Reinforcement Learning for Energy Optimization Under Human Fatigue Constraints of Power-Assisted Wheelchairs

Guoxi Feng, Lucian Bușoniu, Thierry Marie Guerra, Sami Mohammad

***Abstract* — In the last decade, Power-Assisted Wheelchairs (PAWs) have been widely used for improving the mobility of disabled persons. The main advantage of PAWs is that users can keep a suitable physical activity. Moreover, the metabolic-electrical energy hybridization of PAWs provides more flexibility for optimal control design. In this context, we propose an optimal control for minimizing the electrical energy consumption under human fatigue constraints, including a human fatigue model. The electrical motor has to cooperate with the user over a given distance-to-go. As the human fatigue model is unknown in practice, we use model-free Policy Gradient methods to directly learn controllers for a given driving task. We verify that the model-free solution is near-optimal by computing the model-based controller, which is generated by Approximate Dynamic Programming. Simulation results confirm that the model-free Policy Gradient method provides near-optimal solutions.**

I. INTRODUCTION

Improving the mobility of disabled and elderly persons is becoming an important challenge in ageing societies [1]. Power-assisted wheelchairs (PAW) are a promising solution to address these mobility issues. Compared to traditional wheelchairs, PAWs such as the motorization kits Duo and Nomad designed by AutoNomad Mobility [9] provide a good compromise between rest and physical exercise for users [6]. Consequently, it prevents disabled people from suffering the issues caused by a long-term use of manual wheelchair, such as rotator cuff tendonitis, lateral epicondylitis and calcific tendonitis [8]. Meanwhile, PAWs can also enable users to maintain a desired physical activity level, which cannot be provided by a fully electric powered wheelchair.

The first major novelty in this paper is a control strategy for PAWs that optimizes electrical energy while also taking into account human fatigue. In particular, we consider a scenario where the PAW must drive for a desired distance, allowing a desired fatigue variation on the whole trajectory, and minimizing electrical energy consumption. Due to the initial-to-final fatigue constraint, the obtained policy is expected to provide a good compromise between transforming metabolic energy into force and rest to users. The aim is a suitable cooperation of the electrical motor with the users, supplying an appropriate assistance.

Our assistive control design is formulated mathematically as optimal control. The state of fatigue (S_{of}) of the human is described by the single-state model [7], and the maximum available human force is limited by the fatigue. The human applies forces alongside the electrical motor, acting as a separate, nonlinear velocity controller that varies with the assistance. Specifically, a proportional control law is assumed, in which the desired velocity depends on the user's motivation, which is in turn affected by the assistance and state of fatigue. The objective function includes an electrical energy cost at each step, and terminal costs that penalize deviations from the desired distance and state of fatigue.

The second major contribution of the paper is the application of an online, model-free reinforcement learning method to solve the optimal control problem. The solution is learned by treating the entire system dynamics (wheelchair, human fatigue, and human controller) as a black box, and the algorithm only needs state measurements and costs. This model-free nature is crucial in practice, since the human dynamics is unavailable. The optimal control method of choice that we evaluate is Policy Gradient (PG), so far mostly used in robotics, *e.g.* [17]-[19]. To verify the quality of the PG solution, we compare it to a baseline model-based solution, computed with a finite-horizon extension of the interpolation-based dynamic programming technique in [21]. The PG method reaches close to the near-optimal solution found by the model-based technique.

To our best knowledge, no works in the PAW literature address energy optimization with human fatigue considerations for PAW design. Most existing works only deal with energy consumption without considering human fatigue, for example using fuzzy rule-based approaches [20]; where, of course, no guarantee of optimality can be given. The similar works [3]-[5] on hybrid bicycle design investigate model-based optimal energy management considering physiological human factors. Because of the similar energy storage structure between PAWs and hybrid bicycles, our results might also be applicable to assisted bicycles.

The paper is organized as follows. Section 2 introduces the modeling of human-wheelchair system and the problem statement. Section 3 introduces two existing optimal control approaches in the literature. In Section 4, we apply the model-free presented in Section 3 for PAW design and extend the ADP approach to drive a finite-horizon baseline solution. Simulation results are presented Section 5 for validating the proposed approach. Section 6 gives conclusion and discusses future work.

Guoxi Feng, Thierry Marie Guerra are with Univ. Valenciennes CNRS, UMR 8201 – LAMIH – F59313 Valenciennes, France ({guoxi.feng, guerra}@univ-valenciennes.fr). Lucian Bușoniu is with the Department of Automation, Technical University of Cluj-Napoca, Memorandumului 28, 400114 Cluj-Napoca, Romania (lucian@busoniu.net). Sami Mohammad is with AutoNomad Mobility Le Mont Houy, 59313 Valenciennes Cedex 9, France (sami.mohammad@autonomad-mobility.com).

