{X}_t + B\mathbf{u}_t$$

$$\mathbf{Y}_t = C\mathbf{X}_t$$

where $A \in \mathbb{R}^{6\times6}$, $B \in \mathbb{R}^{6\times2}$ and $C = \mathbf{I}_6$ are the dynamics matrices defined as

$$A = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \qquad B = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}$$

5.2 Data collection and processing

To acquire the saccadic gaze movements, eight subjects were asked to point their eyes over two square objects that are consequently appeared on a 32-inch, 1920×1080 pixels, front screen. The first object was to draw and synchronize the attention of all subjects while the second is to stimulate the launch of the saccadic movement. The Unity3D[1] game engine is used to simulate the virtual environment including the mentioned two objects. An Eyetribe[2] eye tracker is placed under the front screen to retrieve horizontal and vertical gaze positions in pixels at a rate of 60 Hz. Subjects were sat at 60 cm away from the eye tracker and a prior calibration procedure is performed.

Subsequent processing steps are conducted over the retrieved eye-gaze trajectories before applied to the proposed ID-LQR as shown in Fig. 2. In the beginning, a scaling process is conducted to keep the human demonstrations within the allowable range for the control perspective by dividing the measurements by 10. Subsequently, the saccades movements are extracted from the whole eye movement b simply detecting abrupt changes in measured demonstrations since the saccades are known to be fast and abrupt. Then, the saccades trajectories are smoothed with a second order digital Butterworth filter (cut-off frequency of 4 Hz).

Since humans generally plan their gaze in a local coordinate system based on the current-to-target positions, it is much reasonable to project the original gaze trajectories into a new coordinate system that takes this into consideration. The origin of this new coordinates is located at the target position while the x-axis is aligned with the start-to-target direction and the y-axis is orthogonal. This makes the new trajectory goes from negative values to zero in x-axis while deviates up and down around zero in y-axis. Such kind of projection is consistent with the regulation term of the LQR problem, where the objective is to regulate the system, i.e. derives its states to zero. Meanwhile, it enriches the understanding of the saccades behavior by investigating the forward movement in the x-axis and the orthogonal deviations in the y-axis as depicted in Fig. 3. So, the aim of our formulated control problem is to bring both the forward and the orthogonal motions to zeros, i.e. reach the target in a defined straight line perfectly with the given initial condition. To note, the saccade demonstration is formulated as the two gaze positions obtained after preprocessing, appended in one columns, $\boldsymbol{y}_h(k) = [\boldsymbol{x}(k), \ \boldsymbol{y}(k)]^T$.

To summarize the preprocessing steps done over the eye gaze movements, the following steps are carried out sequentially:

- Extraction of the saccades trajectories from demonstrated behavior.
- Smoothing by applying a second-order Butterworth filter.
- Projection into the new target-oriented coordinate system.

6 Results and discussion

In this section, simulation results of applying the proposed the ID-LQR and the incremental ID-LQR approaches for modeling the saccadic gaze behavior are presented. At first, the cost function that best describes the saccadic gaze demonstrations is retrieved. Then, the incremental ID-LQR approach is applied to refine the so far retrieved cost function when new demonstrations are introduced with no need to retrain again. Finally, the proposed ID-LQR is

