Enhancing Lower Limb Blood Flow by Optimizing Intermittent Pneumatic Compression Timing

by

Iara Santelices

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Applied Science
in
Mechanical and Mechantronics Engineering

Waterloo, Ontario, Canada, 2023

© Iara Santelices 2023

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Intermittent pneumatic compression (IPC) is used in the management of vascular disorders. IPC systems apply external pressure to the circumference of the leg to enhance blood velocity (BV). Although IPC device performance can be evaluated by the induced change in mean blood velocity (a representation of the device's ability to vacate venous blood from the legs), commercial IPC systems are often designed with the main considerations of cost and size instead of performance. Previous studies have demonstrated that the application of cardiac-gated compression further enhanced blood velocity (BV) compared to fixed compression timing (CT). However, optimal CT, as a function of the cardiac cycle, is not constant across individuals and may change over time. Additionally, IPC is often being used over the course of days and weeks to promote blood flow. Consequently, at the start of a therapy session, the optimal CT for an individual is unknown. Learning the optimal CT for an individual will maximize BV and make the therapy session more efficient. Current CT modelling methods for IPC are limited to predictions for a single day and one heartbeat ahead, which is not sufficient for typical IPC usage since typical IPC sessions last for 60 minutes and IPC is used over the course of days and weeks.

In this thesis a data-driven predictive model to simulate the blood flow response to compression timing across two days was built. Furthermore, a deep reinforcement learning agent learns the CT that maximizes blood flow for each heartbeat in a custom-built simulated environment. Finally, a deep reinforcement learning agent optimizes compression timing with a human-in-the-loop.

In this thesis, six participants wore a custom IPC system and experienced random cardiac-gated CTs for 1.5 hours per day for two days. Six user-specific nonlinear autoregressive models with exogenous inputs (NARX) were implemented using an artificial neural network (ANN) to estimate the BV response to a CT. The BV response to a CT is estimated using past inputs of BV, electrocardiography (ECG), and CT. The NARX model is trained on the first session. This predictive model can be used online to estimate the optimal cardiac gated CT for a therapy session. The mean R^2 for this model across participants on the second session was 0.74 ± 0.09 and the mean absolute error was approximately 3%, which is a reduction of only 11% compared to the first sessions, for both metrics. This study is the first to show that BV across IPC sessions can be predicted using a pre-trained model.

A deep reinforcement learning (DRL) algorithm learns the optimal compression timing to maximize its cumulative reward, mean beat-to-beat mean blood velocity (BBMV). In this work, participant-specific simulated lower limb environments were built, using the

aforementioned NARX models, for six participants and show that DRL can optimize CT, across 19 CTs, in an IPC framework. We show that the DRL agent can adapt to changes in the physiological state unlike previous methods which employ fixed dynamics. The DRL agent can learn an optimal policy in 15 minutes $\pm$ 2 on average. The proposed DRL agent can be implemented in IPC systems to rapidly learn the optimal CT with a human-in-the-loop. This is particularly valuable as the ideal CT is unknown at the outset of a therapy session, and learning it maximizes the benefit of therapy.

It is shown that a DRL agent can optimize IPC CT with a human-in-the-loop within a 60-minute therapy session. This study included 10 participants. To observe learning within 60 minutes, the DRL agent optimizes across two cardiac gated CTs, diastole and systole. The DRL agent state is composed of previous heart rate and BBMV information. The utilized reward function incentivizes constant increases in BBMV to achieve agent convergence. The DRL agent converged to diastole CT at each heartbeat across 10 participants, selecting diastolic compression 83 % $\pm$ 9 of the time. The mean difference in blood velocity between the DRL CT and cardiac-gated diastolic CT was -0.09 cm/s $\pm$ 0.09 (-4% $\pm$ 4). This DRL agent is the first of its kind to optimize IPC CT with a human-in-the-loop.

Acknowledgements

I would like to thank my supervisors, Dr. Arash Arami and Dr. Sean Peterson, for their mentorship and guidance over these past years. Thank you to Dr. Céderick Landry for generously dedicating time to this project and for his recommendations. Thank you to Dr. Andrew Robertson and Dr. William Melek for taking the time to read my thesis. I would also like to thank my lab mates for their suggestions.

Finally, I would like to thank my partner, parents, and my grandparents for their ongoing support and encouragement.

Table of Contents

Author's Declaration ii

Abstract iii

Acknowledgements v

List of Figures x

List of Tables xiv

List of Abbreviations xv

1 Introduction 1

1.1 Motivation and Objectives 1

1.2 Thesis Outline 2

2 Literature Review 4

2.1 Cardiovascular Physiology and Physiological Measurements 4

2.1.1 Vascular Circulation 4

2.1.2 Electrocardiography 6

2.1.3 Measuring Blood Velocity 9

2.2 Intermittent Pneumatic Compression 9

2.2.1 Passive Compression . . . 9
2.2.2 Active Compression . . . 11
2.2.3 Cardiac-Gated Compression Timing . . . 11
2.3 Data Driven Models . . . 13
2.3.1 Autoregressive Networks . . . 14
2.3.2 Artificial Neural Networks . . . 16
2.3.3 Modelling Blood Flow . . . 21
2.4 Reinforcement Learning . . . 22
2.4.1 Reinforcement Learning Algorithms . . . 23
2.4.2 Reinforcement Learning Applied to Simulated Physiological Systems 25
2.4.3 Reinforcement Learning: Sim-to-Real Transfer . . . 26

3 Modelling Blood Velocity 28
3.1 Methods . . . 28
3.1.1 Experimental Procedure . . . 28
3.1.2 Model Architecture and Training . . . 30
3.1.3 Performance Evaluation . . . 30
3.2 Results . . . 31
3.3 Discussion . . . 32
3.4 Conclusion . . . 35

4 Evaluating Deep Reinforcement Learning to Optimize Intermittent Pneumatic Compression Timing 36
4.1 Methods . . . 36
4.1.1 Environment Formulation . . . 36
4.1.2 Training Strategy . . . 39
4.1.3 Hyperparameter Tuning . . . 39
4.1.4 Policy Validation . . . 40

4.2 Results . . . 40
4.2.1 Training . . . 40
4.2.2 Test . . . 41
4.2.3 Changing Physiological State . . . 43
4.3 Discussion . . . 43
4.4 Conclusion . . . 48

5 Optimizing Intermittent Pneumatic Compression Timing with a Human-in-the-Loop 49
5.1 Methods . . . 49
5.1.1 Compression System . . . 49
5.1.2 Participants . . . 50
5.1.3 Experimental Procedure . . . 50
5.1.4 Reward Function . . . 53
5.1.5 Training . . . 54
5.1.6 Agent Evaluation . . . 55
5.2 Results . . . 56
5.3 Discussion . . . 64
5.3.1 Limitations and Future Work . . . 67
5.4 Conclusion . . . 69

6 Conclusions and Recommendations 70
6.1 Conclusions . . . 70
6.2 Limitations and Recommendations . . . 71

References 74

APPENDICES 87

A Appendix A **88**

A.1 Methods 88

A.1.1 Participants 88

A.1.2 Experimental Procedure 88

A.1.3 Environment Formulation 89

A.2 Results and Discussion 89

List of Figures

2.1 Overview of the circulatory system. Image by Anatomy & Physiology, Connexions Web site, is licensed under CC BY 3.0. [14]. 5

2.2 Mean systemic blood pressure at various locations along the circulatory system. Image by Anatomy & Physiology, Connexions Web site, is licensed under CC BY 3.0. [16]. 6

2.3 Calf compressing the veins to promote venous blood flow (muscle pump). Image by Anatomy & Physiology, Connexions Web site, is licensed under CC-BY-3.0. [15]. 7

2.4 Wiggers diagram that shows Electrocardiography (ECG) signal with labelled waves. Systolic and diastolic regions of the ECG signal are also indicated. Image by Adh30, is licensed under CC-BY-4.0. 8

2.5 Sample common carotid arterial blood velocity (BV) signal measured with Doppler ultrasound. Image is by Daniel W. Rickey and is licensed under CC-BY-SA-2.5 [98]. 10

2.6 Gradient descent. The derivative, orange line, can be used to optimize the cost function, blue line 15

2.7 Bias-variance trade-off. This plot shows that for relatively high variance error bias error will be lower and that the opposite is also true. The goal is to balance both. Image is under the public domain. 16