II. Modeling Of The Human-Wheelchair System And Problem Statement

A. *Human Fatigue Model*

Due to the complexity of human metabolism, many physical or physiological variables are needed to estimate precisely the human fatigue. For the sake of control design, single-state models based on heart rate [10], [11] or oxygen uptake [13], [14] have been used to quantify the human physical task. However, these approaches require sensors which are difficult to implement for practical systems.

In this study, to overcome the above limitations, we apply the human fatigue model from [7] that describes human muscle fatigue as a single-state first order dynamic process, which involves simultaneously the fatigue and the recovery effects. These muscular phenomena have been explained in clinical investigation [12]. Only the human torque measurement is required to estimate the human fatigue. Before introducing the S_{of}, the dynamics of the maximum available force $F_m(t)$ provided by human are needed:

$$\dot{F}_m(t) = -\left(R + \frac{k}{M_{vc}} F_h(t)\right) F_m(t) + R \cdot M_{vc} \tag{1}$$

where M_{vc} is the Maximum Voluntary Contraction force at rest, $F_h(t)$ is the human applied force and k and R represent the fatigue and the recovery coefficients respectively. Of course, $0 \le F_h(t) \le F_m(t) \le M_{vc}$. In the reminder of the paper, the maximum available force $F_m(t)$ is assumed not to be affected by the muscular contraction velocity.

If the user applies constantly the maximum feasible force $F_m(t)$ as $F_h(t)$, then $F_m(t)$ decreases at its maximum rate. This leads (1) to an equilibrium point where the fatigue rate is identical to the recovery rate, and $\dot{F}_m(t) = 0$, *i.e.*:

$$-\left(R + \frac{k}{M_{vc}} F_m(t)\right) F_m(t) + R \cdot M_{vc} = 0 \tag{2}$$

The positive solution is:

$$F_{eq} = \frac{R \cdot M_{vc}}{2k}\left(-1 + \sqrt{1 + \frac{4k}{R}}\right) \tag{3}$$

This equilibrium value F_{eq} is also the minimum threshold that $F_m(t)$ can achieve. Thus $F_{eq} \le F_m(t) \le M_{vc}$. The State of Fatigue is then defined as:

$$S_{of}(t) = \frac{M_{vc} - F_m(t)}{M_{vc} - F_{eq}} \tag{4}$$

Thus, the S_{of} is the normalized value of $F_m(t)$ (since $0 \le S_{of}(t) \le 1$) and is an indicator to quantify the human fatigue.

B. *Wheelchair Model and Human Controller*

The wheelchair dynamics are described as follows:

$$\mathrm{x}(k+1) = Ax(k) + B(U_m(k) + F_h(k) \cdot r) \tag{5}$$

where $A \in \mathbb{R}^{2\times 2}$, $B \in \mathbb{R}^{2\times 1}$, $x(k) = [d(k), v(k)]^T$, $d(k)$ is the wheelchair position, $v(k)$ is the wheelchair velocity. We assume that the human force $F_h(k)$ depends on the fatigue state, the electrical motor assistance and the wheelchair velocity (perceived by the user):

$$F_h(k) = y(U_m(k), S_{of}(k), v(k)) \tag{6}$$

Before formulating this human force, we use the fatigue-motivation model [2] to describe how the fatigue and the assistance affect human motivation. The human fatigue decreases the motivation and the perceived help increases the motivation. The normalized perceived help is:

$$H(k) = U_m(k)/U_{m-max} \in [0,1] \tag{7}$$

The equilibrium point between the perceived fatigue and the perceived help is:

$$f(k) = \frac{H(k) - S_{of}(k)}{H(k) + S_{of}(k)} \in [-1,1] \tag{8}$$

The motivation is:

$$M(k) = \begin{cases} f_n(1 + f(k)) & f(k) < 0 \\ f_n + (1 - f_n) f(k) & f(k) \ge 0 \end{cases} \tag{9}$$

where $M(k) \in [0,1]$ and f_n is the fraction of the maximum velocity V_{max}. When the fatigue and help perceptions are balanced, the motivation $M(k) = f_n$. The user motivation in (9) affects proportionally the desired wheelchair velocity $V_{r-human}$ of the user, so that a higher motivation leads to a higher desired velocity. The desired velocity is therefore:

$$V_{r-human}(k) = V_{max} M(k) \tag{10}$$

Finally, the human force is modeled as a proportional velocity tracking controller:

$$F_h(k) = K_p\big(V_{max} M(k) - v(k)\big) \tag{11}$$

Moreover, the human force should be saturated by $F_m(k)$ and only the positive human force is taken into account:

$$F_h(k) = sat(0, F_m(k), F_h(k)) \tag{12}$$

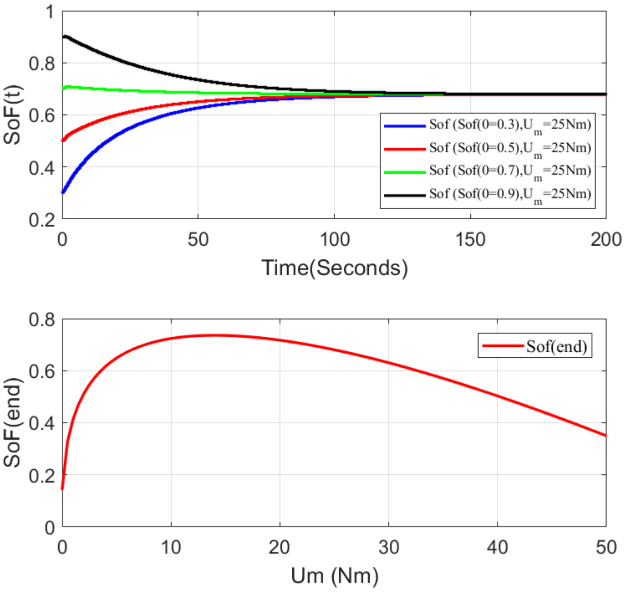

Figure 1: $S_{of}(k)$ evolution with a constant $U_m = 25Nm$ (above) and $S_{of}(end)$ evolution respect to U_m (below)

Remark 1: The human controller represented by (12) is an implicit S_{of} − tracking controller for the interconnected wheelchair/human dynamics. Simulation results (with $f_n = 0.5$) in Fig. 1 show that the S_{of} converges to a specific value for a constant U_m; the proposed model manages the human fatigue depending on the perceived environment. Interestingly, the left part of the second curve, Fig. 1 illustrates the fact that increasing perceived help motivates the user to do more physical exercise. The right part of this curve shows that when the assistive torque is increased, the motor assists the user to decrease his physical workload.

C. Problem Statement

We consider a distance-to-go driving schedule with a predefined time horizon. The main goal is to find the control law U_m so that users have a desired fatigue variation ΔS_{of} and the wheelchair travels the given distance Δd, or equivalently:

$$S_{of}(N) - S_{of}(0) = \Delta S_{of} \qquad (13)$$
$$d(N) - d(0) = \Delta d$$

with given $S_{of}(0)$ and $d(0)$, while minimizing the electrical energy consumption over the driving task. Here, $S_{of}(N)$ is the final S_{of} and $d(N)$ the final distance.

In the interest of simplifying our primary analysis, the electric energy consumption J_{elect} is considered as a quadratic function of U_m:

$$J_{elect} = \sum_{k=0}^{N-1} \frac{1}{2} U_m^2(k) \qquad (14)$$

Considering the terminal constraints (13) and the electric energy consumption (14), we state the finite horizon criterion to minimize as:

$$\min_{U_m} J = [w_1\ w_2] \begin{bmatrix} \left(d(N) - \bar{d}\right)^2 \\ \left(S_{of}(N) - \overline{S_{of}}\right)^2 \end{bmatrix} + \sum_{k=0}^{N-1} \frac{1}{2} U_m^2(k) \qquad (15)$$

subject to the human fatigue dynamics (1), the wheelchair dynamics (5) and the human controller (6).

III. Optimal Control Methods

The first approach is the Policy Gradient reinforcement learning [16] which is largely applied for the control design of physical systems, thanks to its model-free online learning nature and considerable success in high-dimensional systems. The second one is Approximate Dynamic Programming [21], which provides a near-optimal solution used as a baseline.

A. Policy Gradient

Modeling complex processes like human behaviors remains a major challenge for control design. Therefore, model-free PG approaches avoid the need for a model and deliver directly a control policy which tries to collect as much reward as possible. However, it is necessary to have access to the S_{of} measurement to learn its dynamics. In what follows, the S_{of} is assumed to be measured and the cost function (15) to minimize is formulated as a reward function *r* to maximize. The sum of the rewards over a finite-horizon:

$$R(\tau) = T(x_N) + \sum_{k=0}^{N-1} r(x_k, u_k) \qquad (16)$$

is called the return R, where $T(x_N)$ is the terminal reward, $r(x_k, u_k)$ is the stage reward, u is the control input, x is the state of the system and $\tau = (x_0, u_0, x_1, u_1, \dots x_{N-1}, u_{N-1}, x_N)$ is a trajectory of the system. Since exploration is indispensable to learn the unknown dynamics, stochastic policies are needed for model-free policy search methods. Hence, the trajectory probability distribution $p_\pi(\tau)$ for a stochastic control system can be expressed as:

$$p_\theta(\tau) = p(x_0) \prod_{k=0}^{N-1} [p(x_{k+1}|x_k, u_k) \pi_\theta(u_k|x_k, k)] \qquad (17)$$

where $p(x_0)$ is the initial state distribution, and $\pi_\theta(u_k|x_k, k)$ is the policy distribution with the control parameters θ. As the considered state transition is deterministic, see (1) and (5), $p(x_{k+1}|x_k, u_k) = 1$ for $x_{k+1} = f_t(x_k, u_k)$ with the transition function f_t. Hence, (17) is rewritten as:

$$p_\theta(\tau) = p(x_0) \prod_{k=0}^{N-1} \pi_\theta(u_k|x_k, k) \qquad (18)$$

For trajectories τ generated by policy π_θ, expected return is:

$$\tilde{R}_\theta = \int p_\theta(\tau) R(\tau) d\tau \qquad (19)$$

Policy gradient methods update the control parameters θ in the steepest ascent direction so that the expected return (16) is maximized. The update law of the control parameters θ with the learning rate α ($\alpha > 0$) is:

$$\theta_{i+1} = \theta_i + \alpha \cdot \nabla_\theta \tilde{R}_\theta \qquad (20)$$

where i is the iteration index and the policy gradient $\nabla_\theta \tilde{R}_\theta$ is:

$$\nabla_\theta \tilde{R}_\theta = \int \nabla_\theta p_\theta(\tau) R(\tau) d\tau \qquad (21)$$

Since $\nabla_\theta p_\theta(\tau) = p_\theta(\tau) \nabla_\theta \log p_\theta(\tau)$, we have:

$$\nabla_\theta \tilde{R}_\theta = \int p_\theta(\tau) \nabla_\theta \log p_\theta(\tau)\, R(\tau) d\tau \qquad (22)$$

Replacing $p_\theta(\tau)$ by (18), we obtain:

$$\nabla_\theta \tilde{R}_\theta = \int p_\theta(\tau) \nabla_\theta \left[\log p(x_0) \prod_{k=0}^{N-1} \pi_\theta(u_k|x_k, k) \right] R(\tau) d\tau \qquad (23)$$

Finally, by replacing the integral with the equivalent expected value notation, the REINFORCE [15], [16] policy gradient is:

$$\nabla_\theta \tilde{R}_\theta = E_\tau \left[\sum_{k=0}^{N-1} \{\nabla_\theta \log \pi_\theta(u_k|x_k, k)\} R(\tau) \right] \qquad (24)$$

From (24), we derive the policy gradient of GPOMDP (more details can be found in [15], [16]):

$$\nabla_\theta \tilde{R}_\theta = E_\tau \left[\sum_{k=0}^{N-1} \sum_{j=0}^{k} [\nabla_\theta \log \pi_\theta(u_k|x_k, k)]\, r_j \right] \qquad (25)$$

where r_j is the stage reward, in (25) the gradient $\nabla_\theta \tilde{R}_\theta$ depends only on the current policy distribution $\pi_\theta(u_k|x_k, k)$. The gradient calculus is replaced with $\nabla_\theta \log \pi_\theta(u_k|x_k, k)$ calculus. In (24) and (25), the expected value is approximated using Monte Carlo techniques.

B. Approximate Dynamic Programming

To derive a finite-horizon near-optimal policy U_m, we extend the model-based approximate dynamic programming (ADP) algorithm in [21], which is originally used to solve infinite-horizon problems. In the original algorithm, the idea is to find a near-optimal policy that maximizes a predefined infinite-horizon return function. This near-optimal policy can be described by the approximate Q-function, which lies within a bounded distance from the optimal Q-function Q^*:

$$Q^*(x, u) = \rho(x, u) + \gamma \sup_\pi R^\pi\left(\xi(x, u)\right) \qquad (26)$$

where ξ is the transition function, γ is the discount factor, $\rho(x, u)$ is the reward and R is the discounted return from the next state $\xi(x, u)$. The optimal policy π^* can be found from Q^* (*i.e.* $\pi^* = \arg\max_u Q^*(x, u)$). In order to approximate Q^* and π^*, an approximator is used that relies on an interpolation over the state space, and on a discretization of the action space.

The Q-value of the pair (x, u) is approximated by the Q-value of the pair (x, u_d) interpolated over x, where u_d has the closest Euclidean distance to u in the discrete subset of actions. A parameter vector is defined to represent the Q-function. Each individual parameter is associated with a combination between points on the state-interpolation grids and discrete actions. This parameter vector is obtained by iterating the Bellman equation until convergence.

IV. APPLICATION TO POWER-ASSISTED WHEELCHAIR

The objective of the control policy is to maintain the human fatigue level and minimize the energy consumption over a given distance. From an energy point of view and knowing the distance-to-go, negative human torque and negative motor torque are inefficient in terms of metabolic-electrical energy consumption over the task. An energy optimization algorithm should naturally eliminate this kind of solutions. Applying this prior knowledge to the controller parameterization, U_m is saturated between 0 and the maximum torque U_{max}. But we should keep in mind that the model-free controller can be configured general enough in practice to adapt to unknown situations where no prior knowledge is available.

To evaluate the quality of the control policy obtained by the PG model-free approach, we compare it to a near-optimal solution derived using ADP.

A. Model-free Solution

Here, we aim to compute a control policy U_m to minimize the electrical energy consumption (14) subject to the unknown wheelchair/human dynamics (1), (5) and (6) satisfying $0 \leq U_m(k) \leq U_{max}$, and the terminal constraints (13).