[1] https://unity3d.com

[2] https://theeyetribe.com

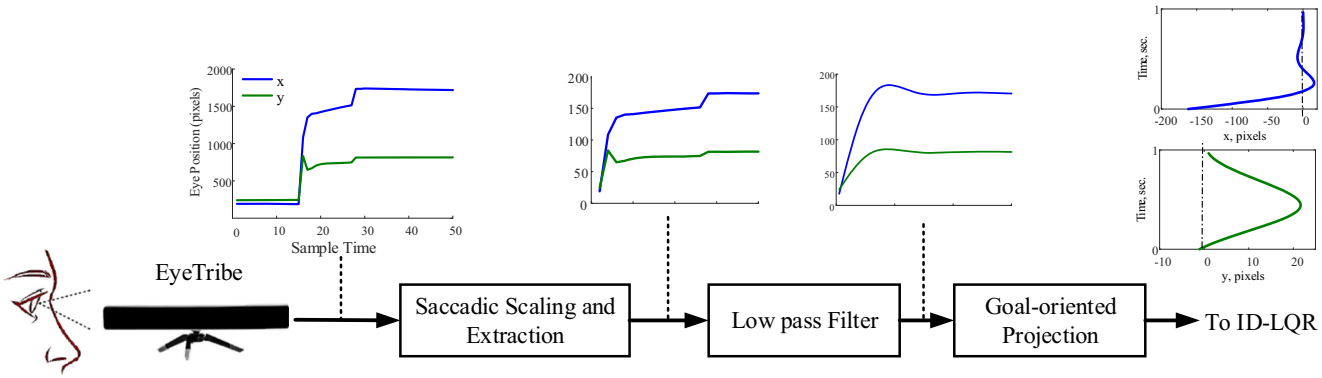


Fig. 2 A schematic showing the subsequent precessing steps conducted over the retrieved eye-gaze demonstrations before applied to the ID-LQR algorithm

compared with the state of the art non-discounted ILQR approach in terms of the accuracy of modeling the given saccadic gaze demonstrations.

6.1 Identification of cost function

The ID-LQR algorithm is applied to retrieve the weighting matrices of the eight demonstrated saccadic behaviors after they have processed. Each retrieved saccade trajectory is labeled as one scenario that is fed to the ID-LQR algorithm alongside the system matrices mentioned earlier. The pre-defined parameters for the algorithm are selected as follows: (i) the maximum eigenvalue constraint, $\gamma = 20$ and (ii) the threshold of convergence ($\epsilon = 1e^{-4}$).

The heat map of the learned cost function in terms of the weighting matrices is shown in Fig. 4 for one demonstration scenario. The Q matrix is on the top-left, while the R matrix is on the bottom-right. The diagonal elements of the discount matrix, $\boldsymbol{\alpha}$, is also shown. In this heat map, each element represents a cost coefficient related to a certain combined quadratic term.

Along the main diagonal (top-left to bottom-right) of the heat map, the cost values are all positive, penalizing state deviations in Q away from zero, and control inputs in R from having large values. For instance, the two top-left values of matrix Q are $xx = 6.93$ and $yy = 2.33$, which penalize the sate position of x and y respectively. As interpreted, during saccades, being away from a target position is much penalized compared to that of being deviated from the straight line connecting to it. In other words, humans show much care for reaching a target, i.e. minimize the $\boldsymbol{x}$ position from the x_0 to 0). Nevertheless, he/she is not caring much for accurately reaching this target in a straight line. Also, as depicted from the R matrix, reaching a goal target is of much interest than minimizing the control efforts in both U_x and U_y. Likewise, it is noted that the higher-order dynamics (i.e. accelerations) are massively penalized especially in the orthogonal direction. This primarily supports the requirement for stable movements by penalizing the growth dynamics.

Indeed, the off-diagonal terms of the weighting matrix Q indicate how each state in association with other states

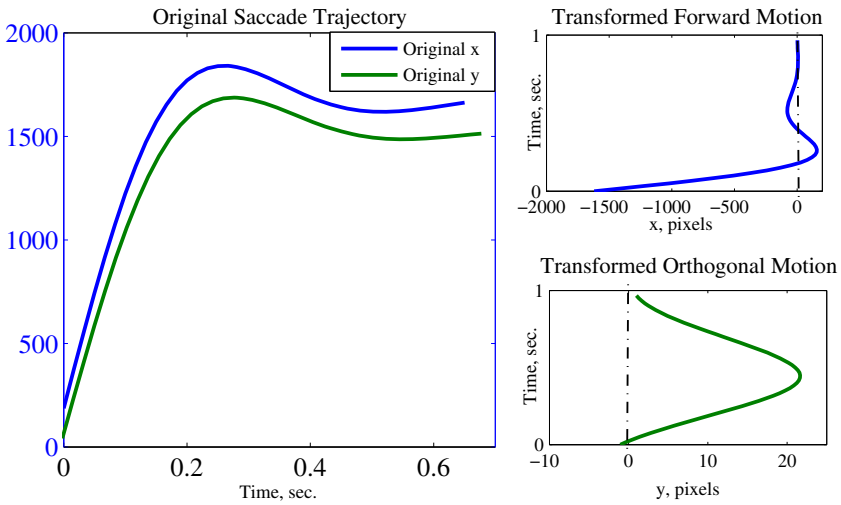


Fig. 3 An example of a captured saccade trajectory on the left and its transformation to forward and orthogonal motions on the right