2.8 Typical feedforward neural network structure where information flows from bottom to top. The left shows the two inputs, x_1 and x_2, separately while the right side shows the inputs as one vector $\mathbf{x}$. 17

2.9 Training and validation set loss during ANN training. 19

2.10 An RNN structure. The connection between the second and the first hidden layer makes this an RNN. 20

2.11 A nonlinear autoregressive network with exogenous inputs (NARX) model. The feedback connection shows how the output is fed back to the input. . . 20

2.12 The agent–environment interaction in a MDP. 22

3.1 (a) Experimental setup. Participant is shown standing for visual clarity, however during data collection participants were seated, (b-d) sample BV, ECG, and computed compression timing (CT) signals that served as inputs for the neural network, respectively. The shaded region shows the time steps fed in as inputs. (e) A typical BV and applied pressure signal. Double-ended arrow shows the CT from the local pulse wave peak (dark red point) to peak inflation of the first cuff. 29

3.2 Beat-to-Beat Mean Blood Velocity (BBMV) as a function of the CT. (a1) Participant 1, day 1; (a2) Participant 1, day 2; (b1) Participant 6, day 1; and (b2) Participant 6, day 2. 31

3.3 Measured (blue) and estimated (red) BV for different CTs of, from left to right, 20%, 40%, 60%, and 80% for Participant 3 on (a) same-day training and testing; and (b) different-day training. 32

3.4 Measured (blue) and estimated (red) BV (mean ± SD). (a1) Participant 2, day 1, Spearman correlation coefficient = 0.68; (a2) Participant 2, day 2, Spearman correlation coefficient = 0.60; (b1) Participant 4, day 1, Spearman correlation coefficient = 0.94; and (b2) Participant 4, day 2, Spearman correlation coefficient = 0.62. 33

4.1 State space for agent. (a) Mean and SD of Heart Rate (HR) is shown in the dark blue square. The faded blue circles show the HR for each heartbeat. (b) Mean and SD for BBMV is shown in dark purple square. The faded purple circles show the BBMV for each heartbeat. (c) CT applied. Compression is applied at each heartbeat and the CT is held constant over one agent time step of 5 heartbeats. (d) Agent action space. Delays from 20% to 80% at 20% intervals are shown for clarity. 38

4.2 Agent environment interaction. The BV estimation model takes inputs of previous ECG, CT, and BV to predict future BV in a closed loop format. T_{N800} is the defined as t-800 ms to t-10 ms, T_{N20} is the previous t-20 ms to t-10 ms. These time steps form the input for the NARX model to estimate future BV. ($NARX_{bv}$) model. 39

4.3 Mean ± SD of BBMV during training for agents with different reward functions across 5 trials. Results are shown for R_1, R_2, and R_3 for different participants. (a) participant 1; (b) participant 2; (c) participant 3; (d) participant 4; (e) participant 5; (f) participant 6. 42

4.4 Plots (a1, b1, c1, d1, e1, and f1) show mean ± SD BV model BBMV response for different CTs. Plots (a2, b2, c2, d2, e2, and f2) show mean ± SD BBMV over 1 hour of compression for trained agents with different reward equations. CT_f and optimal CT are shown for comparison. (a1, a2) participant 1; (b1, b2) participant 2; (c1, c2) participant 3; (d1, d2) participant 4; (e1, e2) participant 5; (f1, f2) participant 6. 44

4.5 Mean ± SD of BBMV during training for agents with different entropy values across 5 trials shown for R_3. The vertical dotted line indicates when the predictive simulator changed. The black line in PS_B shows the BBMV achieved by selecting the CT with the highest BBMV. (a) participant 1; (b) participant 2; (c) participant 3; (d) participant 4; (e) participant 5; (f) participant 6. 45

5.1 (a) Typical mode control signal that alternates every 20 heartbeats. (b) Action selected at each heartbeat. 51

5.2 (a) Typical BV signal. Orange highlighted section shows systole and blue highlighted section shows diastole. (b) Sample pressure applied. First 3 heartbeats shows compression occurring in diastole, CT_d, and remaining heartbeats show compression occurring in systole, CT_s. 52

5.3 BBMV during 5 minutes Intermittent pneumatic compression (IPC) applied at CT_d at each heartbeat. The plot shows mean BBMV for blocks of different heartbeats, where block 1 and block 2 are consecutive blocks. (a) 5 heartbeats; (b) 10 heartbeats; (c) 15 heartbeats; (d) 20 heartbeats; (e) 25 heartbeats; (f) 30 heartbeats. 53

5.4 Agent-human interaction. The environment receives an action, A_t from the agent and applies compression with the desired CT. The measured BBMV and ECG signals compose the state and reward output, S_t and R_t, respectively. These values serve as inputs to the agent and the agent outputs the next action, A_{t+1}, to continue the cycle. 55

5.5 Rewards during Deep Reinforcement Learning (DRL) training for actions CT_d (diastole) and CT_s (systole). The moving average of the reward is shown in dark blue. (a) Participant 1; (b) Participant 2; (c) Participant 3; (d) Participant 4; (e) Participant 5; (f) Participant 6; (g) Participant 7; (h) Participant 8; (i) Participant 9; (j) Participant 10. 57

5.6 Action selection during training for DRL agents. (a) Participant 1; (b) Participant 2; (c) Participant 3; (d) Participant 4; (e) Participant 5; (f) Participant 6; (g) Participant 7; (h) Participant 8; (i) Participant 9; (j) Participant 10. 59

5.7 BBMV and HR over time during DRL training. The faded line shows the measured BBMV and HR and the solid dark line shows the moving average. (a) Participant 3; (b) Participant 7. 61

5.8 Difference between consecutive exposure and control mode blocks during DRL training (scatter points). Dashed line shows the mean difference during training. Solid line shows the moving average of the difference. The blue shaded region is the SD of the exposure blocks. (a) Participant 1; (b) Participant 2; (c) Participant 3; (d) Participant 4; (e) Participant 5; (f) Participant 6; (g) Participant 7; (h) Participant 8; (i) Participant 9; (j) Participant 10. 63

A.1 (a) ECG R-peak amplitude (circles) and heartbeat values (arrows);(b) BV systolic peak amplitudes (squares) and BV systolic peak locations relative to the ECG R-wave location (arrows). 90

A.2 BBMV and HR when fixed CT is applied for 30 minutes on the same day for the same participant: (a) morning HR; (b) afternoon HR; (c) morning BBMV; and (d) afternoon BBMV. 91

A.3 mean and SD BBMV for different CTs: (a) day 1 morning; (b) day 1 afternoon; (c) day 2 morning; (d) day 2 afternoon. 92

List of Tables

3.1 Results of $NARX_{bv}$ model for training and testing on same and different days. Results shown are for each participant for $NARX_{bv}$ model trained on day 1. (A) BBMV R^2; (B) Spearman correlation coefficient; (C) BV MAE (cm/s); (D) Absolute change in BBMV from part A to B; (E) HR and BBMV Pearson correlation coefficient (measured data); (F) % change in BBMV from 10% CT to model optimal CT. The differences in results from same-day to next-day testing were not significant. 34

5.1 Hyperparameters used for DRL agent. 54

5.2 MK test results for slope and BBMV difference. + indicates a positive slope and − indicates a negative slope. # denotes p<0.01 64

A.1 Mean and SD BBMV, SD of BBMV as a percentage of the mean BBMV, HR, and pulse delay for 5 minutes of no compression, 30 minutes of fixed CT, and 30 minutes of DRL CT selection at each heartbeat. 90

List of Abbreviations

$NARX_{bv}$ NARX model to estimate future BV. xi, xiv, 30, 31, 34, 37, 39, 47, 70–72

ANN artificial neural network 14, 16–20, 22, 25, 30, 39, 68

BBMV Beat-to-Beat Mean Blood Velocity xi–xiv, 13, 21, 22, 30–47, 49, 52–56, 60–62, 64–69, 89–92

BV blood velocity x–xiv, 9, 10, 12, 13, 21, 28–37, 39, 40, 44, 47, 48, 52, 53, 65, 67–72, 89–91

CGC cardiac gated compression 11–13, 28–30, 49, 50, 65, 72, 88

CT compression timing xi–xiv, 2, 11–13, 21, 22, 29–41, 43–50, 52–56, 60, 62, 64–73, 88–92

