SOFT ROBOT ASSISTED HUMAN NORMATIVE WALKING: REAL DEVICE CONTROL VIA REINFORCEMENT LEARNING WITHOUT A SIMULATOR

Anonymous authors
Paper under double-blind review

ABSTRACT

This study offers an innovative solution approach to soft robot-assisted human walking. The controller design of the soft robotic exosuit aims at assisting human normative walking with reduced human physical effort. Achieving such optimal interaction between the human and robot agents presents a key challenge to the robot control design due to a lack of robust model of the soft inflatable exosuit and its interaction dynamics with the human user. Moreover, to maximize user comfort, the robot assistance should be personalized to individual users. Toward this goal, we propose an offline to online based approach that is referred to as AIP, which stands for online Adaptation from an offline Imitating expert Policy. Our offline learning mimics human expert actions through real human walking demonstrations without robot assistance. The resulted policy is then used to initialize online reinforcement learning, the goal of which is to optimally personalize robot assistance. In addition to being fast and robust, our online actor-critic learning method also posseses important properties such as learning convergence, system stability, and solution optimality. We have successfully demonstrated our simple and robust solution framework for safe robot control on all four tested human participants.

1 INTRODUCTION

Goal of this study. Wearable robots such as rigid exoskeletons and soft exosuits have been extensively researched and have shown great promise for gait rehabilitation Rodríguez-Fernández et al. (2021) and for assisting human walking to reduce physical efforts Collins et al. (2015). Unlike rigid exoskeletons, soft, garment-like devices made from materials like silicone elastomers and fabrics provide a more comfortable, safer, and adaptable user experience Granberry et al. (2017); Bao et al. (2018); Thalman & Artemiadis (2020). Yet, effectively controlling the wearable robots to seamlessly work with human users in locomotion tasks remain a major challenge. This may be why deployment of the promising wearable technology still have limited success in real-world deployment. In this study, we directly address this soft exosuit control design problem in a real human-robot interactive environment. As a most promising solution approach, reinforcement learning points to two potential solutions: sim-to-real approach or direct design in the real physical environment. We address this important and challenging RL control problem by directly working with the physical process to avoid the elaborate and costly process of first building an accurate simulator of the human-robot physical environment. This is due to the following reasons: 1) It is without a doubt that the ultimate goal of almost all control problems is to implement the designs into physical devices or to influence the physical processes that involve real physical environments. Bypassing the step of first building a simulator will not only avoid the associated overhead, but also inherent modeling errors; 2) Devising a near-perfect simulator of the environment of interest is exceptionally costly if at all possible given the ubiquitous presence of noise, delay, other uncertainty in the system, and changes in use environment. A soft robot attached to a human user only exacerbates these challenges. **Challenges of controlling a soft wearable exosuit.** First off, soft inflatable exosuits lack a robust model of their dynamics Polygerinos et al. (2015a), not to mention modeling the interaction dynamics between the user and the robot, a necessary step in building a high fidelity simulator. These unique challenges stem from that the pneumatic dynamics of the soft inflatable actuators Joshi & Paik (2021) are complicated in part by the nonlinear nature of soft actuators due to material properties and design geometry. The fabric-based actuators result

in highly compliant behavior that enables high levels of deformations Hasan et al. (2022). The manufacturing process of the actuators also introduces significant variations and uncertainties Joshi & Paik (2021). Further wear and tear of the fabric only makes problem more complicated. Lacking a reliable model or simulator of the soft robot has made controlling a soft inflatable exosuit more complex than a traditional rigid exoskeleton Polygerinos et al. (2015a). Unlike rigid exoskeletons where the assistive torque is determined by motor actuators and can be directly used as a control parameter, for soft inflatables, the torque is generated from two collaborative sources: the human and the exosuit, which is nearly impossible to quantify. Additionally, the inflation/deflation of the actuators typically introduces longer actuation delays than motor-actuated exoskeleton, a factor that potentially reduces stability margins in control system design. **An offline-to-online approach to physical device control.** Our approach of online *Adaptation* from an offline *Imitating* expert *Policy* (AIP) provides a framework that enables optimal interaction between a soft inflatable exosuit and the human user. Our offline imitation learning (IL) is to mimic normative human walking via expert demonstrations Xu et al. (2022); Garcıa & Fernández (2015). This expert policy is expected to capture baseline walking dynamics and walking patterns amid the presence of inherent sensor and actuator noise, as well as uncertainties in the environment. With the help of our attention to address these uncertainties via a good data quality improvement method, this expert policy may serve as a best initial guess for online optimal policy, which is expected to be more efficient and effective than random policies learned from simulators Levine et al. (2020); Kumar et al. (2020). **Contributions of this study.** 1) We have devised and demonstrated a data-centric solution approach to the problem of soft robot assisted human normative walking with reduced human effort. We take a less travelled path of directly learning from the physical process without first building a simulator of the human-robot system. The soft robot control method developed in this study is capable of addressing all the challenges that demand safety, effectiveness, time efficiency, and adaptivity simultaneously. 2) The data-centric AIP solution takes full advantage of the physics of the human locomotion to directly capture the typical rhythmic characteristics of human locomotion in the natural physical environment using offline imitation learning, and then to use the offline expert policy as the initial policy for efficient and effective online adaptation and personalization of robot assistance. 3) Our online RL is not only empirically effective by providing robust control policies tailored to individual users, but also retains important qualitative properties such as learning convergence, system stability, and solution optimality.

2 Related Work

RL successes with and without simulated environments. The most celebrated reinforcement learning (RL) achievements with superhuman performance are in playing computer games Silver et al. (2014); Mnih et al. (2015). These successes are largely attributable to the use of unbiased simulation environments, which provide extensive and repeatable training data. However, the simulator-based successes have rarely been duplicated in the real physical world. High-fidelity simulators are often prohibitively expensive or even impossible to construct due to the complex dynamics, limitations in assessing and representing inherently uncertain physical systems, such as sensor and actuator noise, communication delays and other factors Rao et al. (2020); Niu et al. (2022). Nonetheless, RL has been directly applied to physical systems without the use of simulators. For example, Inoue et al. (2017) presents a method to enable industrial robots to perform high-precision assembly tasks (such as the peg-in-hole) by training an LSTM using reinforcement learning directly on the physical device. There have been some other successful demonstrations of RL agents interacting with simulated raw environments instead of simulators providing directly accessible state-action-reward data. For example, Hilleli & El-Yaniv (2018) trains RL agents for autonomous highway steering using raw image sequences from a simulated environment. The VPT Baker et al. (2022) is a semi-supervised imitation learning method, where an inverse dynamics model (IDM) is trained with labeled data to generate pseudo-labels for a vast amount of unlabeled online videos. This allows for training a behavioral prior that exhibits nontrivial zero-shot capabilities and can be fine-tuned using imitation learning and reinforcement learning to perform complex tasks. The method is shown to achieve significant results in the Minecraft Game, especially the crafting diamond tools, which were impossible for RL alone previously. The AR2-D2 Duan et al. (2023) allows users to record themselves manipulating objects, and the data is then used to train a real robot to perform similar tasks. Despite these efforts, these studies still depend on gathering large amounts of data, often requiring hundreds of hours of long sequences of video or episodes, for RL to effectively converge. There is a

significant gap in research addressing RL applications that operate under limited data conditions, particularly those with fewer than a few hundred state transitions. Another compounding factor that significantly complicates the problem in data-scarce environments is the human-in-the-loop effect, which is difficult to model or build a simulator for the human-robot interacting dynamics. These issues remain largely unexplored.

Imitation learning (IL) IL has been demonstrated to be a natural and effective part of reinforcement learning (RL) as it can be used as a reasonable initial policy Taylor et al. (2011). However, it may subject to distribution shift when applied in online environments. Ross et al. (2011); Spencer et al. (2021). Prior work addressing the issue typically falls into the following two categories. **Algorithm-centric approaches** aim to learn robust policies by imposing task-specific assumptions based on specific characteristics of the task Galashov et al. (2022); Guhur et al. (2023); James & Davison (2022), or acquiring additional data to model environment dynamics for the agent to return to in-distribution states Englert et al. (2013); Qi et al. (2022). Some approaches enhance action representation such as using Gaussian or mixture models to capture all expert actions Chi et al. (2023); Mandlekar et al. (2021). Others reduce the task length by employing temporal abstraction of the action spaces Shridhar et al. (2023); Zhao et al. (2023). However, growing evidence has shown the potential of substantial performance improvement in imitation learning by merely modifying the data collection process Belkhale et al. (2024). **Data-centric approaches** prioritize data quality, primarily aiming to maximize state diversity. Numerous studies focus on modifying data collection processes to expose the expert to a diverse set of state transitions through shared control Cui et al. (2019); Kelly et al. (2019); Ross et al. (2011). Some methods allow human intervention to correct robot behavior when necessary Gandhi et al. (2023); Mandlekar et al. (2020). Active learning guides data collection toward more informative samples by prioritizing questions that maximize information gain while minimizing the difficulty of selecting queriesBıyık et al. (2019); Cui & Niekum (2018). We focus on improving **data quality** to address our unique problem challenge within a data-centric framework to improve action divergence Belkhale et al. (2024) between the learned policy and the demonstration policy, thereby to improve task success rates. In doing so, we have two specific considerations: 1) to avoid long data collection processes, and 2) effectively deal with environmental noise.

Control of Soft Exosuit. A fundamental control challenge with wearable devices is modeling the interaction dynamics between the user and the robot for optimal coordination Polygerinos et al. (2015b); Nesler et al. (2018); O'Neill et al. (2022). Several studies have achieved successful coordination through human-in-the-loop optimization methods. In Siviy et al. (2020), offline optimization of a cable-driven ankle exosuit is performed to generate the assistive torque profile. In Ding et al. (2018), the authors perform human-in-the-loop optimization through a Bayesian optimization to identify the peak and offset timing of hip extension assistance with a cable-driven hip exosuit. In Kim et al. (2019b), the authors advance this framework by coupling Bayesian optimization with a Kalman filter metabolic estimator to deliver plantar flexion assistance to the ankle with a cable-driven ankle exosuit. In Li et al. (2022a), the authors developed a hierarchical human-in-the-loop controller of a cable-driven exosuit for impedance adaptation to different terrains. An offline cable control parameter optimization was developed in Li et al. (2022b), which relies on an impedance model based on the geometric relationship of ankle joint. While these studies have achieved coordination between the robot and the user, they have a strong prerequisite that the wearable robot possesses a robust dynamical model. In the presence of unknown dynamics of the robot, these human-in-the-loop optimization methods face a greater challenge to achieve optimal coordination.

For further details on issues related to the modeling of soft exosuits, control strategies for rigid lower limb exoskeletons and powered prostheses, as well as personalized control systems tailored to individual users, please refer to Appendix B.