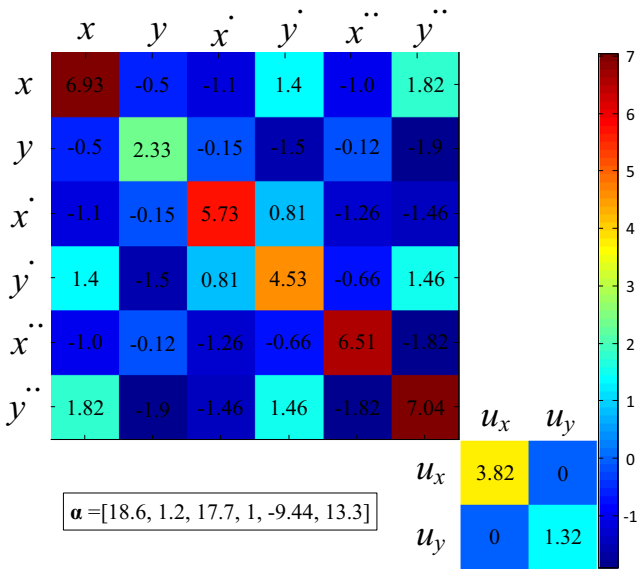


Fig. 4 A heat map for the obtained weighting matrices Q on the top-left and R on the bottom-right while the discount matrix α is emphasized in a box

is penalized towards zero, where positive values correspond to costs while negative correspond to rewards (or badly penalized). For example, the $x\dot{x}$ penalizes velocities in forward direction that are away from the target location ($x < 0$ and $\dot{x} < 0$ or $x > 0$ and $\dot{x} > 0$) with a coefficient of -1.1. This could be particularly interpreted as: at early stages of saccades (i.e. $\boldsymbol{x} < 0$), being away from the target velocity (i.e. $\dot{x} = 0$) is not preferred which is reasonable to reach that target. In the same philosophy, x and y are crossly penalized with a coefficient of $xy = -0.5$. Thus, the orthogonal deviation is "unintentionally" rewarded at early stages of saccades.

The retrieved discount matrix is

$$\boldsymbol{\alpha} = diag([18.6, 1.2, 17.7, 1, -9.44, 13.3])$$

where higher positive values for $\boldsymbol{x}$, $\dot{\boldsymbol{x}}$ and $\dot{\boldsymbol{y}}$ states show how far the associated closed loop poles are located compared to the other states. Definitely, the feedback controller during saccades is much faster in deriving $\boldsymbol{x}$, $\dot{\boldsymbol{x}}$ and $\dot{\boldsymbol{y}}$ to zero, while it acts slowly towards deriving $\boldsymbol{y}$ and $\dot{\boldsymbol{y}}$ to zero. The negative value of -9.44 indicates that the controller behaves in a lazy way towards deriving the acceleration $\ddot{\boldsymbol{x}}$ to zero. In some extent, this could be aligned with the speed-accuracy trade-off exists in biological movements [47]. Although this results could facilitate preliminary understanding for the saccadic behavior, a quantitative clinical study have to be performed to have much better scientific understanding.

6.2 Validation of incremental ID-LQR

To prove the adaptability of the proposed Incremental ID-LQR, two different learning models have been conducted. In the first, the ID-LQR learns the weighting matrices in an *amnesic* way, where at each given demonstration D_i, no prior knowledge is available about the learned weighting matrices learned from demonstration D_{i-1} as depicted in Fig. 5a. On contrary, in the second learning approach, the *incremental* learning is applied over the given demonstrations subsequently by invoking the proposed incremental ID-LQR explained above. Thus, in this model, the weighting matrices learned in demonstration D_i are based on what have learned so far from the previous demonstration D_{i-1} as highlighted in Fig. 5b. We have claimed that in the incremental ID-LQR approach, the learned cost function will be refined further with more coming demonstrations which in turn increase the robustness toward unseen scenarios.

To proceed, a cross-validation method was carried out in both learning models to show how amnesic and incremental approaches could perform over unseen scenarios. The Normalized Squared Errors, $NSE_{i,j}$ between the estimated trajectory $\hat{\boldsymbol{y}}_i$ and the actual demonstrations $\boldsymbol{y}_j$ are calculated