DRL Deep Reinforcement Learning xiii, xiv, 2, 3, 22, 26, 27, 36, 37, 39–41, 43, 46–50, 52–55, 57, 59, 61–73, 88–91

ECG Electrocardiography x, xi, xiii, 2, 6, 8, 11, 21, 28–30, 33, 35, 37, 39–41, 43, 48, 50, 68, 71, 72, 89, 90

HR Heart Rate xi, xiii, xiv, 7, 12, 14, 22, 30, 34, 36, 38, 60, 61, 89–91

IPC Intermittent pneumatic compression xii, 1, 2, 4, 9, 11–13, 28–30, 32–35, 37, 43, 47–50, 53, 64, 67–72

NARX nonlinear autoregressive network with exogenous inputs xi, 19–21, 30, 32–35, 64, 65, 70

Chapter 1

Introduction

Deep vein thrombosis (DVT) is a condition where clots form in the deep veins, commonly in the calf, principally caused by surgery, reduced mobility, and cancer [69]. A conservative estimate placed annual American health care costs associated with DVT, and its subsequent complications, between $7 and 10 billion [41]. Peripheral arterial disease (PAD) is a chronic atherosclerotic cardiovascular disease where narrowing or blockages of the peripheral arteries limit blood flow to the legs. On average, annual American healthcare costs for those with PAD are $7000 greater than those without [106] and one in every 20 Americans over the age of 50 has PAD [90]. Intermittent pneumatic compression (IPC) is commonly used to prevent DVT and PAD in clinical settings [113, 36], and has also been shown to be beneficial for exercise recovery [131] and reducing the effects of orthostatic stress [139, 34].

IPC systems mimic the natural increased blood flow from the muscle pump action by applying external pressure to the circumference of the leg. IPC has been shown to enhance lower limb blood flow in healthy subjects and those with arterial disease [123, 85]. It has been hypothesized that this increase in blood flow is how IPCs prevent DVT [85, 122, 56].

1.1 Motivation and Objectives

This thesis introduces a novel IPC timing strategy that optimizes blood flow for an individual. Commercial IPC systems generally compress, at most, once a minute, and are often designed with the main considerations of cost and size instead of effectiveness or efficacy [108]. Previous research in IPC optimization has focused on venous refill time, the time required for veins to refill after blood has been evacuated to the rest of the body, as

a metric to control IPC compression timing (CT) and improve blood flow. However, this method increases peak blood velocity and there is no evidence to suggest that peak velocity is a valid metric for DVT prevention [85]. Recent research has shown that by timing compression with respect to the cardiac cycle, applying compression approximately 60 to 100 times per minute, there is a significant increase in arterial blood flow in comparison with commercial devices [26]. In this study, the compression rate of the other devices was 1/minute, 0.8/minute to 1.9/minute (dependent on vascular refill time), and 1/minute. A study by Landry et al. showed that applying compression at different points in the cardiac cycle yields a different response across individuals [63]. These differences in blood flow response complicate the process of finding the optimal CT for enhanced blood flow, which also can vary with physiological state [63]. If the optimal CT can be applied during a therapy session, where compression is applied for 60 minutes, then blood flow for an individual to optimize the treatment can be maximized and thus therapy efficacy improved or therapy time potentially reduced.

It is hypothesized that deep reinforcement learning (DRL) can determine the CT that will yield increased blood flow compared to fixed timing for an individual over the course of a 60-minute therapy session. The goal of a DRL agent is to map states to the optimal actions that maximize a designed reward function, based on the measured blood flow. Interacting with the environment, the agent learns the optimal actions for different states. In this setting, the agent observes a state (an individual's physiological state) and selects an action (a CT) to maximize the reward, which is related to the individual's blood flow. To achieve this goal, three objectives are defined.

1. Create a data-driven predictive model to simulate blood flow from electrocardiography (ECG) and compression signals.
2. Build a DRL agent that learns the CT that maximizes blood flow for each heartbeat. This agent will be tested on simulated data produced using the data-driven model in Objective 1.
3. Test and verify the DRL agent developed in Objective 2 on healthy human participants.

1.2 Thesis Outline

Chapter 2 presents a review of the literature relevant to cardiovascular physiology, IPC systems, data-driven models, and reinforcement learning (RL). The core contributions of

this thesis are presented in Chapters 3 to 5. In Chapter 3, a model tracks blood flow during compression over two days. Chapter 4 introduces a custom lower limb simulated environment where it is demonstrated that a DRL agent can optimize mean blood flow faster than previous methods. In Chapter 5, a DRL agent is trained to optimize blood flow with a human-in-the-loop. Chapter 6 presents the major conclusions drawn from this work and discusses directions for future research.

Chapter 2

Literature Review

This chapter introduces the background needed to understand and optimize IPC devices. First, an overview of the cardiovascular system and common physiological sensors are presented. Second, IPC systems and optimization are discussed. Third, data-driven models are explained in the context of this research project. Last, a review of RL theory and RL applied to physiological systems is presented.

2.1 Cardiovascular Physiology and Physiological Measurements

2.1.1 Vascular Circulation

The content in this section is based upon the textbook by Silverthorn [112]. The main function of the circulatory system is to transport oxygen and other vital materials to cells in the body. It consists of two circulatory loops; the pulmonary and systemic circulation systems. The pulmonary system sends deoxygenated blood from the right ventricle of the heart through the pulmonary arteries to the lungs for oxygenation. Afterwards, the oxygenated blood travels through the pulmonary veins to return to the left atrium of the heart. Blood then travels to the left ventricle of the heart, which pumps this oxygenated blood to the systemic circulatory system. In the systemic loop, arteries carry oxygenated blood from the heart to the body. Veins return deoxygenated blood from the tissues to the right atrium of the heart, which feeds the right ventricle and the circulatory loop continues, as shown in Fig. 2.1.

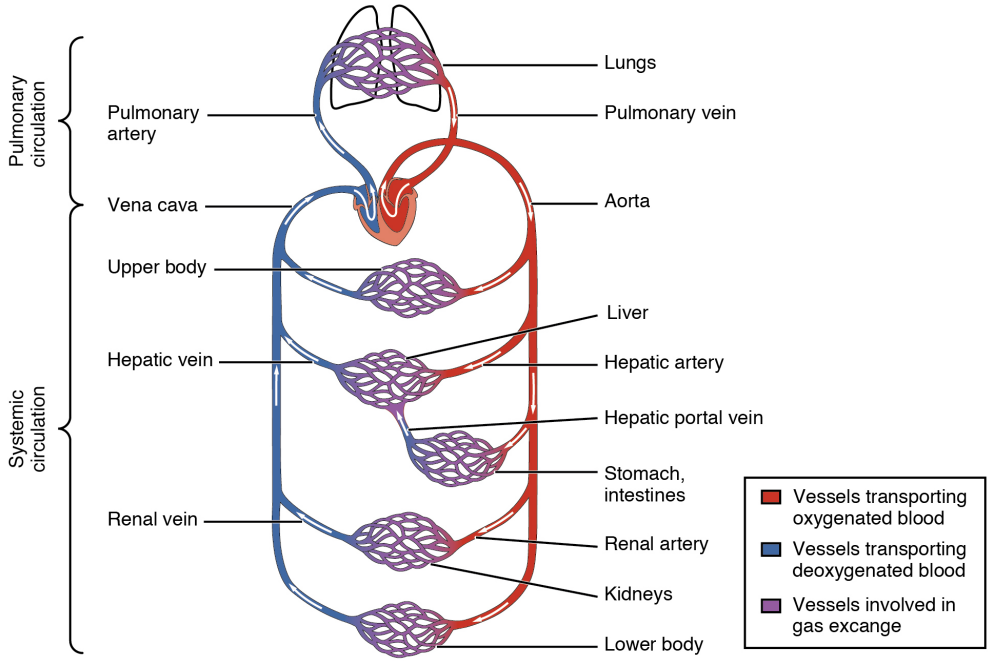


Figure 2.1: Overview of the circulatory system. Image by Anatomy & Physiology, Connexions Web site, is licensed under CC BY 3.0. [14].