3 METHOD

Our online *Adaptation* from an offline *Imitating* expert *Policy* (AIP) procedure consists of two main phases: 1) offline imitation learning from an expert demonstration to capture baseline normative walking policy, which is used to initialize online learning, and 2) personalized online training by fine-tuning the RL controller to achieve optimal interaction, thereby to minimize human effort during normative walking. The AIP approach as proposed is to address the following challenges: 1) learning from limited data due to involving human in experiments and human fatigue; 2) effectively handling

the uncertainties inherent in the physical environments due to sensor and actuator noise and delay, an issue that can be exacerbated as human-robot interact in real time and real life; and 3) ensuring human safety when learning with real robots.

3.1 Physical Setup

Refer to Figure 1, the AIP solution involves two main phases, offline imitation learning (Figure 1.a) and online personalized RL (Figure 1.b). In both phases, a participant walks on the treadmill at a constant speed of 1 m/s. The inertial measurement unit (IMU) sensors collect kinematic data while the electromyography (EMG) sensors measure muscle activity simultaneously. A Motion Capture (MoCap) System, which provides ground truth measurement of the human joint motion, is time synced with the IMU and EMG sensors in the offline phase and the ground truth walking profiles were used to train an offline human policy as an initial policy for online training. Details of the system design are shown in Appendix D

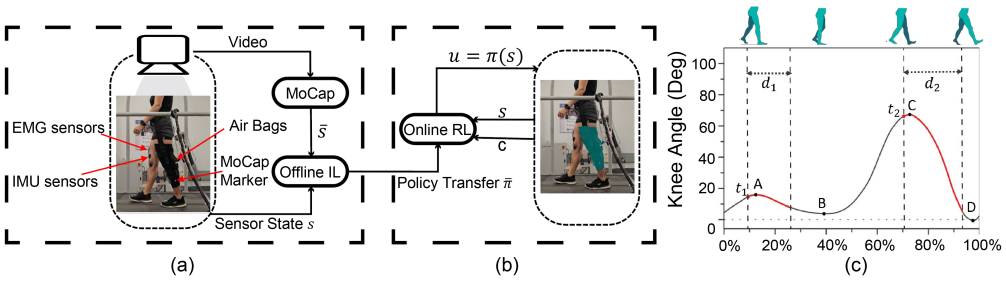

Figure 1: (a): Offline imitation learning using normative human walking data with ground truth provided by MoCap. The learned policy is then used to initialize online RL. (b): Online, personalized RL control of the soft exosuit to achieve human-robot normative walking while minimizing human effort measured by EMG activity. Sensor data are acquired via IMU for real-time control. (c): Knee angle profile of a complete gait cycle (in %) with the four gait phases as shown.

3.2 State and Control Variables

An analysis of knee joint kinematics reveals two critical regions associated with knee stance extension (from point A to B) and swing extension (from point C to D) during a gait cycle (Figure 1.c). The four extrema mark the transition from one gait phase to another: marker A is the maximum knee flexion during mid stance, B the maximum knee extension during terminal stance phase, C the maximum knee flexion during mid swing phase, and D the maximum knee extension adjacent to heel strike. These transition points and their related characteristics are therefore considered for inclusion in the state representation from the human walking profile. Specifically, the state variables include the peak knee flexion angle at point C as denoted by θ_f, the time instances t_A and t_C of peak knee flexion at the stance and swing phases, respectively, the duration of the stance phase (between point A and point B) defined as $d_A = t_B - t_A$, and the duration of the swing phase (between point C and point D) defined as $d_C = t_D - t_C$. Thus, we define the state variable as follows:

$$s = [t_A, d_A, t_C, d_C, \theta_f]^T. \tag{1}$$

Unlike rigid exoskeletons where the torque is generated by electrical motors and can be directly used as a control parameter, for soft inflatable actuators, the amount of assistive torque is determined by both human knee torque and the actuator pressure, the two collaborative sources. It is therefore not feasible to use torque directly as the control variable for the exosuit. Instead, only properly timed inflation and deflation of the exosuit will provide the necessary and optimal assistance to the human user (Figure 1.c). On the contrary, if the exosuit is not properly operated, it may cause discomfort or even injury to the human user. Toward this end, the RL controller must determine the optimal timings to operate the exosuit and these control parameters are:

$$u = [t_1, d_1, t_2, d_2]^T, \tag{2}$$

where t_1 is the onset timing of inflation of the exosuit to assist stance extension, the duration for which the air pressure is maintained during this phase is d_1. Similarly, t_2 represents the onset timing

of inflation of the exosuit to assist swing flexion, and the corresponding duration for maintaining air pressure during this phase is given by d_2. As it takes time for the exosuit to inflate and deflate, it is expected that an optimal RL controller should successfully learn the optimal timings of t_1 and t_2, which are expected to be close to or ahead of the maximum flexion timings t_A and t_C. By precisely adjusting these timings and durations, the RL controller ensures that the exosuit provides optimal assistance to the user's knee movement, enhancing overall gait efficiency and reducing muscle effort quantified by EMG measurement (EMG Effort). Note that, it is natural to maintain consistency in the state and action spaces during both offline and online learning.

3.3 Safety Constraints

To ensure human participants walk continuously and safely, we consider several safety constrains: 1) the actuator pressure is limited to 206.8 kpa; 2) the control timings and inflation/deflation durations are constrained by taking reference of those during participant's normative walking as shown in Table 3, which are within realistic ranges Zhang et al. (2020b). These physical constraints help prevent significant misalignment between controller timing and the respective gait phase during human walking. Without these constraints, it may trigger soft actuator deployment and cause discomfort or injury to the user; and 3) the online training objective is set for the control timings to approach those during normative walking, and thus in a safe state. Details of the safety constraints and their physical representations are in Appendix C. Additionally, we provide a theoretical performance analysis of the online learning process to ensure learning convergence, optimal timing solution, and human-robot interaction dynamic stability under reasonable conditions and within these safety constraints (Appendix F).

3.4 Offline Human Normative Walking Policy

Our imitation learning approach utilizes Behaviour Cloning (BC) Torabi et al. (2018); Bain & Sammut (1995); Daftry et al. (2017) to derive an effective imitation policy based on data $\overline{\mathcal{D}} = \{\overline{s}(k)|, k = 1, 2, ..., N\}$, obtained from normative walking demonstrations under natural walking condition of a human participant, where N represents the total number of gait cycles over which the state variable data is collected from the MoCap system. In this study, we aim to demonstrate the generalizability and data efficiency of our method by ONLY collecting offline walking data from a SINGLE participant with $N = 150$. Detailed information about the offline data collection process can be found in Appendix E

A good offline normative walking policy should serve the following two purposes. First, it provides a reasonable initial policy for online tuning tailored for individual users while both offline and online learning are subject to similar environmental uncertainties such as sensor and actuator noise. As such, this offline policy helps online policy tuning to be kept in a reasonable and meaningful range. Second, this offline learned policy should capture key human locomotion characteristics even under intra- and inter-person variations Zhang et al. (2020a); Ahn & Hogan (2012) as human locomotion (such as knee angle) exhibits similar patterns as shown in Figure (1.c).

Improve data quality. Based on most recent results that data-centric approaches have greater impact than algorithm-centric approaches on the effectiveness of imitation learning Belkhale et al. (2024), we aim to improve data quality and expect that to be especially effective in addressing our unique challenges associated with the human-robot system under study. First note that, typical data-centric techniques, such as collecting more data, diversifying state transition, actively learning human walking dynamics, or human intervention of natural and normative walking, are entirely unfeasible for easy to understand reasons. We therefore propose a reducing intra-person and inter-person variation (RIIV) method to improve measured data quality as it is likely to be the one and the most effective and efficient approach. As a result, we compare our RIIV with a common benchmark approach that normalizes the raw measurements of state variables into the [-1,1] range.

Specifically, for each gait cycle of length T, let the original measurements of each state variable of s, as in Equation (1), be denoted by ζ. The following computations are performed component-wise for each of the state variables ($j = 1, 2, ..., 5$).

1) The benchmark DIRECT method normalizes the raw sensor measurements (ζ) of states into [-1,1] by the following procedure,

$$s = 2\left(\frac{\zeta - min(\zeta)}{max(\zeta) - min(\zeta)}\right) - 1. \tag{3}$$

2) The RIIV method.

The first step of RIIV reduces intra-person step length variations by converting gait timing from actual time into gait percentage by normalizing over a gait cycle T, that is

$$\xi = \frac{\zeta}{T}. \tag{4}$$

The second step reduces inter-person variation by transferring state variables into the range of $[-1, 1]$,

$$s = 2\left(\frac{\xi - \inf(\xi)}{\sup(\xi) - \inf(\xi)}\right) - 1, \tag{5}$$

where the values of inf(ξ) and sup(ξ) are from established studies of biomechanics literature Zhang et al. (2020b), which is shown in Table 5 in Appendix C.

Imitation policy. Once real time measurements for offline policy training are obtained during normative human walking, BC is utilized to train an offline imitation human walking policy $\pi(s_k)$, which maps human state from IMU sensors to control timings and durations of normative walking with the ground truth provided by MoCap, namely, $\overline{\pi} = \{\overline{t}_A, \overline{d}_A, \overline{t}_C, \overline{d}_C\} \in \overline{\mathcal{D}}$. We use "action divergence" to measure offline cost $\overline{c}_k$ in BC learning,

$$\overline{c}_k = \frac{1}{2}(\pi(s_k) - \overline{\pi}_k)^2. \tag{6}$$

Therefore the actor with policy parameter (ϕ) minimizes a supervised loss as:

$$L(\phi) = \frac{1}{N}\sum_{k=0}^{N} \overline{c}_k. \tag{7}$$

which is the distance between the RL policy and that used in human demonstration.

3.5 Reinforcement Learning for Online, Personalized Exosuit Control

Personalizing soft robot control for individuals face the following **Offline to online learning challenges**: 1) Data out of distribution (OOD) due to inter- and intra-human variance; 2) Limited availability of human walking data; and 3) Hardware limitations including communication delays, sensor noise, and significant delay in actuation. Our solution relies on a good data quality improvement procedure and an efficient online reinforcement learning algorithm.

Once a reliable initial controller is established, the online training phase commences. During this phase, the RL controller is fine-tuned through direct interaction with the human subject. This personalized training process adapts the controller to the specific needs and characteristics of the user, ultimately minimizing their muscle effort by improving the effectiveness of the exosuit assistance. In this online training phase, we consider two necessary performance metrics: gait normalcy thus safety constraint, and muscle effort.

In this paper, we employed an established policy gradient algorithm, the direct heuristic dynamic programming (dHDP) Si & Wang (2001), that has been successfully demonstrated in wearable robotics research Wen et al. (2017c; 2019; 2017a;b) and other rather significant real-time control applications Enns & Si (2003); Lu et al. (2008). Unlike traditional RL formulations that aim to maximize the expected reward, in wearable robotics, the objective is to minimize the overall cost over policy π, defined as follows:

$$Q^{\pi}(s_k, u_k) = \mathbb{E}[\sum_{t=k}^{\infty} \gamma^{t-k} c_t | s_k, u_k], \tag{8}$$

where $s_k \sim p(\cdot \mid s_{k-1}, u_{k-1})$, $u_k = \pi(s_k)$, $c_k = c(s_k, u_k)$ is the stage cost, and the discount factor $0 < \gamma < 1$. The stage cost c_k in the above is formulated to take into consideration of two important

performance measures in online learning of personalized optimal policy to achieve robot-assisted normative walking with reduced energy expenditure.