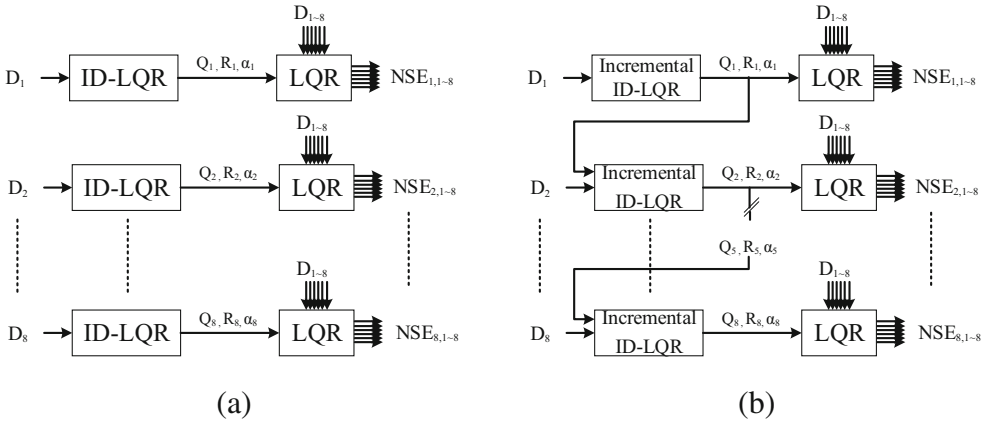


Fig. 5 A Schematic showing the difference between cost function learning in **a** Amnesic and **b** Incremental approaches

to fairly compare among the amnesic and the incremental approaches as follows,

$$NSE_{i,j} = \frac{1}{N_i}\sum_{k=0}^{N_i}(\hat{\mathbf{y}}_i(k) - \mathbf{y}_j(k))^2 \qquad j = 1, 2, \ldots 8 \quad (12)$$

where the estimated saccade $\hat{\mathbf{y}}_i$ is calculated based on the learned cost function from demonstration i by solving the forward LQR problem while the actual saccades $\mathbf{y}_j$ are the measured demonstrations assigned as $D_1, D_2, \ldots D_8$ in Fig. 5.

Results for both learning models are shown in Fig. 6, where the horizontal axis in each plot represents the demonstration i that is used to estimate the cost function, while the vertical axis shows the means and the standard deviations of $NSE_{i,j}$ computed between the estimated saccade and the measured saccades from other demonstrations. As noted, the $NSE_{i,j}$ means and standard deviations of the first scenario in the incremental learning model are identical to the amnesic one. This due to the fact that at the beginning, the incremental model has no previous scenarios to learn from as the amnesic one. After subsequent scenarios, it is apparent that the incremental model affords a generalization behavior towards unseen demonstrations where it becomes conservative more and more toward new demonstrations. On contrary, the amnesic model learns locally form the current scenario while a more changeable behavior is afforded toward subsequent demonstrations. In addition, the simulated and the measured behaviors in terms of eye position and velocity are plotted in Fig. 7 for both the amnesic and the incremental learning models. As depicted, the incremental model exhibits many similar trends to the measured behavior in terms of position and velocity compared to that of the amnesic model. This perhaps supports the key objective of the proposed Incremental ID-LQR approach to generalize for the unseen new demonstrations.

6.3 Comparison with the conventional ILQR

In the end, the benefit behind introducing the discount matrix $\boldsymbol{\alpha}$ in the proposed ID-LQR is assessed in comparison to the non-discounted ILQR, i.e. $\alpha = 0$, that is proposed in [10] and [34].

In this context, the threshold, γ, of the maximum eigenvalue of block diagonal matrix QR, i.e. $\lambda_{max}(QR) \leq \gamma$ is altered. Hence, the minimum allowable condition number of the retrieved weighting matrices Q and R is changed as explained in Section 4. Subsequently, the associated fitness values in terms of Frobenius norm are calculated for both the non-discounted (ILQR) and the proposed discounted (ID-LQR) approaches. As depicted in Fig. 8, for the same threshold γ, the proposed ID-LQR approach has minimum Frobenius norm values compared to that of the ILQR approach. In addition, the ILQR performance is largely affected by the threshold γ values as shown. On contrary, the performance of proposed ID-LQR, is rapidly increased by increasing the threshold at the beginning; and then, it is quietly unaffected by increasing the values of γ. Hence, the proposed ID-LQR outperforms the previously proposed ILQR by invoking that the precision of using a matrix in subsequent operations is highly degraded by its higher condition numbers [7]. This is likely explained as follow: since the saccade is a fast and an abrupt movement, matching a behavior like this requires high penalization terms in the matrix Q. This, in turn, increases the condition number of Q and could affect its precision as in the non-discounted ILQR [10, 34] case. In opposite, in the proposed ID-LQR, the discount matrix found in (2) takes part of this penalization. So, the demonstrations are quite matched with lower values of discrepancies while lower terms are found in the Q without increasing its condition number.