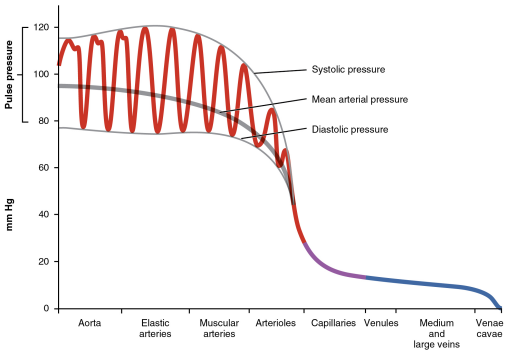


Figure 2.2: Mean systemic blood pressure at various locations along the circulatory system. Image by Anatomy & Physiology, Connexions Web site, is licensed under CC BY 3.0. [16].

Contraction of the left ventricle produces a pressure wave that drives arterial blood flow. The cross-sectional area of the vessels also increases dramatically from the aorta toward the micro-circulation. Also, veins are much more compliant. Thus pressure inside them is lower compared to the arteries, see Fig. 2.2.

Veins are 10 to 20 times more compliant than arteries [33] and hold 60 to 70% of the body's blood volume [32]. When an individual is standing, gravity shifts blood volume towards the lower body. The veins act as a reservoir and blood pools in the lower limbs. One method to act against this pooling is the muscle pump action, see Fig. 2.3. During ambulation, the calf muscles contract which applies force on the compliant veins [12]. During muscle contraction there there is expulsion and central mobilization of peripheral venous blood, which promotes venous return and increases both stroke volume and cardiac output. Valves in the veins prevent back flow, since venous flow is not primarily driven by the pressure of the cardiac cycle. The muscle pump also is also hypothesized to increase blood blood flow in the active muscle within 1 second following a single contraction. This increase may be because the muscle pump results in decreased venous pressure which increases the pressure gradient across the muscle vascular bed.

2.1.2 Electrocardiography

The content in this section is based upon the textbook by Silverthorn [112]. An ECG is a recording of the summed electrical activity generated by the cells in the heart measured by electrodes on the skin surface, a typical signal is shown in Fig. 2.4. The two main elements of an ECG are waves and segments. Waves are sections of the signal that are greater than

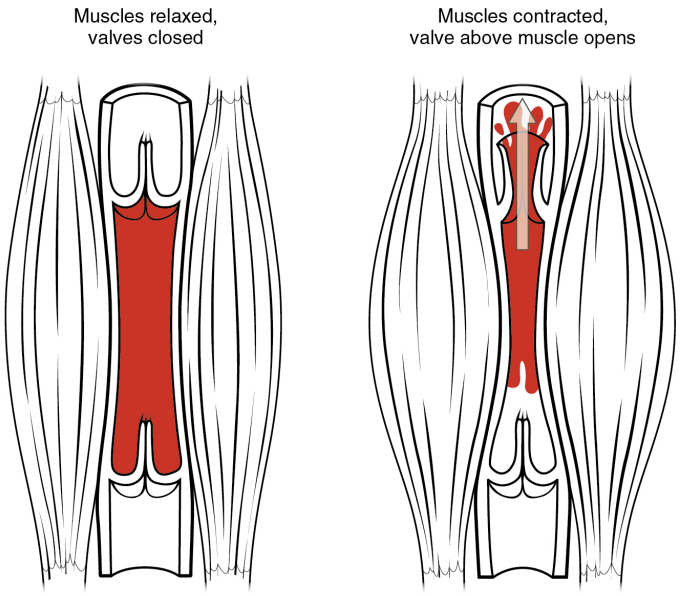


Figure 2.3: Calf compressing the veins to promote venous blood flow (muscle pump). Image by Anatomy & Physiology, Connexions Web site, is licensed under CC-BY-3.0. [15].

or less than the baseline; segments are the parts of the baseline in between waves. Heart rate (HR) can be timed from peak of one R wave to the subsequent R wave peak, also referred to as the RR interval. The R wave peak is chosen because of its prominence. Each cardiac cycle has two phases: diastole, when the cardiac muscle relaxes and refills with blood, and systole, when the muscle contracts and pushes blood out of the heart. These cycles are shown in Fig. 2.4.

A healthy heart does not maintain a constant HR [107]. This variability between successive RR intervals is called heart rate variability (HRV). HRV reflects the body's ability to maintain homeostasis (i.e., adjust to changes in the tissue or cell environment. HR is impacted by the sympathetic and parasympathetic nervous systems. During inspiration, HR increases and during expiration it decreases. Sympathetic activity increases HR while parasympathetic does the opposite. By taking the fast Fourier transform (FFT) of the HRV signal, the HRV signal can be decomposed into different frequency bands. The high frequency (HF) band, 0.15 to 0.4 Hz represents parasympathetic activity. The low frequency (LF) band, 0.04 to 0.15 Hz, may be produced by both parasympathetic and sympathetic activity [91, 68], or by primarily the parasympathetic nervous system [96] or baroreflex activity alone [37].

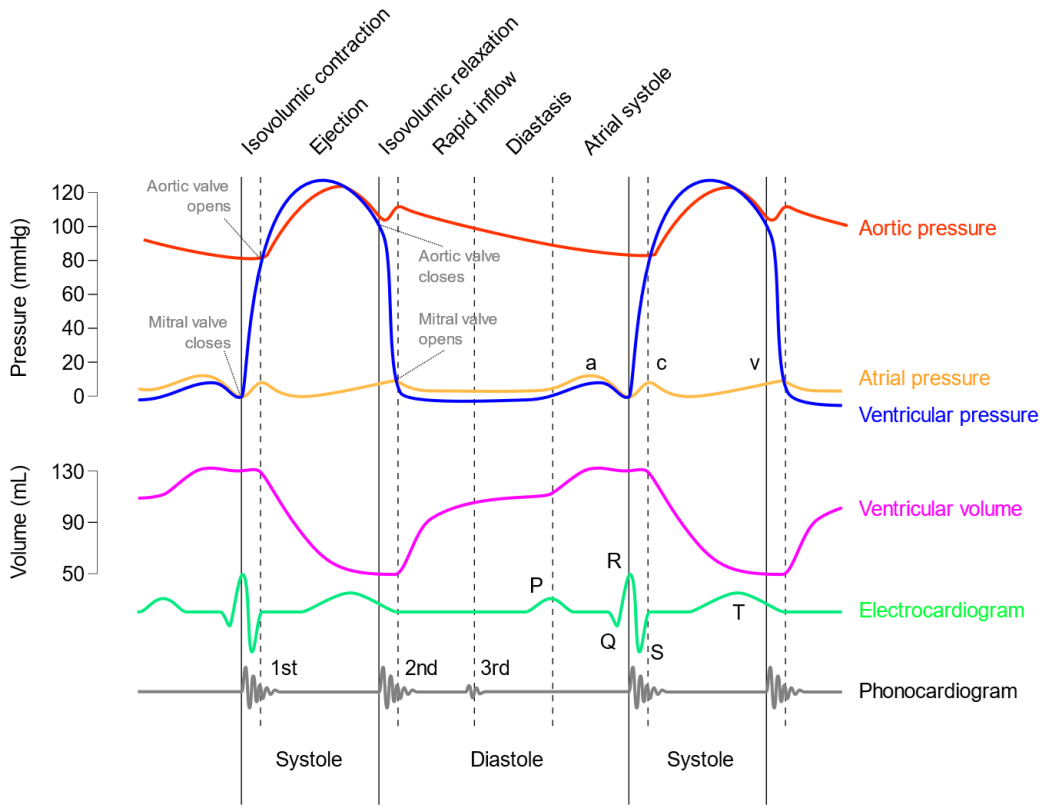


Figure 2.4: Wiggers diagram that shows ECG signal with labelled waves. Systolic and diastolic regions of the ECG signal are also indicated. Image by Adh30, is licensed under CC-BY-4.0.

2.1.3 Measuring Blood Velocity

Often, IPC performance is evaluated by superficial femoral arterial blood flow. Arterial blood flow is assumed to be proportional to blood velocity (BV), as previous studies have shown that arterial diameter is not impacted by IPC [141].

Doppler ultrasound can measure instantaneous BV [28]. A high frequency wave, f is beamed into the blood stream. The frequency of the back scattered wave is shifted in frequency by f_d by moving red blood cells. BV can be calculated by using the following equation from [51]:

$$BV = \frac{f_d c}{2f\cos(\theta_b)} \tag{2.1}$$

where c is the speed of sound in the scattering medium, f is the emitted beam frequency, and θ_b is the beam-vessel angle. The equation shows that when θ_b is smaller than 60°, uncertainty in its value does not impact the velocity as much as when θ_b is greater than 60°. For example, a 5° error in estimating the Doppler angle at 60° will cause a 5.4% error in the velocity calculation. However, at 60° this error increases to 12% [73].