First, we embed normative walking and safety constraint ϵ^s as one of the important performance considerations in the performance index (Equation 10). Specifically, $\epsilon^s = (s - \tilde{s})^2$, where target state $\tilde{s}$ is defined in Appendix C, and is extracted from the offline normative walking profile using MoCap data as in Table 3. Additionally, ϵ^s is bounded within safety constraint provided in Table 4. It ensures that the subject does not deviate significantly from the target, thereby preventing potential falls or discomfort. The **second** consideration of reducing human energy expenditure is reflected by reduced muscle activity, which is measured per gait cycle, namely, the EMG effort ϵ^e is determined by

$$\epsilon^e = \frac{1}{2}(\sum_{t=0}^{T} f_E(t))^2 \tag{9}$$

which $f_E(t)$ is the EMG sensor value at time t of a gait and $\sum_{t=0}^{T} f_E(t)$ simulates the integral of the EMG signal under a complete gait cycle of length T.

We thus have the stage cost c_k formulated by balancing the reduction of EMG effort and adherence to state error tolerance and safety constraints, and it is consequently used in formulating the total cost in Equation (8):

$$c_k = \epsilon^s + \epsilon^e. \tag{10}$$

The dHDP is then used to provide online learning of a personalized optimal policy for individual users. Further details about dHDP, its actor and critic network realizations, and its implementation can be found in Appendices E and F.

3.6 Qualitative Properties of the Learning Process and Control Performance Associated with dHDP Online Learning

In this study, we provide a theoretical analysis to characterize properties of the learning process and the control performance, specifically those related to learning convergence, solution optimality, and control system stability as a result of online dHDP learning initialized by an offline policy obtained via imitation learning. We obtain these results under reasonable conditions. Details are provided in Appendix F.

4 Results and Analysis

This study of directly learning to control a physical device, the robotic exosuit, to assist normative human walking aims at exploring the feasibility of RL in achieving stable and efficient learning without a simulator. As a result, we have shown promising first steps in addressing key challenges of RL control for real life applications. Furthermore, our AIP as a data-centric, offline to online approach reveals its practical usefulness to address environment uncertainty due to variations in human, sensor and actuator noise and delay that are unavoidable in real physical environments.

Participants. Four healthy individuals (2 males and 2 female) participated in the study under a protocol approved by the Institutional Review Board. The complete anthropometric data of the subjects and IRB can be seen in Appendix A. During experimentation, the soft exosuit is strategically attached to a human leg behind the knee (the popliteal fossa area) as such placement maximizes the assistive benefits while minimizing any potential discomfort or interference with the user's natural gait. Further details about online learning algorithm implementation, hyper parameters, measured data processing, and convergence criteria are provided in Appendix E. Additional details on the placement of the soft inflatable exosuit, its manufacturing, wearable sensors, etc. can be found in Appendix D.

Performance Criteria. The results reported in this study were based on the following performance metrics: 1) The stage cost as shown in Equation 10 to reflect online learning performance; 2) Peak knee error as a kinematic measure of normative walking and also to reflect walking safety; 3) EMG activity (Equation 9) which reflects human effort during walking; 4) Time to convergence of RL online learning (influencing human physical fatigue); 5) "Action divergence" to measure offline

policy optimality as in Equation (6). For all the metrics, better performance is associated with smaller/shorter outcomes.

Questions Addressed. Our real experimental results aim at answering the following questions:
1) Is RIIV an effective method for improving IMU sensor data quality in our data-centered solution framework?
2) Can offline normative human walking policy be further adapted and customized for individual participants via robot online learning to achieve optimal human-robot interaction?
3) Can online learning address significant delays (not present in offline learning) in the inflatable actuators, that were not present in offline learning, while maintaining safety and optimal interaction between the user and the robot?
4) Is there evidence that both human and robot co-adapted to achieve optimal interaction
5) To achieve optimal interaction between the human and the robot, what are essential control cost objectives to be considered in RL design?

Performance Evaluations	Beginning of Online Training (Offline IL policy)				End of Online Training			
Human Participant	1	2	3	4	1	2	3	4
Stage Cost	0.94 ± 0.16	0.95 ± 0.23	1.1 ± 0.47	1.42 ± 0.57	0.43 ± 0.06	0.49 ± 0.12	0.37 ± 0.06	0.35 ± 0.08
Training Peak knee error	N/A	N/A	N/A	N/A	0.39 ± 0.05	0.23 ± 0.04	0.16 ± 0.14	0.37 ± 0.27
Training EMG Effort	N/A	N/A	N/A	N/A	0.54 ± 0.03	0.59 ± 0.13	0.43 ± 0.09	0.57 ± 0.14
Evaluation Peak knee error	0.48 ± 0.09	0.33 ± 0.06	0.28 ± 0.05	0.4 ± 0.2	0.41 ± 0.02	0.23 ± 0.04	0.18 ± 0.18	0.38 ± 0.07
Evaluation EMG Effort	0.66 ± 0.01	0.96 ± 0.3	0.55 ± 0.09	0.8 ± 0.27	0.52 ± 0.05	0.32 ± 0.02	0.42 ± 0.07	0.38 ± 0.02

Table 1: Performance of AIP method in terms of stage cost, peak knee error, and EMG effort.

Q1: (Offline Benchmark Study) Our RIIV method is practically effective in capturing invariant normative walking characteristics while directly accounting for sensor and actuator noise in real environments, thereby improving offline policy optimality or action divergence. From Figure 2, we can clearly see advantages of using RIIV procedure over the Direct method (Section 3.4) to process raw IMU sensor data. 1) Firstly, RIIV results in significantly lower training cost and faster convergence than the Direct Method. As illustrated in the four bar charts in Figure 2, the RIIV method (green bar) reduces the action divergence effect more greatly than the Direct Method (orange bar) does, indicating that RIIV more accurately aligns with true human walking characteristics. 2) Next, RIIV has shown to be capable of accounting for significant uncertainties inherent in physical sensing and actuation, as demonstrated by the green bar with its values closer to the ground truth, especially for t_A, t_c, and d_c, where there are notable discrepancies between raw sensor data and the ground truth, and also, a rather significant delay in the actuator due to inflation/deflation time.

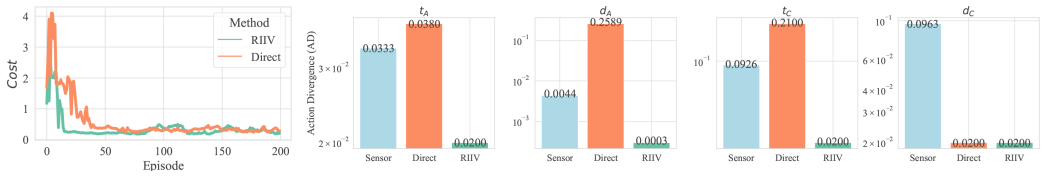

Figure 2: Offline learning outcomes as evidence of the essential role of processing raw sensor measurements in AIP as a data-centric method. (Left): Comparison of cost performance, Equation (7), using Direct and RIIV, respectively along an offline training episode. (Right 4 panels): The MoCap data is used as ground truth in the comparisons, where action divergence (AD) as in Equation (6) was measured (the closer to 0 the better.): "blue" is AD between IMU sensed data and the truth; "orange" is AD between Direct and Truth; and "green" is AD between RIIV and Truth.

Q2: Online learning effectively adapted the initial offline policy to provide personalized control for individual participants and enable robust performance in human-robot normative walking.

From Table 1, although the offline policy enables walking, it does not achieve optimal performance in terms of cost, kinematic error, and EMG measures. As shown in Figure 3 and Table 1, while the offline policy directly benefits participants 2 and 3 in terms of reduced EMG effort (below baseline shown by dashed line), it fails to do so for participants 1 and 4. Through online training, Performance metrics improve for reduced cost and kinematic error, and most importantly, reduced EMG effort for all subjects. Notice additionally that online training resulted in consistent and robust assistance to human walking. From Figure 3 and Table 1, a significant intra-subject variance and inter-subject

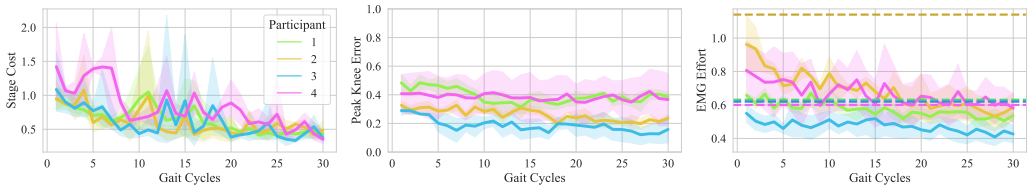

Figure 3: Results of online training for all four participants where the shaded regions indicate the 95% confidence interval for the three online trials. The dashed lines are respectively the baseline human walking EMG effort without exosuit assistance. Participant 1 provided the offline policy.

variance is apparent. At the initial online learning stage (gait cycle 1), the same offline policy produced varying performances across different participants. However, by the end of training, the cost consistently converged to similar values of around 0.5, which indicates that online training has effectively customized the initial offline policy for each individual, allowing all participants to reach normative walking patterns with at least a 20% reduction in EMG effort.

Q3: (Addressing Out-of-Distribution Issue) Online adaptation of physical device control successfully overcame significant actuator delays, which is a key factor causing out-of-distribution issue from offline to online learning, and did so without compromising user safety.

A primary challenge during online training was caused by a significant actuator delay associated with soft actuator inflation and deflation. Specifically, there is an approximately 0.2-second delay to fully inflate and 0.25-second delay to deflate. These delays could not be adequately captured during the offline imitation learning phase as offline policy was obtained without exosuit control. As shown in Figure 4, to compensate for the inflation delay, the control variable t_1 was significantly shifted to an earlier onset, allowing the system to anticipate the slower actuator response time. Similarly, to mitigate the impact of deflation delays, the duration variables d_1 and d_2 were substantially shortened, ensuring that the system could maintain synchronization with the human walking pattern. These adjustments were critical in aligning the actuator responses with the real-time dynamics of human movement, thereby enhancing the overall effectiveness of AIP by achieving normative walking with reduced effort while all safety constraints are met.

Q4: Human and robot co-adapted to achieve normative walking with reduced human EMG Effort

1) Refer to Table 1, online training of robot control has led to normative walking, as measured by the peak knee angle approaching that during normative walking (small peak knee error), and reduced EMG effort for all participants. This is a result of online co-adaptation between the human and the robot. To see that, we show next how robot control has taken effect by looking into measurable human walking states. 2) Let's examine the duration of human stance phase (d_A) and swing phase (d_C) before and after online learning and note that the respective duration has changed little (refer to the top row of Figure 4 above the bar charts and Figure 7.a & b). This is because the participants walk naturally and thus maintains their normative walking patterns. 3) In the meantime, note that the robot has reduced its stance duration (d_1) and swing duration (d_2) to accommodate soft actuator deployment delays (refer to the top row of Figure 4 above the bar charts). 4) Next, if we inspect the robot control onset timing t_1 for stance, it varied around human stance timing t_A (refer to the bar plots in Figure 4). As the soft actuators are to provide leg support for stance, the human responses could vary depending on how they weigh the importance of reducing effort during this less effort demanding phase of walking. 5) The swing phase soft actuator onset timing t_2, however, has adapted to be ahead of the human actual start of swing t_C. Inflating actuators in this phase is critical to reduce human effort of lifting the leg and swing it forward. Note that the initial policy from offline learning also resulted in an even earlier swing onset t_2, an outcome that may be caused by out-of-distribution effect as there was no soft actuator deployment during offline training. Consequently, to accommodate soft actuator delays there has to be an early onset, but cannot be too early so that the soft actuator is in the way of a normal knee swing flexion (reduce the peak knee angle error).