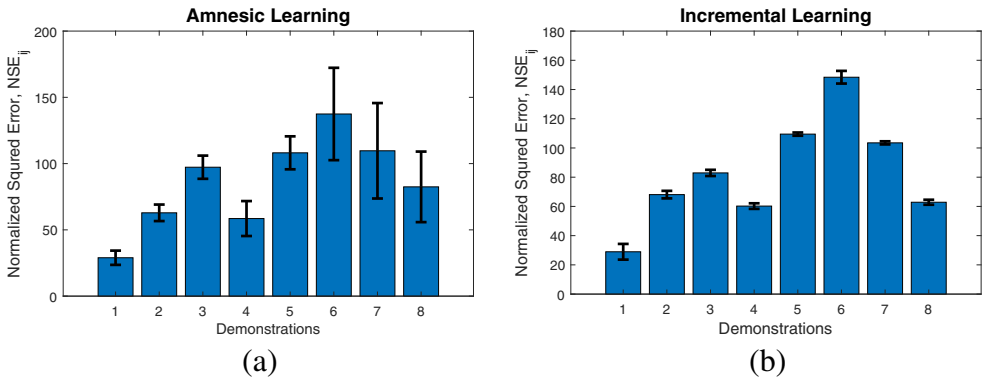
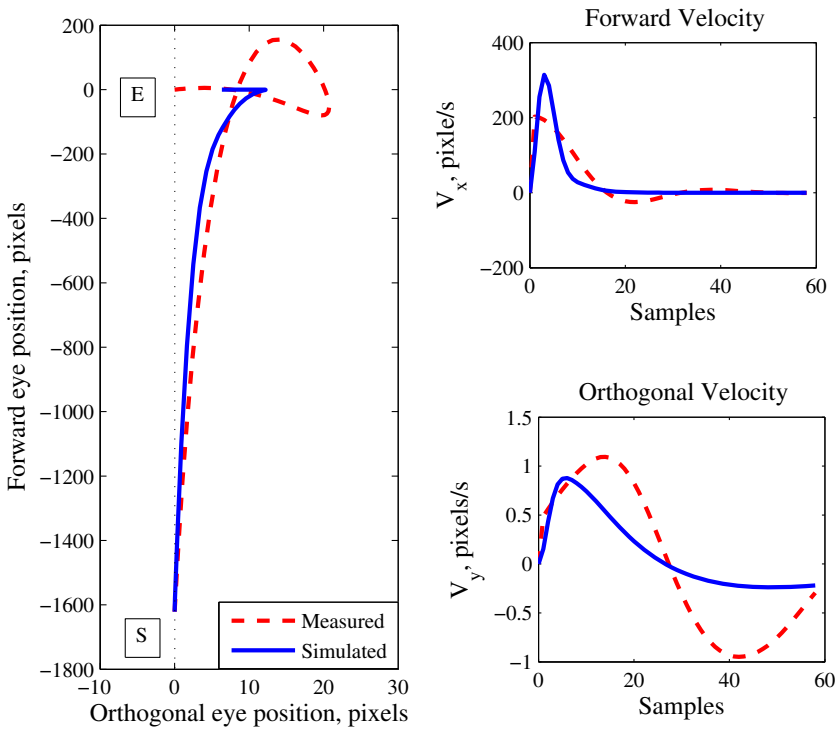
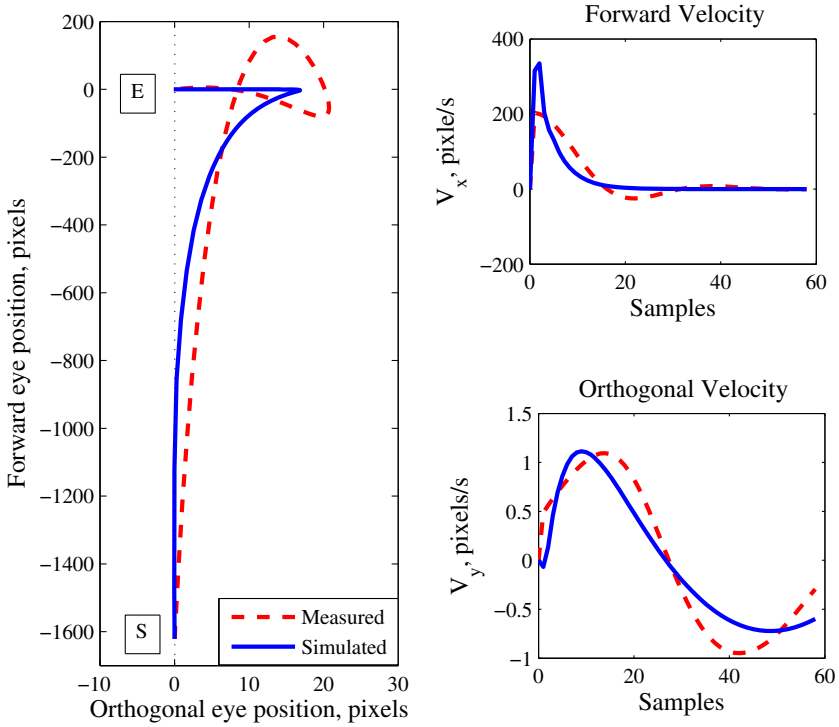