The superficial femoral arterial (SFA) BV signal has two phases, systole and diastole. It is common to characterize systole as from the foot of the wave to the dicrotic notch and diastole as from the dicrotic notch to the foot of the next BV waveform. A sample common carotid arterial BV signal obtained with a Doppler ultrasound device is shown in Fig. 2.5.

2.2 Intermittent Pneumatic Compression

This section discusses two types of mechanical compression applied to the lower limbs, passive and active compression. These compression systems aim to enhance arterial blood flow.

2.2.1 Passive Compression

Passive compression systems are either bandages, applied in a manner to achieve graduated compression (GCS). The mechanism of action of these systems is to apply the greatest amount of compression at the ankle and decrease compression proximally. This pressure

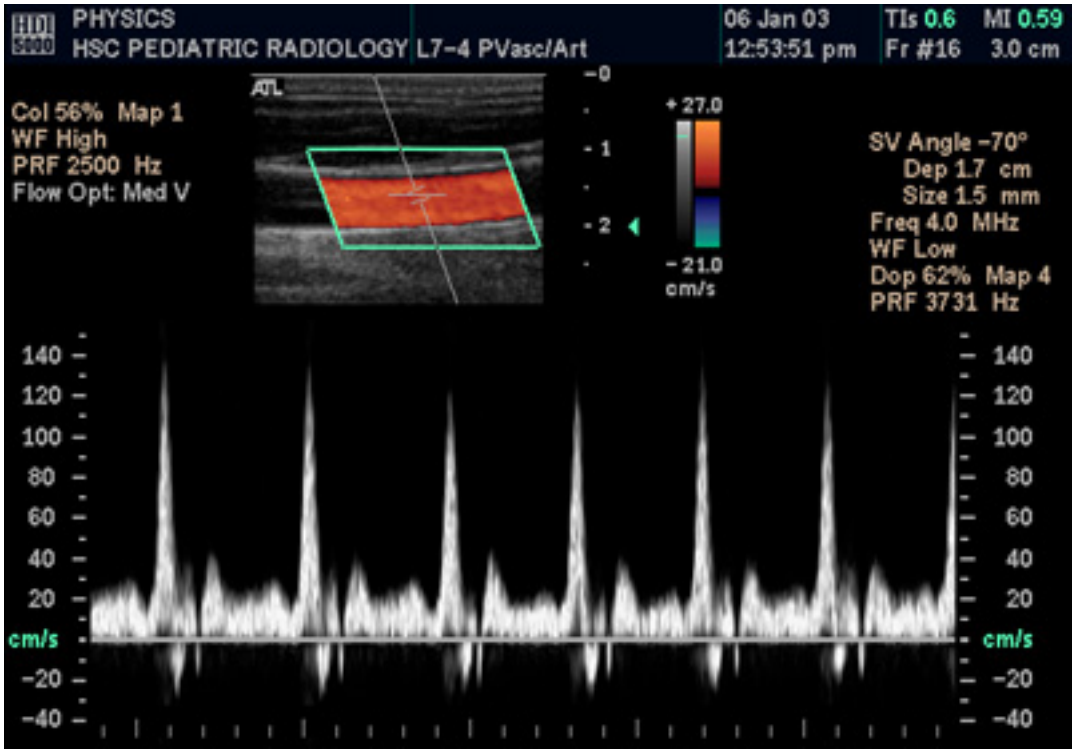


Figure 2.5: Sample common carotid arterial BV signal measured with Doppler ultrasound. Image is by Daniel W. Rickey and is licensed under CC-BY-SA-2.5 [98].

gradient intends to promote venous blood flowing towards the heart instead of pooling in the foot.

Compression garments are used by patients in hospital settings to treat lymphatic and venuous insufficiency [126], along with those seeking increased blood flow for exercise recovery [75]. While these garments have been shown to be effective at treating patients with venous leg uclers [83], previous work has shown that these garments impact neither the blood flow nor muscle oxygenation in healthy participants [9] and their effectiveness at treating venuous thromboembolism is unclear [19]. Another limitation is that these garments must be properly fitted, otherwise they can cause discomfort or in the worst case, skin necrosis [13].

2.2.2 Active Compression

IPC is an active compression therapy that has been shown to enhance blood flow and be effective in preventing DVT [85]. IPC devices consist of pneumatic cuffs that wrap around the lower limb to apply compression. Device design varies across manufacturer from cuff coverage (calf alone, calf and thigh, calf and foot), pressure applied (80 to 120 mmHg) [119], and compression timing. Cuff inflation can be slow and last 10 to 12 seconds or be fast and take place within 0.3 to 0.5 seconds [78]. Peak cuff inflation can be held for 1 to 13 seconds [119]. These devices generally employ "fixed timing", where compression is applied with a set frequency, often 1 to 3 times per minute.

Recent studies have focused on implementing "smart timing" algorithms to enhance the blood flow response to IPC. For example, the Kendall SCD Express (Cardinal Health, USA) IPC estimates the venous refill time by measuring an individual's calf circumference during calibration and uses this value to control compression timing [56, 55, 84]. When compared against other IPCs that use slow inflation and long compression periods, and rapid inflation and short compression periods, the Kendall SCD resulted in a increased venous flow over a longer time period [40, 55].

2.2.3 Cardiac-Gated Compression Timing

A promising strategy to augment arterial blood flow with IPC devices is to use a measurement of the cardiac cycle to control CT, referred to as cardiac gated compression (CGC).

Tochikubo et al. developed an IPC device (PS-AM) that triggers compression in diastole at each heartbeat based on an earlobe photoplesthymography (PPG) signal [120].

Peripheral venous blood flow, measured by using a laser Doppler flow meter attached to the underside of the third toe, was evaluated for 3 conditions: no compression; PS-AM at 1 second intervals; and PS-AM with cardiac synchronized compression. When compression was applied for 15 minutes in a supine position the reported increases in blood flow, compared to no compression, for PS-AM at 1 second intervals and PS-AM with cardiac synchronized compression were 23% and 65%, respectively. While compression occurs within diastole, details about the timing with respect to local blood flow, as a percentage of the local blood pulse, were not reported. Specifically, whether the compression occured at the beginning of diastole, or the end of diastole is unknown. They also reported a positive relationship between the percent increase in blood flow and the percent increase in the HF band power of the ECG signal between no compression and compression.

Jonas et al. [53] compared the impact of a cardiac gated pulsating suit, Stendo, (Stendo, France) to a typical IPC device (TP05 or TP07 Eureduc, France) for decreasing total limb volume in lymphedema patients. The Stendo device covers the entire lower body and abdomen and consists of 12 compartments that pulsate sequentially (distal to proximal). A compression cycle completes within a heartbeat and compression is timed to the beginning of diastole based on finger PPG. However, more details about compression duration and exact CT were not provided. The typical IPC device had 40-second inflate and 21-second deflate periods. In the study, patients were in supine position and compression was applied for 1 hour a day for 5 consecutive days. Stendo patients experienced a greater decrease in total limb volume (11.2% versus 8.8%) and they reported a larger increase in quality of life compared to the typical IPC patient group. A limitation of the Stendo suit is that local blood flow cannot be measured during a therapy session because the entire lower limb is covered by the suit. Consequently, forearm blood flow [97] is used to measure the blood flow in studies. While both the PS-AM and Stendo devices use cardiac gated compression timing, neither system applies compression in local diastole.

The CGC, a custom-built device made at the University of Waterloo that is used in this thesis, is a rapid-inflation IPC device that applies compression in local diastole based on a measurement of the local blood pulse. The CGC device consists of five cuffs, each encasing a separate air bladder, that inflate sequentially from ankle to knee. The compression cycle lasts < 200 ms so that a compression cycle can be completed within diastole. Average resting HR for a healthy individual is often between 60 to 100 beats per minute (BPM) [102]. For HRs of 60 bpm, 80 bpm, and 100 bpm, the RR interval spans 1000 ms, 750 ms and 600 ms, respectively. The length of diastole for these HRs is approximately 750 ms, 560, and 450 ms, respectively.

Two studies investigated the application of IPC with the CGC device with exercise. The studies tested the hypothesis that IPC with the CGC device would enhance blood flow during light exercise (single-leg plantar flexion exercise) and post exercise recovery [140] and walking and post walking recovery [141]. These studies showed that the CGC device augmented exercise and post exercise SFA BV [140, 141] and tissue oxygenation during recovery [141].