The block diagram Fig. 2 illustrates the black box including the dynamics of the wheelchair/human (*i.e.* human metabolism, wheelchair and human controller). An efficient method to deal with this issue is to learn the controller directly without knowing the system dynamics. The PG algorithm provides a state feedback control law approximately maximizing a given return function.

Using the proposed PG approach, the optimization problem (15) is solved without knowing completely the dynamics (1), (5) and (6). We approximate first the deterministic motor torque by the following basis functions φ (BF):

$$\widetilde{U_m}(k) = \sum_{h=1}^{M} \theta_h \varphi_h(d(k), v(k), Sof(k)) \quad (27)$$

where M is the number of BFs and ϕ_d is the weight of BF. We use radial basis functions (RBFs) in $X(k) = [d(k), v(k), Sof(k)]^T$ to fit the general formalism (27). The resulting parameterization is:

$$\widetilde{U_m}(k) = \sum_{d=1}^{M} \theta_d \exp(-\beta \|X(k) - c_d\|^2) \quad (28)$$

In (28), c_d is the center vector for RBF d. The exploration is carried out by the exploration noise σ which adds directly a random value to the executed action and renders the policy (28) stochastic. The applied exploration noise σ has a normal distribution (29). In order to prevent the executed action from violating the constraint (29), the stochastic human force is selected as:

$$U_m(k) = U_{max}\, q_{sat} \left[\frac{1}{U_{max}} N\left(\widetilde{U_m}(k), \sigma^2\right)\right] \quad (29)$$

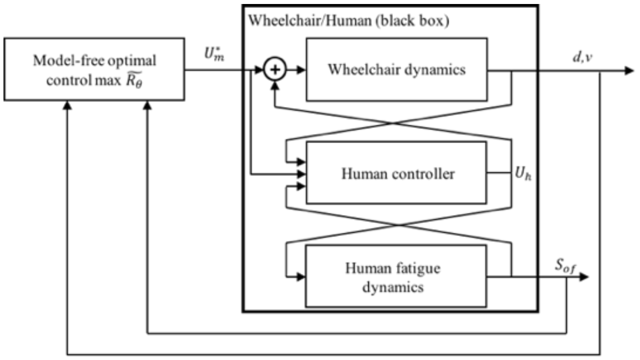

Figure 2: Power assistance algorithm with PG model-free approach

Thus, the motor torque $U_m(k)$ is saturated smoothly by the smooth saturation function q_{sat} between $[0, U_{max}]$, as depicted in Fig. 3. The variance σ of the exploration noise is chosen such that the algorithm explores efficiently the unknown system dynamics described by (1), (5) and (6). Using these new trajectories generated with the exploration noise σ, the PG algorithm (25) updates the parameters θ with (20) to improve the performance of control policy in terms of the expected return (19). According to the normal distribution, the policy distribution can be written as:

$$\pi_\theta\left(U_m(k)\middle|X(k)\right) = \frac{1}{\sqrt{2\pi\sigma^2}} * \exp\left(-\frac{\left(U_{max}\, {q_{sat}}^{-1}(\frac{U_m(k)}{U_{max}}) - \sum_{d=1}^{N} \theta_d \exp(-\beta\|X(k) - c_d\|^2)\right)^2}{2\sigma^2}\right) \quad (30)$$

The derivative of each policy parameter θ_d $(d = 1 \ldots M)$ is:

$$\nabla_{\theta_d} \log \pi_{\theta_d}\left(U_m(k)\middle|X(k)\right) = \frac{\exp(-\beta\|X(k) - c_d\|^2)}{\sigma^2}\left[U_{max}\, {q_{sat}}^{-1}(\frac{U_m(k)}{U_{max}}) - \sum_{d=1}^{M} \theta_d \exp(-\beta\|X(k) - c_d\|^2)\right] \quad (31)$$

To travel the given distance with a desired variation ΔS_{of} defined in (13), the terminal reward in (16) is defined as:

$$T\left(d(N), S_{of}(N)\right) = -[w_1\ w_2]\begin{bmatrix}\left(d(N) - \bar{d}\right)^2 \\ \left(S_{of}(N) - \overline{S_{of}}\right)^2\end{bmatrix} \quad (32)$$

Here, w_1, w_2 are the reward function weights, the distance-to-go is $\bar{d}$, and the desired final S_{of} is $\overline{S_{of}}$.