Q5 (Ablation Study) Both safe regulation of joint kinematics and reduction of human EMG effort are necessary to achieve stable human-robot normative walking. We performed an ablation

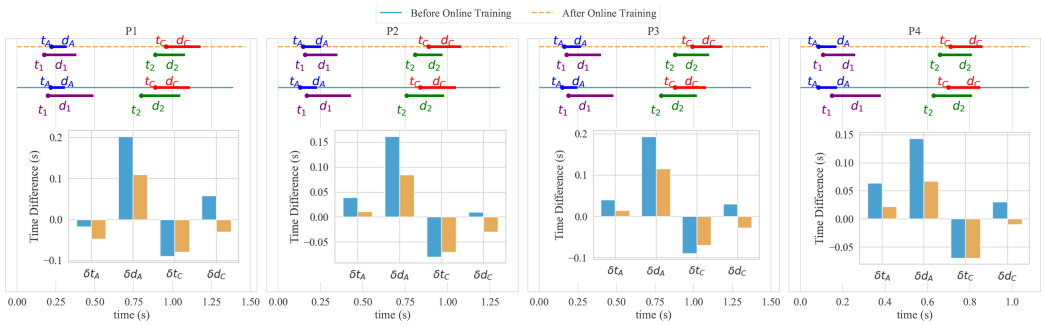

Figure 4: Timing and duration in state and control variables to demonstrate adaptation taking place during online learning. In the top panel above the bar charts, the blue line segment is d_A, the red is d_C, the purple is d_1 and the green is d_2. The bar plots show the mean differences in timing and duration between respective actual human walking measurements and those of the robot control. Specifialy, $\delta t_A = t_1 - t_A$, $\delta d_A = d_1 - d_A$, $\delta t_C = t_2 - t_C$, and $\delta d_C = d_2 - d_C$.

study on the cost objective function as shown in Appendix G. Our proposed performance index, which incorporates both EMG effort and kinematic error or state error, demonstrates superior performance. By balancing the reduction of EMG effort and adherence to state error tolerance and safety constraints, the RL controller optimizes both aspects of the user's walking behavior. Refer to Figure 3, this balanced approach leads to convergence, stability, and significant improvements in the user's mobility, as evidenced by lower stage cost, peak knee error, and decreased EMG effort. However, if only EMG effort or Kinematic error was used in Equation 10, not only the EMG did not reduce but also it resulted in a significant learning variance. The absence of state kinematic error in the cost function resulted in failure to maintain normative walking patterns, which led to increased EMG levels and overall less effective assistance (Figure 6). The absence of EMG effort in the performance index leads to a lack of focus on reducing muscle activity. Consequently soft exosuit failed to provide the necessary support to reduce muscular strain, resulting in increased EMG levels (Figure 5).

5 Conclusion

This study presents an innovative solution AIP for the real-time soft robot control to provide personalized assistive walking with a goal to reduce human physical effort during normative walking. Our RL-based, data-centric approach is conceptualized, implemented, and demonstrated directly in the physical environment. Without a human-robot interactive dynamic model and a simulator to provide meaningful data, achieving optimal interaction between human and robot is particularly challenging. Our AIP learning solution shows a promising first step that may shed light on future studies addressing real life RL control applications.

Our AIP is a data-centric approach. The RIIV method is shown effective to improve data quality. Our solution framework, that includes RIIV, offline imitation learning and online dHDP learning, has shown effective to personalize individual assistance in a data-efficient manner. We have successfully demonstrated our simple and robust solution framework for safe robot control on all four tested human participants, providing robust control policies tailored to individual users and leading to reduced EMG effort.

In conclusion, the proposed AIP framework offers a viable and effective solution for personalized robotic assistance in human locomotion. The co-adaptation between the human and the robot to address actuator delays has resulted in reduced muscular effort, highlighting the synergistic interaction within the human-robot system. This work opens new avenues for personalized robotic assistance in rehabilitation and performance enhancement, contributing to more efficient and less physically demanding walking. Future research could focus on scaling this approach to a larger and more diverse population of users, integrating more complex locomotion patterns, and exploring the long-term adaptation and learning capabilities of the system. Finally, information about the code for this study is in Appendix E

REFERENCES

Jooeun Ahn and Neville Hogan. Walking is not like reaching: evidence from periodic mechanical perturbations. *PloS one*, 7(3):e31767, 2012.

Alejandro F Azocar, Luke M Mooney, Jean-François Duval, Ann M Simon, Levi J Hargrove, and Elliott J Rouse. Design and clinical implementation of an open-source bionic leg. *Nature biomedical engineering*, 4(10):941–953, 2020.

Michael Bain and Claude Sammut. A framework for behavioural cloning. In *Machine Intelligence 15*, pp. 103–129, 1995.

Bowen Baker, Ilge Akkaya, Peter Zhokov, Joost Huizinga, Jie Tang, Adrien Ecoffet, Brandon Houghton, Raul Sampedro, and Jeff Clune. Video pretraining (vpt): Learning to act by watching unlabeled online videos. *Advances in Neural Information Processing Systems*, 35:24639–24654, 2022.

Guanjun Bao, Hui Fang, Lingfeng Chen, Yuehua Wan, Fang Xu, Qinghua Yang, and Libin Zhang. Soft robotics: Academic insights and perspectives through bibliometric analysis. *Soft robotics*, 5 (3):229–241, 2018.

Suneel Belkhale, Yuchen Cui, and Dorsa Sadigh. Data quality in imitation learning. *Advances in Neural Information Processing Systems*, 36, 2024.

Erdem Bıyık, Malayandi Palan, Nicholas C Landolfi, Dylan P Losey, and Dorsa Sadigh. Asking easy questions: A user-friendly approach to active reward learning. *arXiv preprint arXiv:1910.04365*, 2019.

Magdo Bortole, Anusha Venkatakrishnan, Fangshi Zhu, Juan C Moreno, Gerard E Francisco, Jose L Pons, and Jose L Contreras-Vidal. The h2 robotic exoskeleton for gait rehabilitation after stroke: early findings from a clinical study. *Journal of neuroengineering and rehabilitation*, 12:1–14, 2015.

Gwendolyn M Bryan, Patrick W Franks, Seungmoon Song, Ricardo Reyes, Meghan P O'Donovan, Karen N Gregorczyk, and Steven H Collins. Optimized hip-knee-ankle exoskeleton assistance reduces the metabolic cost of walking with worn loads. *Journal of neuroengineering and rehabilitation*, 18:1–12, 2021.

Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *arXiv preprint arXiv:2303.04137*, 2023.

Steven H Collins, M Bruce Wiggin, and Gregory S Sawicki. Reducing the energy cost of human walking using an unpowered exoskeleton. *Nature*, 522(7555):212–215, 2015.

Yuchen Cui and Scott Niekum. Active reward learning from critiques. In *2018 IEEE international conference on robotics and automation (ICRA)*, pp. 6907–6914. IEEE, 2018.

Yuchen Cui, David Isele, Scott Niekum, and Kikuo Fujimura. Uncertainty-aware data aggregation for deep imitation learning. In *2019 International Conference on Robotics and Automation (ICRA)*, pp. 761–767. IEEE, 2019.

Shreyansh Daftry, J Andrew Bagnell, and Martial Hebert. Learning transferable policies for monocular reactive mav control. In *2016 International Symposium on Experimental Robotics*, pp. 3–11. Springer, 2017.

Ye Ding, Myunghee Kim, Scott Kuindersma, and Conor J. Walsh. Human-in-the-loop optimization of hip assistance with a soft exosuit during walking. *Science Robotics*, 3(15):eaar5438, 2018. doi: 10.1126/scirobotics.aar5438. URL `https://www.science.org/doi/abs/10.1126/scirobotics.aar5438`.

Jiafei Duan, Yi Ru Wang, Mohit Shridhar, Dieter Fox, and Ranjay Krishna. Ar2-d2: Training a robot without a robot. *arXiv preprint arXiv:2306.13818*, 2023.

Peter Englert, Alexandros Paraschos, Marc Peter Deisenroth, and Jan Peters. Probabilistic model-based imitation learning. *Adaptive Behavior*, 21(5):388–403, 2013.

Russell Enns and Jennie Si. Helicopter trimming and tracking control using direct neural dynamic programming. *IEEE Transactions on Neural networks*, 14(4):929–939, 2003.

Alexandre Galashov, Josh S Merel, and Nicolas Heess. Data augmentation for efficient learning from parametric experts. *Advances in Neural Information Processing Systems*, 35:31484–31496, 2022.

Kanishk Gandhi, Siddharth Karamcheti, Madeline Liao, and Dorsa Sadigh. Eliciting compatible demonstrations for multi-human imitation learning. In *Conference on Robot Learning*, pp. 1981–1991. PMLR, 2023.

Xiang Gao, Jennie Si, and He Huang. Reinforcement learning control with knowledge shaping. *IEEE Transactions on Neural Networks and Learning Systems*, 35(3):3156–3167, 2024.

Javier Garcıa and Fernando Fernández. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(1):1437–1480, 2015.

Rachael Granberry, Julia Duvall, Lucy E Dunne, and Bradley Holschuh. An analysis of anthropometric geometric variability of the lower leg for the fit & function of advanced functional garments. In *Proceedings of the 2017 ACM international symposium on wearable computers*, pp. 10–17, 2017.

Pierre-Louis Guhur, Shizhe Chen, Ricardo Garcia Pinel, Makarand Tapaswi, Ivan Laptev, and Cordelia Schmid. Instruction-driven history-aware policies for robotic manipulations. In *Conference on Robot Learning*, pp. 175–187. PMLR, 2023.

Ibrahim Mohammed Hasan, Emiliano Quinones Yumbla, and Wenlong Zhang. Development of a soft inflatable exosuit for knee flexion assistance. In *2022 9th IEEE RAS/EMBS International Conference for Biomedical Robotics and Biomechatronics (BioRob)*, pp. 1–6, 2022. doi: 10.1109/BioRob52689.2022.9925474.

Bar Hilleli and Ran El-Yaniv. Toward deep reinforcement learning without a simulator: An autonomous steering example. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

He Helen Huang, Jennie Si, Andrea Brandt, and Minhan Li. Taking both sides: seeking symbiosis between intelligent prostheses and human motor control during locomotion. *Current opinion in biomedical engineering*, 20:100314, 2021.