Two studies assessed the impact of IPC with the CGC device to reduce orthostatic stress [139, 34]. Gibbons et al. found that CGC of the lower legs after an active orthostatic exercise enhanced local, central, and cerebral hemodynamics in healthy participants [34]. Zuj et al. found that CGC of the lower limbs had the potential to be a therapeutic tool for individuals vulnerable to orthostatic stress [139].

Another study compared four different IPC devices each with different compression cy-

cles. Three devices are from the manufacturers ArjoHuntleigh (Flowtron ACS800), Kendall (SCD Express), and DJO Global (VenaFlow Elite) respectively, whereas the other device is an aforementioned device made at the University of Waterloo (CGC). The Flowtron and VenaFlow both apply compression once every minute. Alternatively, the SCD applies compression based on venous refill time and the CGC applies compression based on the cardiac cycle. The cuffs of these four devices inflate sequentially, from ankle to knee, except for the Flowtron, which applies uniform compression. The Flowtron and SCD both use slow inflation, inflation lasting more than 10 seconds, while VenaFlow and CGC use rapid inflation, inflation lasting less than 10 seconds. The CGC device increased SFA BV compared to the aforementioned commercially available IPCs [26]. This study also found that the SFA BV had the greatest increase, compared to no compression, when participants were in seated position compared to supine.

Landry et al. [63] investigated the SFA BV response to different CTs within a heartbeat since previous pilot studies suggested that the response to a particular CT was not the same across individuals. However, the underlying hemodynamics that govern the optimal CT for an individual are unknown and further data collection was required. Using the rapid-inflation CGC device, 20 random CTs where applied to four individuals for 1 minute each, followed by 5 minutes of no compression, and then 10 random CTs for 1 minute each. CTs are defined by the start of inflation. The results suggested that the optimal CT, as defined by SFA beat-to-beat mean blood velocity (BBMV), was not constant across participants and may change over time.

2.3 Data Driven Models

System identification (SI) is concerned with building mathematical models of dynamic systems by using measurements from the system [72]. Machine learning (ML) and SI are often confused and there is a lack of a definitive distinction between these two methods. In this document, both will be referred to as data-driven models.

Data-driven models can broadly fall into two approaches: supervised and unsupervised learning. Supervised learning uses a dataset where the input data is paired with the desired output. The algorithm aims to create a model that learns the relationship between the inputs and outputs of this data and make predictions of the output for the inputs of a new dataset. The two types of supervised learning are classification and regression. Classification outputs discrete variables, these are often referred to as labels, whereas regression outputs a continuous variable. Unsupervised learning is used to analyze and draw inferences about datasets that have unknown outputs. Clustering is the most common

type of unsupervised learning where the goal is to find hidden groups or patterns in the data.

This document focuses on regression models learning dynamical systems. To that end, auto-regressive models and neural networks are discussed. Physiological signals like HR [3] and blood flow [57] have been shown to have auto-correlation. Auto-regressive networks are designed to model signals with auto-correlation. Additionally, recent advancements in data-driven models have shown excellent performance in characterizing physiological signals with artificial neural network (ANN)s [64, 46].

2.3.1 Autoregressive Networks

The information in this subsection is summarized from [86]. Many measured time series have serial autocorrelation, a linear relationship between lagged observations. This linear relationship suggests that past observations may predict future observations and forms the basis of autoregressive (AR) models. These models are commonly expressed as:

$$y(k) = -a_1 y(k-1) - ... - a_m y(k-m) + v(k) \tag{2.2}$$

where k is the discrete time, y is the output time series, $y(k)$ is the k^{th} sample of said time series, $v(k)$ is the noise, a_1 to a_m are the coefficients, and m is the order of the time series.

The most common linear dynamic model for time series is the autoregressive exogenous input (ARX) model. This is an extension of the AR model with an external input. The ARX model is commonly expressed as:

$$y(k) = b_1 u(k-1) - ... - b_n u(k-n) - a_1 y(k-1) - ... - a_m y(k-m) + v(k) \tag{2.3}$$

where $u(k)$ is the exogenous input, b_1 to b_n are the coefficients, and n is the order of the exogenous time series.

While other more complicated approaches exist, like autoregressive moving average (ARMA) models and autoregressive moving average with exogenous inputs model (ARMAX), the ARX model is commonly used because the parameters are easy to compute. These parameters can be solved with linear least squares estimation (LSE). However, LSE is too computationally expensive to solve ARMA and ARMAX models. Ridge regression, lasso, and instrumental variable estimation are other methods to solve ARX and ARMA models. LSE for an ARX model uses the following equation:

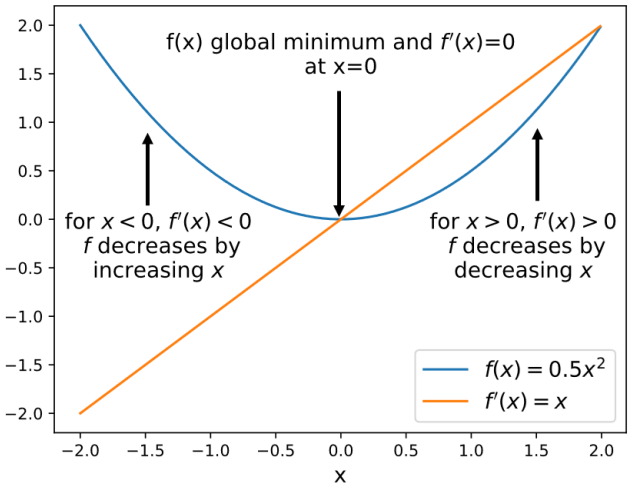


Figure 2.6: Gradient descent. The derivative, orange line, can be used to optimize the cost function, blue line

$$\hat{\Theta} = (X^\top X)^{-1} X^\top y \tag{2.4}$$

where y is the vector of the desired time series output, also called the labelled data. $\hat{y}$ is the best estimate of y using least squares $\hat{y} = X\hat{\Theta}$. X is the regression matrix, where each row is composed of the inputs to the model $u(k-1), ..., u(k-n), y(k-1), ...y(k-m)$, see equation. 2.3. $\hat{\Theta}$ is the optimal estimate of the ARX model coefficients $b_1, ..., b_n, a_1, ..., a_m$.

LSE can solve equation 2.4. However, inverting the Hessian $(X^\top X)$ can be computationally expensive when the number of coefficients is large because the size of $(X^\top X)$ is (m+n) $\times$ (m+n). Iterative methods, like gradient descent, are commonly used in these situations. The intuition behind gradient descent is to use the derivative, which gives information about how to change the input, x, to create a small improvement in y [38]. Consequently, the error of the output is reduced by moving is small steps with the opposite sign of the gradient. Fig. 2.6 shows how gradient descent reduces error when used with a simple equation, until the error is at a minimum. Additionally, LSE can generate biased results for these models. A standard approach to this problem is instrumental variables estimation.

An important consideration when building ARX models is which delayed inputs to include, as determined by the values of m and n in equation 2.3. Autocorrelation, cross-

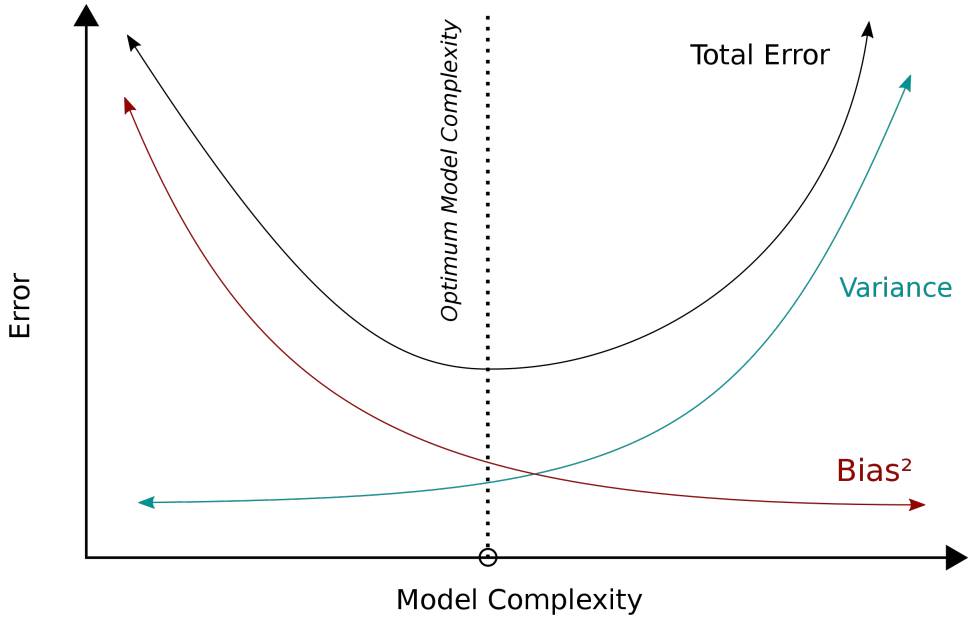


Figure 2.7: Bias-variance trade-off. This plot shows that for relatively high variance error bias error will be lower and that the opposite is also true. The goal is to balance both. Image is under the public domain.

correlation and mutual information can help inform these decisions. In practice, successively including greater delays and observing the results is often and successfully employed.