Fig. 6 Comparison between the Amnesic and the Incremental learning models in terms of Normalized Squared Error (NSE)



(a) Amnesic Learning Model



(b) Incremental Learinig Model

Fig. 7 Comparison of the simulated and measured gaze positions and velocities for the two learning model **a** Amnesic and **b** Incremental. $\boxed{S}$ and $\boxed{E}$ represent the start and the endpoints

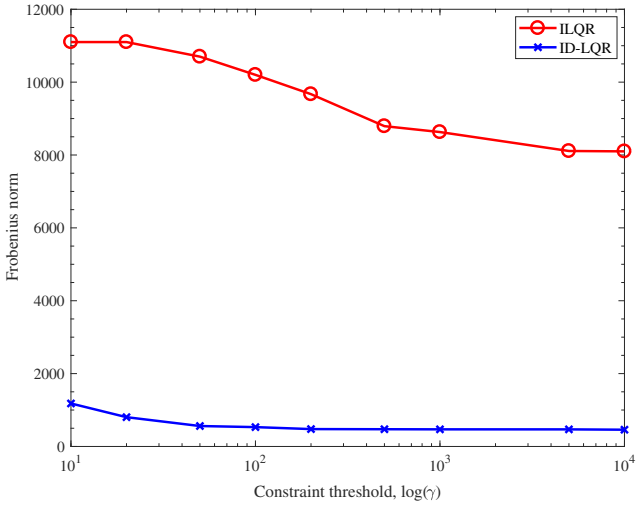


Fig. 8 Comparison between ILQR [10, 34] and the proposed ID-LQR in terms of Frobenius norm values for different thresholds, $\log(\gamma)$

7 Conclusion

A new Inverse Optimal Control (IOC) algorithm called ID-LQR has been developed for learning the optimization criteria underlying a given human behavior. The key idea of the proposed algorithm is to retrieve the discounted cost function that could reproduce the given human behavior perfectly. The discount matrix proposed in ID-LQR not only allows for parameters tuning, it also affords further investigations about the degree of aggressiveness afforded for each state by the feedback controller representing the human. In addition, an incremental ID-LQR algorithm has been developed to incrementally refine the retrieved cost function when receiving other untrained demonstrations even with different initial conditions and lengths.

The developed ID-LQR has been validated through the application of retrieving the cost function of the saccadic eye gaze movements. Although the obtained weighting matrices seem to be reasonable, a further biomedical investigation has to be conducted to confirm the resulting saccades model. Additionally, in the modeling of saccadic eye movements, the incremental ID-LQR of the weighting matrices proves generalization towards the new unseen demonstrations. This perhaps is considered a step towards avoiding the well-known the over-fitting problem. Moreover, the proposed ID-LQR has quietly outperformed the previously ILQR approaches proposed in the literature in terms of fitting errors at different specified condition number constraints.

In future, we are looking for further incorporate human behaviors that include moving targets, i.e., the eye gaze movements towards moving instead of stationary goal objects. In addition, the Linear Quadratic Gaussian (LQG) could be used to free the assumption of having full measurable states with an additional discount matrix as proposed in ID-LQR.

Acknowledgements This research was partially supported by the project "Toward the Next Generation of Robotic Humanitarian Assistance and Disaster Relief: Fundamental Enabling Technologies (10069072)" and by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP)(No. NRF-2016R1E1A1A02921594).