Tadanobu Inoue, Giovanni De Magistris, Asim Munawar, Tsuyoshi Yokoya, and Ryuki Tachibana. Deep reinforcement learning for high precision assembly tasks. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 819–825. IEEE, 2017.

Stephen James and Andrew J Davison. Q-attention: Enabling efficient learning for vision-based robotic manipulation. *IEEE Robotics and Automation Letters*, 7(2):1612–1619, 2022.

Sagar Joshi and Jamie Paik. Pneumatic supply system parameter optimization for soft actuators. *Soft Robotics*, 8(2):152–163, 2021. doi: 10.1089/soro.2019.0134. URL https://doi.org/10.1089/soro.2019.0134. PMID: 32598232.

Michael Kelly, Chelsea Sidrane, Katherine Driggs-Campbell, and Mykel J Kochenderfer. Hg-dagger: Interactive imitation learning with human experts. In *2019 International Conference on Robotics and Automation (ICRA)*, pp. 8077–8083. IEEE, 2019.

Jinsoo Kim, Giuk Lee, Roman Heimgartner, Dheepak Arumukhom Revi, Nikos Karavas, Danielle Nathanson, Ignacio Galiana, Asa Eckert-Erdheim, Patrick Murphy, David Perry, et al. Reducing the metabolic rate of walking and running with a versatile, portable exosuit. *Science*, 365(6454):668–672, 2019a.

Myunghee Kim, Charles Liu, Jinsoo Kim, Sangjun Lee, Adham Meguid, Conor J. Walsh, and Scott Kuindersma. Bayesian optimization of soft exosuits using a metabolic estimator stopping process. In *2019 International Conference on Robotics and Automation (ICRA)*, pp. 9173–9179, 2019b. doi: 10.1109/ICRA.2019.8793817.

Jeffrey R Koller, Deanna H Gates, Daniel P Ferris, and C David Remy. 'body-in-the-loop'optimization of assistive robotic devices: A validation study. In *Robotics: Science and Systems*, volume 2016, pp. 1–10, 2016.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 33:1179–1191, 2020.

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

Minhan Li, Yue Wen, Xiang Gao, Jennie Si, and He Huang. Toward expedited impedance tuning of a robotic prosthesis for personalized gait assistance by reinforcement learning control. *IEEE Transactions on Robotics*, 2021.

Zhijun Li, Xiang Li, Qinjian Li, Hang Su, Zhen Kan, and Wei He. Human-in-the-loop control of soft exosuits using impedance learning on different terrains. *IEEE Transactions on Robotics*, 38(5): 2979–2993, 2022a. doi: 10.1109/TRO.2022.3160052.

Zhijun Li, Xiang Li, Qinjian Li, Hang Su, Zhen Kan, and Wei He. Human-in-the-loop control of soft exosuits using impedance learning on different terrains. *IEEE Transactions on Robotics*, 38(5): 2979–2993, 2022b.

Yi Long, Zhijiang Du, Lin Cong, Weidong Wang, Zhiming Zhang, and Wei Dong. Active disturbance rejection control based human gait tracking for lower extremity rehabilitation exoskeleton. *ISA transactions*, 67:389–397, 2017.

Chao Lu, Jennie Si, and Xiaorong Xie. Direct heuristic dynamic programming for damping oscillations in a large power system. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 38(4):1008–1013, 2008.

Ajay Mandlekar, Danfei Xu, Roberto Martín-Martín, Yuke Zhu, Li Fei-Fei, and Silvio Savarese. Human-in-the-loop imitation learning using remote teleoperation. *arXiv preprint arXiv:2012.06733*, 2020.

Ajay Mandlekar, Danfei Xu, Josiah Wong, Soroush Nasiriany, Chen Wang, Rohun Kulkarni, Li Fei-Fei, Silvio Savarese, Yuke Zhu, and Roberto Martín-Martín. What matters in learning from offline human demonstrations for robot manipulation. *arXiv preprint arXiv:2108.03298*, 2021.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei a Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, Feb 2015. ISSN 0028-0836. doi: 10.1038/nature14236.

Christopher R Nesler, Tim A Swift, and Elliott J Rouse. Initial design and experimental evaluation of a pneumatic interference actuator. *Soft robotics*, 5(2):138–148, 2018.

Haoyi Niu, Yiwen Qiu, Ming Li, Guyue Zhou, Jianming Hu, Xianyuan Zhan, et al. When to trust your simulator: Dynamics-aware hybrid offline-and-online reinforcement learning. *Advances in Neural Information Processing Systems*, 35:36599–36612, 2022.

Ciarán T O'Neill, Connor M McCann, Cameron J Hohimer, Katia Bertoldi, and Conor J Walsh. Unfolding textile-based pneumatic actuators for wearable applications. *Soft Robotics*, 9(1):163–172, 2022.

Panagiotis Polygerinos, Zheng Wang, Johannes T. B. Overvelde, Kevin C. Galloway, Robert J. Wood, Katia Bertoldi, and Conor J. Walsh. Modeling of soft fiber-reinforced bending actuators. *IEEE Transactions on Robotics*, 31(3):778–789, 2015a. doi: 10.1109/TRO.2015.2428504.

Panagiotis Polygerinos, Zheng Wang, Johannes TB Overvelde, Kevin C Galloway, Robert J Wood, Katia Bertoldi, and Conor J Walsh. Modeling of soft fiber-reinforced bending actuators. *IEEE Transactions on Robotics*, 31(3):778–789, 2015b.

Carl Qi, Pieter Abbeel, and Aditya Grover. Imitating, fast and slow: Robust learning from demonstrations via decision-time planning. *arXiv preprint arXiv:2204.03597*, 2022.

Kanishka Rao, Chris Harris, Alex Irpan, Sergey Levine, Julian Ibarz, and Mohi Khansari. Rl-cyclegan: Reinforcement learning aware simulation-to-real. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11157–11166, 2020.

Antonio Rodríguez-Fernández, Joan Lobo-Prat, and Josep M. Font-Llagunes. Systematic review on wearable lower-limb exoskeletons for gait training in neuromuscular impairments. *Journal of NeuroEngineering and Rehabilitation*, 18:22, 12 2021. ISSN 1743-0003. doi: 10.1186/s12984-021-00815-5. URL https://jneuroengrehab.biomedcentral.com/articles/10.1186/s12984-021-00815-5.

Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 627–635. JMLR Workshop and Conference Proceedings, 2011.

Di Shi, Wuxiang Zhang, Wei Zhang, and Xilun Ding. A review on lower limb rehabilitation exoskeleton robots. *Chinese Journal of Mechanical Engineering*, 32(1):1–11, 2019.

Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Perceiver-actor: A multi-task transformer for robotic manipulation. In *Conference on Robot Learning*, pp. 785–799. PMLR, 2023.

Jennie Si and Yu-Tsung Wang. Online learning control by association and reinforcement. *IEEE Transactions on Neural networks*, 12(2):264–276, 2001.

David Silver, Guy Lever, Nicolas Heess, Thomas Degris, Daan Wierstra, and Martin Riedmiller. Deterministic policy gradient algorithms. In *International conference on machine learning*, pp. 387–395. Pmlr, 2014.

Christopher Siviy, Jaehyun Bae, Lauren Baker, Franchino Porciuncula, Teresa Baker, Terry D. Ellis, Louis N. Awad, and Conor James Walsh. Offline assistance optimization of a soft exosuit for augmenting ankle power of stroke survivors during walking. *IEEE Robotics and Automation Letters*, 5(2):828–835, 2020. doi: 10.1109/LRA.2020.2965072.

Christopher Siviy, Lauren M Baker, Brendan T Quinlivan, Franchino Porciuncula, Krithika Swaminathan, Louis N Awad, and Conor J Walsh. Opportunities and challenges in the development of exoskeletons for locomotor assistance. *Nature Biomedical Engineering*, 7(4):456–472, 2023.

Jonathan Spencer, Sanjiban Choudhury, Arun Venkatraman, Brian Ziebart, and J Andrew Bagnell. Feedback in imitation learning: The three regimes of covariate shift. *arXiv preprint arXiv:2102.02872*, 2021.

S. Sridar, S. Poddar, Y. Tong, P. Polygerinos, and W. Zhang. Towards untethered soft pneumatic exosuits using low-volume inflatable actuator composites and a portable pneumatic source. *IEEE Robotics and Automation Letters*, 5(3):4062–4069, 2020. doi: 10.1109/LRA.2020.2986744.

Matthew Edmund Taylor, Halit Bener Suay, and Sonia Chernova. Using human demonstrations to improve reinforcement learning. In *2011 AAAI Spring Symposium Series*, 2011.

Carly Thalman and Panagiotis Artemiadis. A review of soft wearable robots that provide active assistance: Trends, common actuation methods, fabrication, and applications. *Wearable Technologies*, 1:e3, 2020.

Faraz Torabi, Garrett Warnell, and Peter Stone. Behavioral cloning from observation. *arXiv preprint arXiv:1805.01954*, 2018.

Xikai Tu, Minhan Li, Ming Liu, Jennie Si, and He Helen Huang. A data-driven reinforcement learning solution framework for optimal and adaptive personalization of a hip exoskeleton. In *2021 IEEE international conference on robotics and automation (ICRA)*, pp. 10610–10616. IEEE, 2021.

Yue Wen, Andrea Brandt, Ming Liu, He Huang, and Jennie Si. Comparing parallel and sequential control parameter tuning for a powered knee prosthesis. In *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 1716–1721. IEEE, 2017a.

Yue Wen, Andrea Brandt, Jennie Si, and He Helen Huang. Automatically customizing a powered knee prosthesis with human in the loop using adaptive dynamic programming. In *2017 International Symposium on Wearable Robotics and Rehabilitation (WeRob)*, pp. 1–2. IEEE, 2017b.

Yue Wen, Jennie Si, Xiang Gao, Stephanie Huang, and He Helen Huang. A new powered lower limb prosthesis control framework based on adaptive dynamic programming. *IEEE Transactions on Neural Networks and Learning Systems*, 28(9):2215–2220, 2017c.

Yue Wen, Jennie Si, Andrea Brandt, Xiang Gao, and He Huang. Online reinforcement learning control for the personalization of a robotic knee prosthesis. *IEEE transactions on cybernetics*, 2019.

Ruofan Wu, Zhikai Yao, Jennie Si, and He Helen Huang. Robotic knee tracking control to mimic the intact human knee profile based on actor-critic reinforcement learning. *IEEE/CAA Journal of Automatica Sinica*, 9(1):19–30, 2021.

Haoran Xu, Li Jiang, Li Jianxiong, and Xianyuan Zhan. A policy-guided imitation approach for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 35:4085–4098, 2022.

Juanjuan Zhang, Pieter Fiers, Kirby A Witte, Rachel W Jackson, Katherine L Poggensee, Christopher G Atkeson, and Steven H Collins. Human-in-the-loop optimization of exoskeleton assistance during walking. *Science*, 356(6344):1280–1284, 2017.

Li Zhang, Geng Liu, Bing Han, Zhe Wang, Yuzhou Yan, Jianbing Ma, and Pingping Wei. Knee joint biomechanics in physiological conditions and how pathologies can affect it: a systematic review. *Applied bionics and biomechanics*, 2020(1):7451683, 2020a.