While AR models can generalize a system's behavior, to achieve this goal models must balance bias and variance. Increasing the model complexity, such as increasing the embedded dimension, can result in overfitting. An overfit, or biased, model performs well on the training data, however preforms poorly when presented with unseen data, or test data. Alternatively, a model with low bias will have similar performance for training and unseen data. Fig. 2.7 presents this trade-off, where the optimal complexity for a model balances both bias and variance.

2.3.2 Artificial Neural Networks

The information in this section is summarized from [38]. This subsection describes the structure of ANNs and how to train these models. The aforementioned AR models are linear models. However, time series with lagged observations that have non-linear rela-

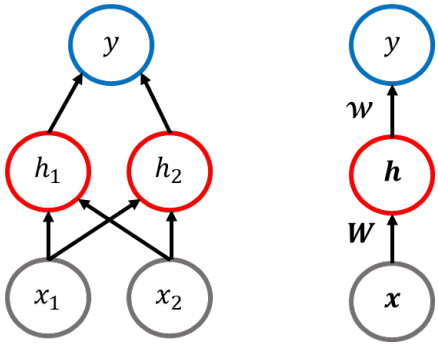


Figure 2.8: Typical feedforward neural network structure where information flows from bottom to top. The left shows the two inputs, x_1 and x_2, separately while the right side shows the inputs as one vector $\mathbf{x}$.

tionships can also be modelled using non-linear methods by using an ANN. An ANN is a collection of neurons that work together to optimize a function. A common feedforward ANN architecture is a multi-layer perceptron, see Fig. 2.8. Generally, the goal of this ANN is to optimize some function f^*. For example, a regression model, $y = f^*(\mathbf{x})$ maps an input $\mathbf{x}$ to a continuous variable y. A feedforward network learns a mapping $\mathbf{y} = f(\mathbf{x}; \theta)$ and the parameters, θ, that result in the best function approximation. These are called feedforward models because information flows in the forward direction, no feedback loops exist in this architecture. These feedforward ANNs can have multiple layers, which define the depth of the model. The first layer of a feedforward ANN is the input final, the final layer is output layer and any layers in between are the hidden layers ($\mathbf{h}$ is a hidden layer in Fig. 2.8). During training, the goal is produce agreement between $f(x)$, the output of the ANN, and the optimal solution, $f^*(x)$. Training data provides noisy, approximate examples of $f^*(x)$ from different points in the input space along with their labels $y \approx f^*(x)$. Generally, feedforward networks can be thought of as function approximation systems that aim to achieve statistical generalization.

Deep learning aims to learn $y = f(x; \theta; \mathbf{w})$, where θ and $\mathbf{w}$ are parameters to help learn the non-linear mapping f that best matches f^*. In Fig. 2.8 the hidden units can be computed by $h = f^{(1)}(x; \mathbf{W}, \mathbf{c})$ where $\mathbf{W}$ are the weights and $\mathbf{c}$ are the biases. These hidden units serve as inputs for the second layer, which in this case is the output layer. The mapping for the second layer can be computed as $y = f^{(2)}(\mathbf{h}; \mathbf{w}, b)$. When combined, the mapping reads as $f(\mathbf{x}; \mathbf{W}, \mathbf{c}, \mathbf{w}, b) = f^{(2)}(f^{(1)}(x))$. In Fig. 2.8, vector $\mathbf{W}$ describes the mapping from x to h, and a vector $\mathbf{w}$ describes the mapping from h to y. The non-linear functions $f^{(1)}$ and$f^{(2)}$ are called activation functions, g. This activation function is

often applied element-wise, $h_i = g(x^\top \mathbf{W}_{:,i} + c_i)$. $\mathbf{W}$ is a matrix where the number of rows corresponds to the number of neurons in the previous layer and the number of columns corresponds to the number of neurons in the next layer. Consequently, $\mathbf{W}_{:,i}$ refers to the column of the weight matrix that corresponds to the given output neuron, h_i.Hidden layers often employ sigmoidal, hyperbolic tangent (tanh) or rectified linear unit (ReLU) activation functions. Output layers generally employ linear activation.

For a simple problem with a limited number of hidden layers and few training examples, it is possible to find the parameters, $\mathbf{W}$ and $\mathbf{c}$, that optimize the problem by hand. However, in applied problems with potentially billions of model parameters and training examples, a gradient-based optimization algorithm, like steepest descent or gradient descent, is a practical method to learn the parameters that result in minimal error. The global minimum is the point that obtains the absolute lowest value of $f(x)$. There is no restriction on the number of global minima for a given function. For a model with many parameters, finding the global minimum is very difficult. Since there is often a value of f that is low, but not necessary the minimum that produces good performance; this is often the end result of ANN training.

When training an ANN, the input, x, provides the initial information that flows through the hidden units at each layer to produce the output y, this is called forward propagation. The x produces a scalar cost, $J(\theta)$, that is based on the difference between the estimated output $\hat{y}$ from input x and the labelled output y from the training data set. Back propagation allows the information from the cost to flow backwards through the network to compute the gradient, which is then used to update the network parameters.

The learning rate, a positive scalar, is used to determine the rate of change in the parameters of the ANN that is proportional to the computed gradient. A constant learning rate forces users to sacrifice speed or performance, since a small learning rate will require many training iterations to converge whereas a larger learning rate may not be able to converge to a local minimum.Consequently, researchers have developed adaptive learning rates optimizers like ADAGRAD [21] and Adam [61].

Hyperparameters, parameters whose value controls the learning process, are important to model performance. In research, these values are often tuned to achieve a desired result. Examples of hyperparameters include the aforementioned η and the optimizer (ADAGRAD, Adam etc). Another hyperparameter is the minibatch size, or batch size. Optimizers often compute the update to the model parameters based on a subset of the training data, because this method is less computationally expensive than evaluating the model on every example in the training data individually. The number of examples in this subset is referred to as the batch size. A smaller batch size may be biased towards the last

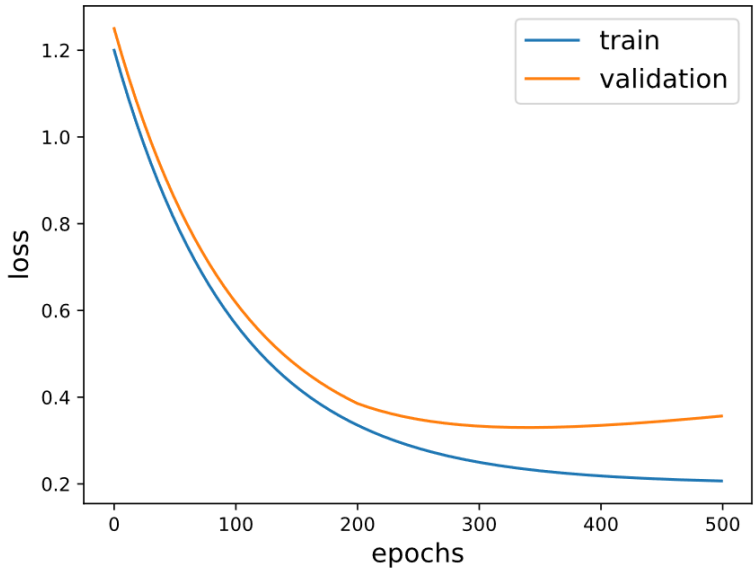


Figure 2.9: Training and validation set loss during ANN training.

subsets, whereas a larger batch size is more likely to get stuck in a local optima.