Owing to the exploration in action space for the policy (29), if the optimal control value $U_m(k)$ is located at the borders of the interval $[0, U_{max}]$, the algorithm permanently increases/ decreases the values $\widetilde{U_m}(k)$ to obtain more expected return. Therefore, a saturation penalty is needed so that the control parameters θ converge. Finally, the energy consumption has to be taken into account in the reward function. Then, the stage reward is:

$$r(x_k, u_k) = -\left[\frac{1}{2}U_m^2(k) + w_3 P_s(U_m(k))\right] \quad (33)$$

where $P_s(U_m(k))$ is the saturation penalty function. The energy consumption is described by the quadratic function mentioned previously in (14), and w_3 is the constraint penalty weight. Overall, the return is defined as follows:

$$R = -w_1\left(d(N) - \overline{d}\right)^2 \tag{34}$$
$$-w_2\left(S_{of}(N) - \overline{S_{of}}\right)^2 - \sum_{k=0}^{N-1}\left[\frac{1}{2}U_m^2(k) + w_3 P_s(U_m(k))\right]$$

where P_s, illustrated Fig. 3, is defined as:

$$P_s = \begin{cases} \sin\left(\frac{\pi}{0.04U_{max}}(U_m - U_{max})\right) + 1 & 0.98U_{max} \le U_m \le U_{max} \\ 0 & 0.02U_{max} \le U_m \le 0.98U_{max} \\ \sin\left(\frac{\pi}{0.04U_{max}}(-U_m)\right) + 1 & 0 \le U_m \le 0.02U_{max} \end{cases} \tag{35}$$

Now, by tuning the learning rate α, the parameters (β, c, M) of the basis functions, the variance σ and the parameters (w_1, w_2, w_3) of the reward function, we have all the conditions to compute $\nabla_\theta \log\pi_\theta\left(U_m(k)\middle|X(k)\right)$ and so the gradient $\nabla_\theta \tilde{R}_\theta$. In this paper, we apply the GPOMDP algorithm (25). The complete algorithm is shown as follows:

```
Initialize the policy parameters $\theta_i$ $(i = 1, 2, 3, \ldots M)$
repeat
  generate trajectories $\tau$ using current policy
  use (31) to compute $\nabla_{\theta_d} log\pi_{\theta_d}\left(U_m(k)\middle|X(k)\right)$
  compute the gradient for each parameter $\theta_d$ with (25)
  update each control parameter $\theta_d$ $(d = 1, 2, \ldots M)$ by (20)
until the return $R$ converges
```

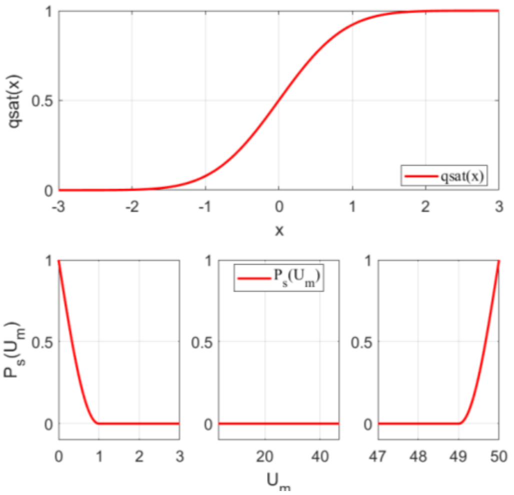

Figure 3: Smooth saturation function q_{sat} (above) and penalty function P_s for $U_{max} = 50N$ (below)

B. Model-based Baseline Solution

To solve our finite-horizon problem, we use the backward iteration of the original algorithm [21] and choose the discount factor γ as 1. For a horizon of 10s with a sampling time 0.05s, the number of the backward iteration is 200. To represent the finite-horizon return (34), the terminal cost is used firstly to compute the Q-function of the last time step, and then each stage is gradually added via the backward dynamic programming iterations. In total, 200 Q-functions are generated to represent a time-varying Q-function for a horizon of 10s. Moreover, we derive the policy from the obtained time-varying Q-function in the forward direction, by choosing the action which maximizes the Q-function of that step and apply it to the system.

V. Simulation Results

For the following simulations, we choose the recovery coefficient $R = 0.0063s^{-1}$, the fatigue coefficient $k = 0.153s^{-1}$, the MVC $M_{vc} = 100N$, the wheel radius $r = 0.33m$ and the control gain K_p of (11) is 30. The parameter f_n of the motivation model is chosen as 0.5. The system matrices of (5) are:

$$A = \begin{bmatrix} 1 & 0.05 \\ 0 & 0.9406 \end{bmatrix}, B = \begin{bmatrix} 0 \\ 0.0059 \end{bmatrix}$$

The reward weights are: $w_1 = 4000$, $w_2 = 10^7$ $w_3 = 800$. Time horizon is 10s, the initial state of fatigue $S_{of}(0) = 0.5$, the desired final human fatigue $\overline{S_{of}}$ is also 0.5 and the distance to go $\overline{D}$ is 20 rad. $d \in [0,30]$ (rad), $v \in [0,7](rad/s)$ and $S_{of} \in [0.35, 0.7]$. The state vector is $X = [d \quad v \quad S_{of}]^T$. The motor torque is bounded ($U_m \in [0,50]$ (Nm)). To apply the model-based ADP approach, we use an equidistant three dimensional $10 \times 10 \times 41$ interpolation grid. 15 discrete actions are chosen for U_m.