Li Zhang, Geng Liu, Bing Han, Zhe Wang, Yuzhou Yan, Jianbing Ma, and Pingping Wei. Knee joint biomechanics in physiological conditions and how pathologies can affect it: A systematic review. *Applied Bionics and Biomechanics*, 2020:1–22, 4 2020b. ISSN 1176-2322. doi: 10.1155/2020/7451683.

Tony Z Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. *arXiv preprint arXiv:2304.13705*, 2023.

Mengjia Zhu, Shantonu Biswas, Stejara Iulia Dinulescu, Nikolas Kastor, Elliot Wright Hawkes, and Yon Visell. Soft, wearable robotics and haptics: Technologies, trends, and emerging applications. *Proceedings of the IEEE*, 110(2):246–272, 2022.

A Participant Information and IRB approve

Four healthy individuals (2 males and 2 female, participated in the study under a protocol approved by the Institutional Review Board (IRB ID#: STUDY00011110) The average height, weight, and age of the recruited participants were 163 ± 8 cm, 66.1 ± 11.6 kg, and 28 ± 1.9 years, respectively. The complete anthropometric data of the subjects can be seen in Table 2.

Table 2: Subject participants' anthropometric data.

Subject	Gender	Age	Weight (kg)	Height (m)
S1	M	26	76	1.75
S2	F	27	52	1.54
S3	M	28	79	1.65
S4	F	31	57.5	1.58

B Additional Related Work

Modeling of soft exosuit A pertinent example of these challenges can be seen in the development and application of soft inflatable exosuits. On one hand, modeling a simulator for these devices is highly complex Polygerinos et al. (2015b). This is mainly due to the compliant nature of soft robots, which introduce physical properties that are difficult to model. For instance, several studies have shown that even obtaining a quasi-static model of torque for a soft actuator is not trivial Nesler et al. (2018); O'Neill et al. (2022). On the other hand, the dynamic interaction between the human wearer and the robot creates a highly coupled and complex system, making it even more challenging to model accurately Zhu et al. (2022).

Control of rigid lower limb exoskeleton and powered prosthesis. Rigid robotic lower-limb exoskeletons and prostheses and their controls are being actively researched or even commercially available Huang et al. (2021); Siviy et al. (2023); Shi et al. (2019). A typical **control strategy** of these devices often focus on mimicking the kinematics of biological joints via position control Bortole et al. (2015); Long et al. (2017). However, another control strategy, referred to as finite state machine impedance control, is often preferred especially for consideration of achieving compliant lower limb behaviors. This stratagy provides safe human-exoskeleton interactions, as biological systems are capable of in order to adapt to various environments Azocar et al. (2020). Unfortunately, neither of the above two strategies are applicable to soft exosuit. For **control actuation** in either control strategy, rigid device control torques are generated by mechanical joint motor actuators, which can be directly used as a control parameter. In soft inflatables, however, the torque is generated from both human and the exosuit. Differentiating the two sources is difficult or nearly impossible. The natural control parameters for exosuit instead is the timing of inflation and deflation, which introduces additional delays to actuation and thus reduced stability margins have to be considered in the control design.

Personalized controls for individual users are common to all wearable lower limb devices, and human-in-the-loop (HIL) optimization represents several important design approaches Koller et al. (2016); Zhang et al. (2017); Ding et al. (2018); Kim et al. (2019a); Bryan et al. (2021). They are used in open or closed-loop force or torque control to operate individual joints. In those applications, Bayesian optimization plays a key role to provide optimal controls. These methods search for an extremum on the system response surface to determine the optimal control parameters. Although these methods can customize the control strategies or parameters, they are time-consuming and lack of adaptation. A small change of the wearer requires a re-design of the control Tu et al. (2021). It is noted that a large class of wearable rigid device controls aims at achieving reducing **metabolic cost** reflected by oxygen intake. However, it usually takes long walking time to be able to extract reliable measurements. This prohibits online and real time requirement that is highly desired for wearable robot applications. To overcome the limitations due to control strategy or optimizaiton method, data-driven **reinforcement learning** (RL)-based optimal adaptive control methods have been developed and successfully demonstrated for robot control of exoskeletons and prostheses. Wen et al. (2019; 2017c); Wu et al. (2021); Li et al. (2021). In these applications, robot control relies on optimal cumulative cost/reward related to producing normative walking using directly measurable human-robot walking variables. However, these methods have been principally implemented in rigid robot devices Huang et al. (2021), not soft inflatable exosuits, the control problem formulation is different and the solution presents unique challenges due to discussions in the above.

C Safety Constraints

The problem under investigation requires human physical safety and control system stability of the human-robot system. In this study, physical safety refers to that the human participants do not fall or endure injury as a result of robot control. This is ensured by imposing safety bounds to limit the soft suit inflation and inflating duration timing. The control system stability is in the same classical control sense, and we set one of the RL design objective in Equation 11 that the state regulation errors approach 0 (or practically error tolerance bounded). Our assurance of stability and safety is embedded in learning quantitatively, and guaranteed by analysis qualitatively. Our systematic data have shown that these constraints are met and objectives achieved.

To ensure human participants walk continuously and safely, we consider several safety constrains: 1) the soft actuator pressure is limited to 206.8 kpa; 2) the control timings and inflation/deflation duration

are constrained by taking reference of those during participant's normative walking as shown in Table 3, which are within realistic ranges Zhang et al. (2020b). These physical constraints help prevent significant misalignment between controller timing and the respective gait phase during human walking. Without these constraints, it may trigger soft actuator deployment and cause discomfort to the user; and 3) the online training objective is set for the control timings to approach those during normative walking, and thus in a safe state. Equation 11 renders such state constraint where the target state variables $\tilde{s} = [\tilde{t}_A, \tilde{d}_A, \tilde{t}_C, \tilde{d}_C, \tilde{\theta}]$ (Table 3) are obtained from normative walking profile without soft actuator deployment to assist human walking:

Control	t_1	d_1	t_2	d_2	Target State	$\tilde{t}_A$	$\tilde{d}_A$	$\tilde{t}_C$	$\tilde{d}_C$	$\tilde{\theta}$
Safety constraints (% gait phase)	[0,20]	[0,20]	[60,80]	[0,20]	normal walking (% gait phase)	14%	15%	68%	15%	60^o

Table 3: Safety constraint for the control timing and target state for control regulation to reach.

$$\epsilon_s = (s - \tilde{s})^2, \tag{11}$$

where the respective error tolerance for each state variable is as shown in Table 4. They represent realistic sensing and actuation errors inherent in physical systems, and they are physically meaningful, human physiologically realistic, and validated in studies of human biomechanics such as Zhang et al. (2020b).

Error Tolerance	$t_A - \tilde{t}_A$	$d_B - \tilde{d}_B$	$t_C - \tilde{t}_C$	$d_C - \tilde{d}_C$	$\theta - \tilde{\theta}$
Tolerance Range	[-5%,5%]	[-5%,5%]	[-5%,5%]	[-5%,5%]	[0, 40 deg]

Table 4: Ranges of state error tolerance that are used in learning for achieving normative walking.

State	t_A	d_B	t_C	d_C	θ
inf value	10 %	10%	60%	10%	53^o
sup value	20 %	30%	75%	30%	78^o

Table 5: Tolerance values that ensure human normative walking.

D HARDWARE DETAILS

Soft inflatable actuators are designed to generate extension torque, fabricated from nylon fabric and thermoplastic polyurethane to ensure a transparent interaction with the user. When the knee is flexed and the actuators are inflated, they apply extension torque to the knee joint. These actuators are strategically positioned in the popliteal fossa to aid in knee extension. Further design details of the soft inflatable exosuit can be found in Sridar et al. (2020).

The electro-pneumatic system that controls the real-time inflation of the exosuit includes:
a microcontroller (Raspberry Pi),
solenoid valves (MHE3-MS1H, Festo, Hauppauge, NY) for switching between inflation and deflation,
pressure sensors (ASDXAVX100PGAA5, Honeywell International Inc., Morris Plains, NJ) for monitoring the internal pressure of the actuators.

Kinematic data were collected using:
a camera-based motion capture system (T40s, VICON Inc., Los Angeles, CA) sampling at 100 Hz.
EMG sensors (Delsys Trigno, Delsys, Natick, MA) were used to capture muscle activity, sampled at 2000 Hz.
An instrumented treadmill (Bertec Inc., Columbus, OH) was used as the platform for the walking trials. The treadmill is equipped with force plates that measure the user's ground reaction forces at a sample rate of 2000 kHz.
An IMU motion capture system (Ultium Motion, Noraxon Inc., Scottsdale, AZ) was used to detect the maximum knee joint angle.

EMG sensors were placed on both legs over three muscles of interest: *vastus lateralis* (VL), *biceps femoris* (BF) and *rectus femoris* (RF). The raw EMG data were first band-pass filtered (Butterworth, 4th order, 20 Hz and 450 Hz cutoff frequencies). The profile of the signal was obtained by computing the root-mean-square envelope using a moving window of 250 ms. The integral of the envelope was computed for each gait cycle to quantify the overall muscular effort.

E EXPERIMENTATION, HYPERPARAMETERS AND IMPLEMENTATION DETAILS

We use PyTorch for all implementations. All results were obtained using our desktop with Intel Core i9-12900K processor.**Experimentation.** The experiments consisted of two sets of walking: an offline, normative walking session with the exosuit attached but not inflated, and an online walking session with RL-controlled exosuit inflation and deflation. Data of one participant was recorded during offline walking for one experiment session. The offline session lasted around 10 minutes of about 170 steps. For the online human-robot collaborative walking, three sessions were performed for each participant. Each online session began with the controller initialized to the learned offline policy from IL, and lasted 10 minutes of about 150 steps. All participants walked at a constant speed of 1 m/s on the treadmill during all experimental sessions, **Safety constraints** were imposed as discussed in Section 3.3. **Data Collection for Offline Training.** After a two-minute warm-up period for participant 1 to get accustomed to the experimental setup and walking speed, data collection began. MoCap video data of the state variables s in Equation (1) were collected. MoCap data were synchronized with the time sequences of the state variables to provide control target for offline training. **Data Collection for Online Training.** After a two minute warm up period, the learned offline policy (refer to Section 3.4) was used for initializing the online RL controller for all participants. To mitigate environment noise and intra-human variance, a consecutive 5 steps were used to obtain one gait sample, resulting in a total of 30 gait samples. The RL policy update was performed for every gait sample. **Performance Evaluations.** Evaluation sessions were performed after online learning convergence upon meeting criteria (Table 4). Each participant rested for 5 minutes after online learning prior to evaluation which involves walking of 100 steps in about 6 minutes. Evaluation data were processed using similar procedures to those used in processing online learning data.

E.1 OFFLINE TRAINING PROCEDURE

The offline training consist with 200 episode. An Episode start with the first offline data in the dataset $\overline{\mathcal{D}}$ to the end of the dataset with total data points of 150. For each training trial, we use an off-policy exploration strategy, adding Gaussian noise $\mathcal{N}(0, 0.05)$ to each control. The algorithm hyperparameter for offline training is as Table 6.