References

1. Abaid N, Cappa P, Palermo E, Petrarca M, Porfiri M (2012) Gait detection in children with and without hemiplegia using single-axis wearable gyroscopes. PloS One 8(9):e73,152–e73,152
2. Abbeel P, Ng AY (2004) Apprenticeship learning via inverse reinforcement learning. In: Proceedings of the twenty-first international conference on machine learning. ACM, p 1
3. Ahmad BI, Murphy JK, Langdon PM, Godsill SJ, Hardy R, Skrypchuk L (2016) Intent inference for hand pointing gesture-based interactions in vehicles. IEEE Trans Cybern 46(4):878–889
4. Atkeson CG, Schaal S (1997) Robot learning from demonstration. In: ICML, vol 97, pp 12–20
5. Attia A, Dayan S (2018) Global overview of imitation learning. arXiv:1801.06503
6. Bijl H, Schön TB (2017) Optimal controller/observer gains of discounted-cost lqg systems. arXiv:1706.01042
7. Cheney E, Kincaid D (2012) Numerical mathematics and computing. Nelson Education
8. Deb K, Pratap A, Agarwal S, Meyarivan T (2002) A fast and elitist multiobjective genetic algorithm: Nsga-ii. IEEE Trans Evol Comput 6(2):182–197
9. Dragan AD, Srinivasa SS (2013) A policy-blending formalism for shared control. Int J Robot Res 32(7):790–805
10. El-Hussieny H, Abouelsoud A, Assal SF, Megahed SM (2016) Adaptive learning of human motor behaviors: an evolving inverse optimal control approach. Eng Appl Artif Intel 50:115–124
11. El-Hussieny H, Asker A, Salah O (2017) Learning the sit-to-stand human behavior: an inverse optimal control approach. In: 2017 13th international computer engineering conference (ICENCO), pp 112–117. https://doi.org/10.1109/ICENCO.2017.8289773
12. Enderle JD, Wolfe JW (1987) Time-optimal control of saccadic eye movements. IEEE Trans Biomed Eng, 43–55
13. Englert P, Vien NA, Toussaint M (2017) Inverse kkt: learning cost functions of manipulation tasks from demonstrations. Int J Robot Res 36(13–14):1474–1488
14. Finn C, Levine S, Abbeel P (2016) Guided cost learning: deep inverse optimal control via policy optimization. In: International conference on machine learning, pp 49–58
15. Freedman EG (2001) Interactions between eye and head control signals can account for movement kinematics. Biol Cybern 84(6):453–462
16. Galiana H, Guitton D (1992) Central organization and modeling of eye-head coordination during orienting gaze shifts. Ann N Y Acad Sci 656(1):452–471
17. Giorgino T (2009) Computing and visualizing dynamic time warping alignments in R: the dtw package. J Stat Softw 31(7):1–24. http://www.jstatsoft.org/v31/i07/
18. Harris CM (1998) On the optimal control of behaviour: a stochastic perspective. J Neurosci Methods 83(1):73–88
19. Harris CM, Wolpert DM (1998) Signal-dependent noise determines motor planning. Nature 394(6695):780