For the model-free PG approach, an equidistant three dimensional $5 \times 5 \times 8$ grid is selected as the centers of the RBFs. In total, 200 RBFs ($M = 200$ and $\beta = 0.5$), together with a parameter vector $\theta \epsilon \mathcal{R}^{200}$ are used to approximate the controller (27). The learning rate α is chosen as 10^{-5}.

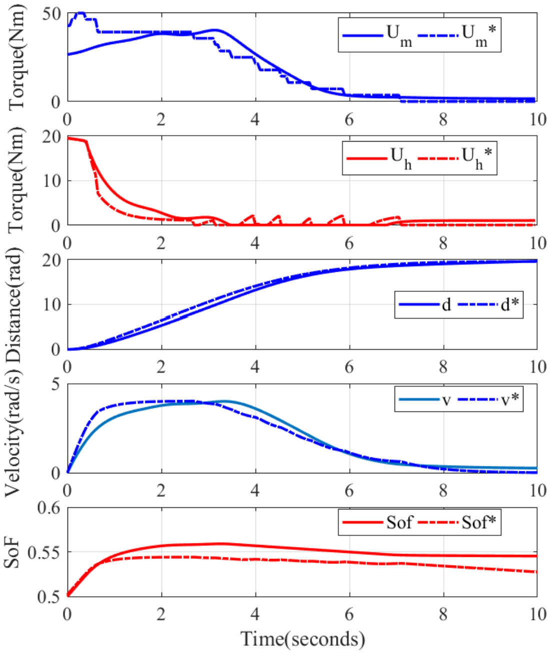

Figure 4: Simulation results provided by GPOMDP algorithm and ADP algorithm

A number of 1600 iterations (5 trajectories of 10s for each iteration) are performed to learn the control parameters θ. We compare the solution of PG with the solution obtained by the ADP. As shown in Fig. 4, PG approach (solid line) has a similar performance with ADP approach (dotted line). The obtained simulation results as follows: The final return is $-8.2017e4$ for PG (the energy consumption: $-5.3893e4$ and the terminal penalty: $-2.8124e4$). The final return is $-6.9837e4$ for ADP (the energy consumption: $-6.4726e4$ and the terminal penalty: $-5.1108e3$). The PG approach provides 12.7% less return than ADP. However, the PG approach eliminates the need for a model by accepting this loss in return. It is important to mention this 12.7% difference includes both an electrical energy component and a difference in the final *Sof* reached by the two methods.

From a practical point of view, first the user cooperates with the motor to push the wheelchair. After reaching a suitable velocity between $1s$ and $3.5s$, the user reduces his applied force to reduce his fatigue. During this time, the electrical motor provides the main input to maintain this velocity. In the reminder of the driving, the motor assistance is reduced gradually to minimize the energy consumption. Moreover, the user tries to attain the desired final fatigue level by reducing his force. The system uses the kinetic energy given previously by the user and the motor to end the mission. During the driving task, the provided assistive algorithm tries to provide an energy-efficient assistance to the user so that his final fatigue level reaches the desired one.

For the model-free PG approach, we have a terminal error of 0.05 between the final $S_{of}(N)$ and the desired final value $\overline{S_{of}}$ (0.02 for the ADP approach). This error can be reduced by increasing the weighting factor w_2. However, the energy consumption should have a significant weight in the return function (34) to fulfill the optimization objective. The weight parameters w_1, w_2 and w_3 must be tuned to have a tradeoff between reaching the terminal conditions and minimizing the energy consumption. The learning rate α tuning also depends on the weighting factors and parameters (β, c, M). Since no prior knowledge about the optimal policy is available, an equidistant grid on the given intervals is chosen for the centers of the RBFs. If we increase the number of RBFs, the approximate controller may tend to the optimal solution after receiving enough training. Roughly speaking, around 20-30 preliminary experiments are required to fix the 4 parameters and the RBFs used in this paper.

Remark 2*:* Models (1), (5) and (6) do not fully represent the real dynamics of the system. However, The PG approach treats the whole wheelchair/human dynamics (including the negative torque of the user) as unknown. In real-time applications, we expect the good performance of the algorithm observed here to generalize also to other dynamics.

VI. Conclusion and Future Work

In this paper, a novel design of PAW control has been proposed based on an energy optimization under human fatigue constraint. The idea is to maintain a suitable fatigue level for users while reducing the electrical energy consumption over a driving task. Using a mathematical model to describe the human fatigue dynamics, we consider a distance-to-go driving task to carry out the simulations.

We applied the model-free approach PG for this distance-to-go driving task. The wheelchair, human fatigue and human controller were treated as unknown dynamics. Simulation results illustrated that the assistive algorithm provided by PG tries to improve energy efficiency, despite an acceptable error between the final $S_{of}(N)$ and the desired final value $\overline{S_{of}}$. Moreover, we show that this policy obtained by PG is not far from a near-optimal solution derived by ADP.