Hyperparameter	Value
Exploration noise	$\mathcal{N}(0, 0.05)$
Noise clip	± 0.5
Policy update frequency	2
Batch size	32
Buffer size	200
γ	0.95
τ	0.1
Adam Learning rate	0.001

Table 6: Hyper Parameters used for offline training

E.2 ONLINE TRAINING PROCEDURE

For the online human-robot collaborative walking, three sessions were performed for each participant. Each online session began with the controller initialized to the learned offline policy from IL, and lasted 10 minutes of about 150 steps. All participants walked at a constant speed of 1 m/s on the treadmill during all experimental sessions, Safety constraints were imposed as discussed in Section 3.3. The algorithm hyperparameter for offline training is as Table 7.

Hyperparameter	Value
Exploration noise	$\mathcal{N}(0, 0.01)$
Noise clip	± 0.1
Policy update frequency	2
Batch size	5
Buffer size	20
γ	0.95
τ	0.4
Adam Learning rate	0.001

Table 7: Hyper Parameters used for online training

E.3 Network Structure and optimizer

The actor-critic networks in DHDP are implemented by feedforward neural networks with two layers of weights. Each layer has 256 hidden nodes with rectified linear units (ReLU) for both the actor and critic. The input layer of actor has the same dimension as observation state. The output layer of the actor has the same dimension as action requirement with a tanh unit. Critic receives both state and action as input to THE first layer and the output layer of critic has 1 linear unit to produce Q value. Network parameters are updated using Adam optimizer with a learning rate of 10^{-3}.

E.4 Code

All the code will be provided to GitHub once the paper get accepted

F dHDP solution and properties

To find thd dHDP solution, let the critic value as Equation 8 be Q_θ where θ denotes the critic weights that are to be learned by using dHDP. Specifically, weight updates were performed to minimize the loss as a function of the weights (θ):

$$L(\theta) = \mathbb{E}_{s\sim p_\pi, u\sim\pi}\left[(y - Q_\theta(s_k, u_k))^2\right], \tag{12}$$

where in the above, y denotes the critic target. Accordingly, the actor weights (denoted by (ϕ)) are updated by applying the chain rule to the total return from the start distribution J with respect to the policy parameter (ϕ):

$$\nabla_\phi J(\phi) = \mathbb{E}_{s\sim p_{\pi_\phi}}\left[\nabla_u Q_\theta(s_k, u_k)|_{u_k=\pi_\phi(s_k)} \nabla_\phi \pi_\phi(s_k)\right]. \tag{13}$$

The update rules for the critic and the actor, respectively are:

$$\begin{aligned} \theta &\leftarrow \theta + \alpha\nabla_\theta L(\theta), \\ \phi &\leftarrow \phi + \alpha\nabla_\phi J(\phi), \end{aligned} \tag{14}$$

where α is the learning rate.

Here we analyze and characterize properties of the learning process and the control performance, specifically those related to learning convergence, solution optimality, and stability as a result of online dHDP learning. In the following, we express the exosuit control system with the following general nonlinear dynamics for the ease of discussion although this model is unknow, and our offline to online learning is completely data-driven.

$$s_{k+1} = f(s_k, u_k), k = 0, 1, ... \tag{15}$$

where $s \in \mathbb{R}^5$ and $u \in \mathbb{R}^4$ are defined in Equations 1, 2, respectively, k denotes discrete time steps.

The objective of optimal control is to find a control policy that can stabilize system (15) and minimize the cost-to-go in Equation (8).

According to the Bellman optimality principle, the optimal cost-to-go satisfies the following relationship,

$$Q^*(s_k, u_k) = c_k(s_k, u_k) + \gamma Q^*(s_{k+1}, \pi^*(s_{k+1})), \tag{16}$$

and the optimal control law π^* can be expressed as

$$\pi^*(s_k) = arg\min_{u_k} Q^*(s_k, u_k), \tag{17}$$

where $Q^*(s_k, u_k)$ is the state-action value function corresponding to the optimal control policy $\pi^*(s_k)$.

We need the following definition and assumption to develop our results.

Definition 1. (Stabilizable System) A nonlinear dynamical system is said to be stabilizable on a compact set $\Omega \in \mathbb{R}^n$, if for all initial states $s_0 \in \Omega$, there exists a control sequence $u_0, u_1, \ldots, u_k, \ldots$ such that the state $s_k \to s^e$ as $k \to \infty$ where s^e is a equilibrium point.

Assumption 1. *System (15) is controllable and stabilizable. The system state $s_k = s^e$ is an equilibrium of the system under the control $u_k = \pi(s_k) = u^e$ for $s_k = s^e$, i.e., $f(s^e, u^e) = s^e$. The feedback control sequence u_k is determined from control policy π represented by the actor neural network, and in the most general case is bounded by actuator saturation.*

Assumption 2. *The stage cost function $c_k(s_k, u_k)$ is finite, continuous in s_k and u_k, and positive semi-definite with $c_k(s_k, u_k) = 0$ if and only if $s_k = s^e$ and $u_k = u^e$.*

Note that the above assumptions are reasonable and realistic, as they are under the presumptions that a person who can use a exosuit to assist walking can reach an equilibrium state that they can achieve normative walking while their muscle activities are reduced to a level less than that without wearing assistance.

As a actor-critic method, dHDP solve the Bellman's optimality by learning to approximate both policy and value functions where actor refers to the learned policy and critic refers to the learned value. An actor-critic algorithm starts with an initial value, e.g., $Q_0(s, u) = 0$ and an initial arbitrary policy π_0. Then for $i = 0, 1, 2, ...$, it iterates between policy update and policy evaluation steps.

$$Q_{i+1}(s_k, u_k) = c_k(s_k, u_k) + \gamma Q_i(s_{k+1}, \pi_i(s_{k+1})), \tag{18}$$

and

$$\pi_i(s_k) = \arg\min_{u_k} Q_i(s_k, u_k). \tag{19}$$

Or by combining (18) and (19), we have

$$Q_{i+1}(s_k, u_k) = c_k(s_k, u_k) + \gamma \min_{u_{k+1}} Q_i(s_{k+1}, u_{k+1}). \tag{20}$$

Theorem 1. *Let Assumptions 1 and 2 hold. Let Q_i be the sequence of estimated Q values starting from $Q_0 = 0$ at ith update of RL agent. For policy π_i, its actor network weights are updated based on the policy gradient estimator (14), and the controls are bounded by the output function of the action network. Then*

(1) ***Bounded:*** *there is an upper bound Y such that $0 \le Q_i(s_k, u_k) \le Y$, for $i = 1, 2, ...$*

(2) Q_i is a non-decreasing sequence satisfying $Q_i(s_k, u_k) \le Q_{i+1}(s_k, u_k), \forall i.$

(3) ***Convergence:*** *the limit of the sequence, $Q_\infty(s_k, u_k) = \lim_{i\to\infty} Q_i(s_k, u_k)$, satisfies*

$$Q_\infty(s_k, u_k) = c_k(s_k, u_k) + \gamma \min_{u_{k+1}} Q_\infty(s_{k+1}, u_{k+1}). \tag{21}$$

(4) ***Optimality:*** *the Q-value sequence $Q_i(s_k, u_k)$ and the corresponding policy $\pi_i(s_k)$, with $\pi_\infty(s_k) = \lim_{i\to\infty} \pi_i(s_k)$, converge to the optimal value Q^* and optimal policy π^*, respectively:*

$$\pi_\infty(s_k) = \pi^*(s_k), \tag{22}$$

$$Q_\infty(s_k, u_k) = Q^*(s_k, u_k). \tag{23}$$

Proof. (1) Let $\eta(s_k)$ be a deterministic control policy represented by a neural network which is a continuous mapping from s_k in stochastic environment E. Let $Z_0(\cdot) = 0$, and Z_i be updated by

$$Z_{i+1}(s_k, u_k) = c_k(s_k, u_k) + \gamma Z_i(s_{k+1}, \eta(s_{k+1})), \tag{24}$$

Thus, $Z_1(s_k, u_k) = c_k(s_k, u_k)$.

According to Lemma 2 in Gao et al. (2024), we obtain

$$\begin{aligned} &Z_{i+1}\left(s_k, u_k\right) \\ &= \sum_{j=0}^{i} \gamma^j c_k\left(s_{k+j}, \eta\left(s_{k+j}\right)\right) \leq \sum_{j=0}^{\infty} \gamma^j c_k\left(s_{k+j}, \eta\left(s_{k+j}\right)\right). \end{aligned} \tag{25}$$

If Assumption 1 holds, $c_k(s_{k+j}, u_{k+j})$ is bounded, there exists an upper bound Y such that

$$\sum_{j=0}^{\infty} \gamma^j c_k\left(s_{k+j}, \eta\left(s_{k+j}\right)\right) \leq Y, \tag{26}$$

According to Lemma 1 in Gao et al. (2024), as Q_{i+1} is the result of minimizing the right-hand side of (20), we have

$$Q_{i+1}\left(s_k, u_k\right) \leq Z_{i+1}\left(s_k, u_k\right) \leq Y, \forall i. \tag{27}$$

(2) Define a value sequence Φ_i as

$$\Phi_{i+1}\left(s_k, u_k\right) = c_k\left(s_k, u_k\right) + \gamma \Phi_i\left(s_{k+1}, \pi_{i+1}\left(s_{k+1}\right)\right), \tag{28}$$

and $\Phi_0 = Q_0 = 0$. In the following, a shorthand notation is used for $\Phi_i(s_{k+1}, \pi_{i+1}) = \Phi_i(s_{k+1}, \pi_{i+1}(s_{k+1}))$.

Since $\Phi_0\left(s_k, u_k\right) = 0$ and $Q_1\left(s_k, u_k\right) = c_k(s_k, u_k)$, and c_k is positive semi-definite under Assumption 2,

$$\Phi_0\left(s_k, u_k\right) \leq Q_1\left(s_k, u_k\right). \tag{29}$$

From (18) and (28), we get

$$\begin{aligned} &Q_{i+1}\left(s_k, u_k\right) - \Phi_i\left(s_k, u_k\right) \\ &\qquad = \gamma\left[Q_i\left(s_{k+1}, \pi_i\right) - \Phi_{i-1}\left(s_{k+1}, \pi_i\right)\right] \geq 0. \end{aligned} \tag{30}$$

Therefore,

$$\Phi_i\left(s_k, u_k\right) \leq Q_{i+1}\left(s_k, u_k\right). \tag{31}$$

Further by using Lemma 1 in Gao et al. (2024)

$$Q_i\left(s_k, u_k\right) \leq \Phi_i\left(s_k, u_k\right) \leq Q_{i+1}\left(s_k, u_k\right). \tag{32}$$

This completes the proof of Theorem 1 (2).

(3) From parts (1) and (2) in the above, Q_i is a monotonically non-decreasing sequence with an upper bound. Therefore, its limit exists. Let the limit be $\lim_{i\to\infty} Q_i\left(s_k, u_k\right) = Q_\infty\left(s_k, u_k\right)$.