20. Huston SJ, Jayaraman V (2011) Studying sensorimotor integration in insects. Curr Opinion Neurobiol 21(4):527–534
21. Kardamakis AA, Moschovakis AK (2009) Optimal control of gaze shifts. J Neurosci 29(24):7723–7730
22. Khokar KH, Alqasemi R, Sarkar S, Dubey RV (2013) Human motion intention based scaled teleoperation for orientation assistance in preshaping for grasping. In: 2013 IEEE international conference on rehabilitation robotics (ICORR). IEEE, pp 1–6
23. Kirk DE (2012) Optimal control theory: an introduction. Courier Corporation
24. Kober J, Bagnell JA, Peters J (2013) Reinforcement learning in robotics: a survey. Int J Robot Res 32(11):1238–1274
25. Kwakernaak H, Sivan R (1972) Linear optimal control systems, vol 1. Wiley-Interscience, New York
26. Lee SJ, Popović Z (2010) Learning behavior styles with inverse reinforcement learning. In: ACM transactions on graphics (TOG), vol 29. ACM, p 122
27. Mombaur K, Truong A, Laumond JP (2010) From human to humanoid locomotion—an inverse optimal control approach. Autonom Robots 28(3):369–383
28. Muhammad W, Spratling MW (2017) A neural model of coordinated head and eye movement control. J Intell Robot Syst 85(1):107–126
29. Ng AY, Russell SJ et al (2000) Algorithms for inverse reinforcement learning. In: Icml, pp 663–670
30. Parisi GI, Magg S, Wermter S (2016) Human motion assessment in real time using recurrent self-organization. In: 2016 25th IEEE international symposium on robot and human interactive communication (RO-MAN), pp 71–76. https://doi.org/10.1109/ROMAN.2016.7745093
31. Phaniteja S, Dewangan P, Guhan P, Sarkar A, Krishna KM (2018) A deep reinforcement learning approach for dynamically stable inverse kinematics of humanoid robots. arXiv:1801.10425
32. Powell MJ (2009) The bobyqa algorithm for bound constrained optimization without derivatives. Cambridge NA Report NA2009/06. University of Cambridge, Cambridge, pp 26–46
33. Priess MC, Choi J, Radcliffe C (2014) The inverse problem of continuous-time linear quadratic gaussian control with application to biological systems analysis. In: ASME 2014 dynamic systems and control conference. American Society of Mechanical Engineers, pp V003T42A004–V003T42A004
34. Priess MC, Conway R, Choi J, Popovich JM, Radcliffe C (2015) Solutions to the inverse lqr problem with application to biological systems analysis. IEEE Trans Control Syst Technol 23(2):770–777
35. Ramachandran D, Amir E (2007) Bayesian inverse reinforcement learning. Urbana 51(61801):1–4
36. Ramadan A, Choi J, Radcliffe CJ (2016) Inferring human subject motor control intent using inverse mpc. In: American control conference (ACC), 2016. IEEE, pp 5791–5796
37. Ratliff ND, Silver D, Bagnell JA (2009) Learning to search: functional gradient techniques for imitation learning. Auton Robot 27(1):25–53
38. Roncone A, Pattacini U, Metta G, Natale L (2016) A cartesian 6-dof gaze controller for humanoid robots. In: Robotics: science and systems
39. Saeb S, Weber C, Triesch J (2011) Learning the optimal control of coordinated eye and head movements. PLoS Comput Biol 7(11):e1002,253
40. Soechting J, Lacquaniti F (1981) Invariant characteristics of a pointing movement in man. J Neurosci 1(7):710–720
41. Spiers A, Khan SG, Herrmann G (2016) Human motion. Springer International Publishing, pp 49–74
42. Suleiman W, Yoshida E, Kanehiro F, Laumond JP, Monin A (2008) On human motion imitation by humanoid robot. In: IEEE international conference on robotics and automation, 2008. ICRA 2008. IEEE, pp 2697–2704
43. Todorov E (2004) Optimality principles in sensorimotor control. Nat Neurosci 7(9):907–915
44. Watkins CJCH (1989) Learning from delayed rewards. Ph.D. thesis. King's College, Cambridge
45. Zhifei S, Joo EM (2012) A review of inverse reinforcement learning theory and recent advances. In: 2012 IEEE congress on evolutionary computation (CEC). IEEE, pp 1–8
46. Zhifei S, Meng Joo E (2012) A survey of inverse reinforcement learning techniques. Int J Intell Comput Cybern 5(3):293–311
47. Zimmerman ME (2011) Speed–accuracy tradeoff. In: Encyclopedia of clinical neuropsychology. Springer, pp 2344–2344

Haitham El-Hussieny (M'16) received the B.S. degree in electronics and communication engineering from Benha University, Benha, Egypt, in 2007, and the M.S. and Ph.D. degrees in mechatronics and robotics engineering from Egypt-Japan University of Science and Technology (E-JUST), Alexandria, Egypt, in 2013 and 2016, respectively. He is an Assistant Professor in the Department of Electrical Engineering, Benha University, Egypt. He is also working as an Adjunct Assistant Professor in E-JUST. His research interests include behavior modeling, learning by demonstrations, soft robots and Human-Robot Interaction (HRI).

Jee-Hwan Ryu (M'02) received the B.S. degree in mechanical engineering from Inha University, Incheon, South Korea, in 1995, and the M.S. and Ph.D. degrees in mechanical engineering from Korea Advanced Institute of Science and Technology, Taejon, South Korea, in 1995 and 2002, respectively. He is a Professor in the Department of Mechanical Engineering, KOREATECH, Cheonan, South Korea. His research interests include haptics, telerobotics, exoskeletons, and autonomous vehicles.