In the considered driving task, S_{of} is assumed to be measured, which is not possible in practice. In future research, the fatigue has to be estimated via a physical variable (*i.e.* human input torque). The estimated information would feed into the PG algorithm.

As a considerable amount of data is needed to obtain a high performance controller, more PG learning techniques should be investigated to reduce the learning time. The ultimate objective is to develop an efficient real-time learning control of PAWs.

Reference

[1] World Health Organization. (2011). World report on disability. World Health Organization.

[2] Ronchi E., P.A. Reneke, R.D. Peacock. A conceptual fatigue-motivation model to represent pedestrian movement during stair evacuation. *Applied Math Modelling* 40.7 (2016): 4380-4396.

[3] Guanetti, J., Formentin, S., Corno, M., & Savaresi, S. M. (2015). Optimal energy management in series hybrid electric bicycles. In IEEE *Decision and Control (CDC), 2015 IEEE 54th*,869-874

[4] Corno, M., Berretta, D., Spagnol, P., Savaresi, S.M. (2016). Design, control, and validation of a charge-sustaining parallel hybrid bicycle. *IEEE Transactions on Control Systems Technology*, *24*(3), 817-829.

[5] Wan, N., Fayazi, S. A., Saeidi, H., Vahidi, A. (2014). Optimal power management of an electric bicycle based on terrain preview and considering human fatigue dynamics. IEEE *ACC* 3462-3467

[6] Algood, S.D., Cooper, R.A., Fitzgerald, S.G., Cooper, R. Boninger, M.L. (2004). Impact of a pushrim-activated power-assisted wheelchair on the metabolic demands, stroke frequency, and range of motion among subjects with tetraplegia. *Archives of physical medicine and rehabilitation*, *85*(11), 1865-1871.

[7] Fayazi, S.A., Wan, N., Lucich, S., Vahidi, A., Mocko, G. (2013). Optimal pacing in a cycling time-trial considering cyclist's fatigue dynamics. IEEE *American Control Conference (ACC)* 6442-6447

[8] Levy C.E., Chow, J.W. (2004). Pushrim-activated power-assist wheelchairs: elegance in motion. *American journal of physical medicine & rehabilitation*, *83*(2), 166-167.

[9] Mohammad, S., & Guerra, T. M. (2015). *U.S. Patent Application No. 15/311,769*.

[10] Mohammad, S., Guerra, T.M., Grobois, J.M., Hecquet, B. (2011). Heart rate control during cycling exercise using Takagi-Sugeno models. *IFAC Proceedings Volumes*, *44*(1), 12783-12788.

[11] Mohammad, S., Guerra, T.M., Grobois, J.M., Hecquet, B. (2012). Heart rate modeling and robust control during cycling exercise. Fuzz'IEEE *International Conference*. 1-8.

[12] Liu, J.Z., Brown, R.W., Yue, G.H. (2002). A dynamical model of muscle activation, fatigue, and recovery. *Biophysical journal*, *82*(5), 2344-2359.

[13] Burnley, M., Jones, A.M., Carter, H., Doust, J.H. (2000). Effects of prior heavy exercise on phase II pulmonary oxygen uptake kinetics during heavy exercise. *Journal Applied Physiology*, *89*(4), 1387-1396.

[14] Barstow, T.J., Mole, P.A. (1991). Linear and nonlinear characteristics of oxygen uptake kinetics during heavy exercise. *Journal of Applied Physiology*, *71*(6), 2099-2106.

[15] Deisenroth, M.P., Neumann, G., Peters, J. (2013). A survey on policy search for robotics. *Foundations & Trends® in Robotics*, *2*(1–2), 1-142.

[16] Peters, J., & Schaal, S. (2006, October). Policy gradient methods for robotics. In *Intelligent Robots and Systems, IEEE/RSJ* 2219-2225.

[17] Grondman, I., Busoniu, L., Lopes, G. A., Babuska, R. (2012). A survey of actor-critic reinforcement learning: Standard and natural policy gradients. *IEEE T. SMC, Part C 42*(6), 1291-1307.

[18] Kober, J., Peters, J.R. (2009) Policy search for motor primitives in robotics. *Advances neural information proc. systems* (pp. 849-856).

[19] Grondman, I., Vaandrager, M., Busoniu, L., Babuska, R., Schuitema, E. (2012). Efficient model learning methods for actor–critic control. *IEEE T. Systems, Man, Cybernetics, Part B (Cybernetics)*, *42*(3), 591-602.

[20] Tanohata, N., Murakami, H., Seki, H. (2010). Battery friendly driving control of electric power-assisted wheelchair based on fuzzy algorithm. IEEE *SICE Annual Conference 2010, Proceedings of* 1595-1598

[21] Buşoniu, L., Ernst, D., De Schutter, B., Babuška, R. (2010) Approximate dynamic programming with a fuzzy parameterization. *Automatica*, *46*(5), 804-814