Given i and for any u_{k+1}, according to (18), there is

$$Q_i\left(s_k, u_k\right) \leq c_k\left(s_k, u_k\right) + \gamma Q_{i-1}\left(s_{k+1}, u_{k+1}\right). \tag{33}$$

As Q_i is monotonically non-decreasing, we have

$$Q_{i-1}\left(s_k, u_k\right) \leq Q_\infty\left(s_k, u_k\right), \tag{34}$$

the following then holds

$$Q_i\left(s_k, u_k\right) \leq c_k\left(s_k, u_k\right) + \gamma \min_{u_{k+1}} Q_\infty\left(s_{k+1}, u_{k+1}\right). \tag{35}$$

As $i \to \infty$, we have

$$Q_\infty\left(s_k, u_k\right) \leq c_k\left(s_k, u_k\right) + \gamma \min_{u_{k+1}} Q_\infty\left(s_{k+1}, u_{k+1}\right). \tag{36}$$

On the other hand, since the cost-to-go function sequence satisfies

$$Q_{i+1}\left(s_k, u_k\right) = c_k\left(s_k, u_k\right) + \gamma \min_{u_{k+1}} Q_i\left(s_{k+1}, u_{k+1}\right), \tag{37}$$

applying inequality (34) as $i \to \infty$,

$$Q_\infty\left(s_k, u_k\right) \geq c_k\left(s_k, u_k\right) + \gamma \min_{u_{k+1}} Q_\infty\left(s_{k+1}, u_{k+1}\right). \tag{38}$$

Based on (36) and (38), (21) is true. This completes the proof of Theorem 1 (3).

(4) According to Theorem 1 (3) and by using Equations (18) and (19), we have

$$\begin{aligned} Q_\infty(s_k, u_k) &= c_k(s_k, u_k) + \gamma \min_{u_{k+1}} Q_\infty(s_{k+1}, u_{k+1}) \\ &= c_k(s_k, u_k) + \gamma Q_\infty(s_{k+1}, \pi_\infty(s_{k+1})), \end{aligned} \tag{39}$$

and

$$\pi_\infty(s_k) = \arg\min_{u_k} Q_\infty(s_k, u_k). \tag{40}$$

Observing (39) and (40), and then (16) and (17), we can find that (22) and (23) are true. This completes the proof of Theorem 1 (4).

Theorem 2. *Let Assumptions 1 and 2 hold, and Q_i be the sequence of estimated Q values starting from $Q_0 = 0$. For policy π_i, its actor network weights are updated based on the policy gradient estimator (14). If Q_i converges to Q_∞ as $\pi_i \to \pi_\infty$, then π_∞ is a stabilizing policy.*

Proof. If Assumption 1 holds, let $\mu(s_k)$ be a stabilizing control policy, and let its cost-to-go Λ_i be updated by the following equation from $\Lambda_0(\cdot) = 0$,

$$\Lambda_{i+1}(s_k, u_k) = c_k(s_k, u_k) + \gamma \Lambda_i(s_{k+1}, \mu(s_{k+1})), \tag{41}$$

We have

$$\Lambda_i(s_k, u_k) = \sum_{j=0}^{i} \gamma^j R(s_{k+j}, \mu(s_{k+j})), \tag{42}$$

Because $\mu(s_k)$ is a stabilizing policy, if Assumption 1 and 2 holds, we have $s_k \to s^e$ and $c_k(s_k, u_k) \to 0$ as $k \to \infty$. Therefore, $\Lambda_i(s_k, u_k) \to 0$ as $k \to \infty$.

Next, from Lemma 1 in Gao et al. (2024), π_i minimizes Q_i, we have

$$Q_i(s_k, u_k) \le \Lambda_i(s_k, u_k). \tag{43}$$

Since $\Lambda_i(s_k, u_k) \to 0$ as $k \to \infty$, we have $Q_i(s_k, u_k) \to 0$ as $k \to \infty$.

From Theorem 1 (3), we obtain $c_k(s_k, u_k) = 0$ as $k \to \infty$. Further, under Assumption 2, $c_k(s_k, u_k) = 0$ if and only $s_k = s^e$, we have $s_k \to s^e$ as $k \to \infty$. This completes the proof.

G Ablation Cost Study

Figure 5 illustrates the training performance when EMG effort is excluded from the performance index (Equation 10). The control results exhibit significant variance across different trials. Furthermore, both stage cost and peak knee error show little improvement, and the EMG effort even increases, indicating that the participants did not benefit from the exosuit. This outcome suggests that merely mimicking the human walking pattern without considering EMG effort is insufficient for improving walking performance. The absence of EMG effort in the performance index leads to a lack of focus on reducing muscle activity, which is crucial for enhancing comfort and efficiency in assisted walking. Consequently, the soft suit fails to provide the necessary support to reduce muscular strain, resulting in increased EMG levels and overall less effective assistance.

On the other hand, Figure 6 illustrates the training performance when the state error tolerance is excluded from the performance index (Equation 10). Without it, the RL controller explores the action space without considering normative walking behavior, resulting in divergence and large variance in the control policy. It is thus not surprising that learning performance metrics, such as stage cost and peak knee error, and EMG effort, remain high throughout the training. This indicates that although the RL controller explores a wide range of the action space, it fails to identify an improved policy that reduces these learning metrics. The absence of state error of kinematics in the cost function leads to a failure to maintain normative walking patterns. As a result, the soft suit does not provide the necessary support to reduce muscular strain, leading to increased EMG levels and overall less effective assistance.

H Human Adaptation

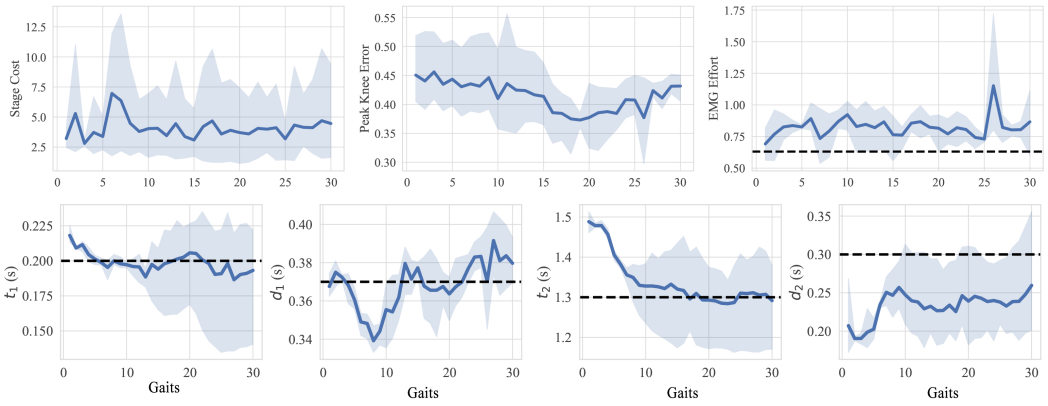

Figure 5: Learning performance without considering EMG effort in the control objective function. The shaded regions represent the 95 % confidence range of the three experiment trials. The x-axis is the number of gaits. The black dash lines are the reference baseline from human normative walking profiles.

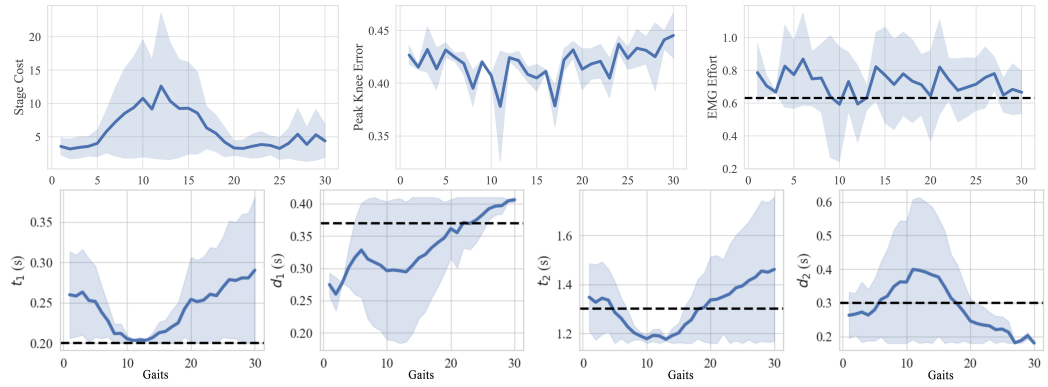

Figure 6: Learning performance without considering knee kinematic errors in the control objective function. The shaded regions represent the 95 % confidence range of the three experiment trials. The x-axis is the number of gaits. The black dash lines are the reference baseline from human normative walking profiles.

Table 8: Anthropometric data of our participants. S5 is the newly added participant

Subject	Gender	Age	Weight (kg)	Height (m)
S1	M	26	76	1.75
S2	F	27	52	1.54
S3	M	28	79	1.65
S4	F	31	57.5	1.58
S5	M	28	80	1.72

Figure 7: Characteristic timings and durations of gait trials: a) Raw gait data in seconds during training. b) Respective RIIV-processed data. c) RL policy during training. The shaded regions represent the 95 % confidence range of the three experiment trials. The x-axis is the number of gaits.

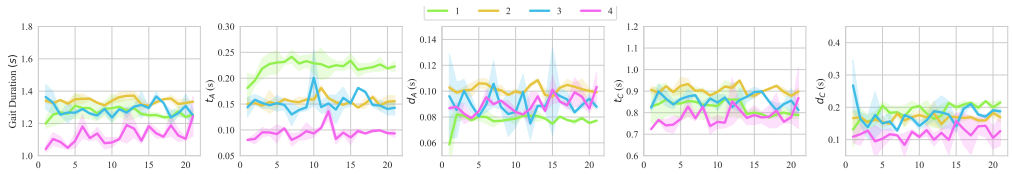

Figure 8: Characteristic timings and durations of gait trials under a fixed policy (i.e., no policy update). This is to observe how human users adapt to biological torques generated from the exosuit. The shaded regions represent the 95 % confidence range of the three experiment trials. The x-axis is the number of gaits.

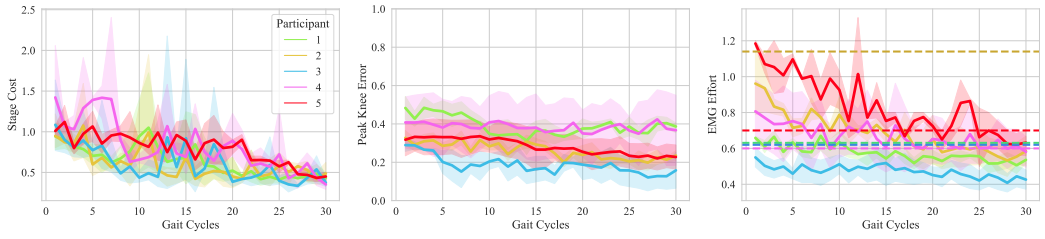

Figure 9: Results of online training for all five participants where the shaded regions indicate the 95% confidence interval for the three online trials. The dashed lines are respectively the baseline human walking EMG effort without exosuit assistance. Participant 1 provided the offline policy.