During ANN training, the goal is to minimize the cost function. However, this can lead to over fitting, where the learned parameters, ANN weights and biases, perform well on the training data but poorly on unseen data. To this end, many techniques split the training dataset into three parts, training, validation and test. During training, the ANN is trained on the training set and is also evaluated on the validation set. When the error of the validation set is at a minimum, ANN training stops to prevent early stopping. An example of overfitting is seen in Fig. 2.9 where the ANN training loss consistently decreases over time, but the validation loss eventually begins to increase over time, resulting in an asymmetric U-shaped loss curve. The performance is reported on the test set.

This section focused on feedforward ANNs. There exists another family of ANNs, termed recurrent neural networks (RNN)s. RNNs are built to process a sequence of values $x^{(1)}, ...x^{\tau}$ because they can recognize sequential patterns in the presence of sequential distortions. RNNs have recurrent connections, see Fig. 2.10, that allow a memory of a previous input to persist in the network's internal state. [39].

Similar to an RNN, there exists nonlinear autoregressive networks with exogenous inputs (NARX). The information presented regarding NARX models is summarized from

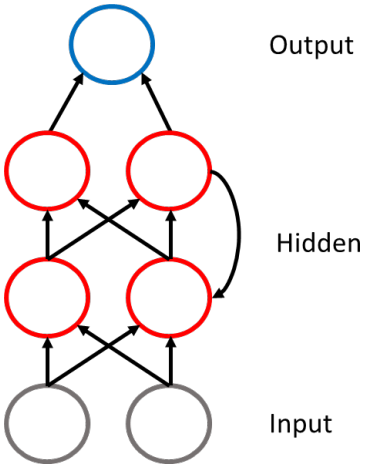


Figure 2.10: An RNN structure. The connection between the second and the first hidden layer makes this an RNN.

[114]. These models are defined as:

$$y(t) = g(y(t-1), ..., y(t-n_y), u(t-1), ..., u(t-n_u)) \tag{2.5}$$

where $u(t)$ is the input and $y(t)$ is the output. n_u and n_y are the input and output order, respectively. Fig. 2.11 shows a diagram of a NARX model.

NARX has a feedback connection from the output to the input whereas an RNN has connections only amongst neurons in the hidden layer. Generally, NARX can be trained as a feedforward ANN, where the measurements are fed as inputs and the estimated output is

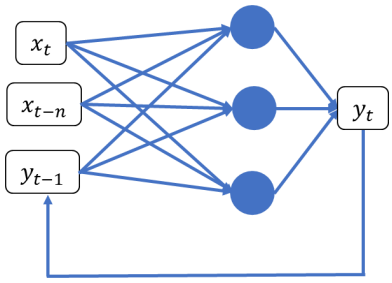


Figure 2.11: A NARX model. The feedback connection shows how the output is fed back to the input.

compared with the corresponding measured output. When used to estimate future values, the NARX outputs can be fed back to the model to continue estimating the sequence.

2.3.3 Modelling Blood Flow

Most approaches that use data-driven models to model blood flow involve imaging blood flow with magnetic resonance or computed tomography. Generally, these studies often focus on modelling blood flow for one individual. Sun et al. [115] predict intra-cardiac blood flow for individuals at rest using long axis cine magnetic resonance imaging with a convolutional neural network (CNN) and validate their results with 4D flow magnetic resonance images. The CNN shows a negative bias, the model predicts velocity than is 30% lower than the measured values. Zhang et al. [136] build a venous finite element model (FEM) for blood flow using computed tomography images from a rabbit. The computed tomography image provided the dimensions of the veins which is necessary for the FEM. The imaging required for both of these methods is resource intensive if building a user-specific blood flow estimation model for a group of participants. However, because blood flow is related to the size of an individual's arteries and veins, these imaging approaches are commonly used in the literature.

Landry et al. predicted BV one heart beat ahead using a NARX model [63]. The data collection protocol is described earlier in this chapter, see section 2.2.3. In this work, a nonlinear autoregressive neural network (NAR) model was trained to predict future ECG using past inputs of ECG and a NARX model was trained to predict future BV using past inputs of BV, applied pressure, and the predicted ECG from the NAR model. This model was trained using 19 minutes of collected data. The NAR model could not accuracy predict the following ECG R-wave, though it had good performance up to that point. Consequently, the predictions were limited to one heartbeat ahead. When predicting BV the model showed good qualitative performance, in that the NARX model could follow changes in the BV shape for different CTs. The mean R^2 for the BBMV was 0.76 ± 0.11 across the 4 participants. The NARX model also has good performance when comparing the mean BBMV for different CTs across participants. A limitation of this study is that only 1 minute of each CT was included in the data collection, and consequently the model training data. It is unknown if 1 minute of the BV response to a given CT is enough to capture the response or if a longer time period is required. The NARX model presented in this study showed that BBMV can be modeled using a nonlinear function of past ECG, applied pressure, and BV.

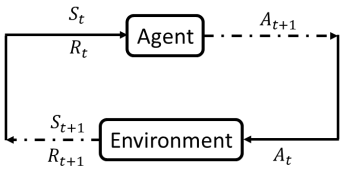


Figure 2.12: The agent–environment interaction in a MDP.

2.4 Reinforcement Learning

The information herein is summarized from [116]. RL is learning how to map a situation to an action to maximize the cumulative reward. During training there are no explicit instructions for behavior, rather this mapping is learned by interaction with the environment.

RL problems are formulated as a Markov decision process (MDP); a mathematical framework of sequential decision making. The agent is the learner and it interacts with the environment. In addition, there are three other main elements of an RL system: (1) policy π; (2) reward signal R_t; and (3) value function V. The policy, π defines how the agent will behave at a given time step; it maps observations to states. The agents' objective is to maximize the reward, R_t that it receives from the environment and the reward is also a key signal in the policy update. For example, if the policy selects an action which is followed by a high reward, the policy may be updated to select that same action in that same situation in the future. The value function, $V^{\pi}(s)$, tells us, starting from state s and following policy π, the cumulative reward the agent can expect. While a reward evaluates the immediate quality of a state, value functions evaluates a state's long term quality considering likely future states. In an RL problem, the agent should be able to observe its state within the environment; a measurement of how the environment is at a particular time step. DRL uses an ANN to approximate the policy or value functions.

At each time step, the agent observes its state, S_t and selects an action A_t, the environment responds by updating the agent's state, S_{t+1} and giving a reward R_{t+1} to the agent. Fig. 2.12 presents the agent-environment interaction in an MDP.

Consider the task at hand, maximizing BBMV by selecting the optimal CT. To formulate it as an MDP, the environment is the lower limb physiological environment. The action is the CT selected by the agent at each heartbeat such that average BBMV $\sum R_t$ is maximized over an episode. A state representation could include information about HR, BBMV, and previous CT selection.

MDPs assume that the agent can observe the entire state. However, physical systems

have noisy sensors or signals that cannot be observed such that the MDP assumption fails. To solve this problem of uncertainty in observing the environment, partially observable Markov decision processes (POMDP)s are a mathematical framework of sequential decision making under uncertainty [93]. In a POMDP, the assumption made in MDPs that the entire state can be observed is relaxed. Instead agent receives some observation, that either does not include the entire state or the observation is correlated with the state. This uncertainty makes these problems more challenging to solve and is an ongoing area of research [92, 4].

2.4.1 Reinforcement Learning Algorithms

This subsection summarizes information presented in [134, 1]. There exist many methods to solve RL problems. While there is no universal taxonomy for RL algorithms, we will attempt to classify the different algorithm types for clarity. Broadly, RL algorithms can fall into 2 categories, model-free and model-based. The model refers to the dynamics of the environments, the transition probabilities, and the reward function. In a model-based environment, the agent either has access to a model of the environment or this model is learned through interaction. For example, the AlphaGo algorithm uses the rules of the game Go such that the agent knows the true environment. A benefit of model-based RL algorithms is that the agent can plan ahead. The limitation is that often, and especially for real-world environments, the agent cannot access the true environment. Consequently, during training the agent can capitalize on biases in the learned model such that when deployed in the real world, agent performance is suboptimal despite optimal performance in simulation. Because of the challenges associated with deploying models trained in simulated environment in the real world (sim-to-real), model-free methods have grown more popular in recent years.

Within model-free methods, there exist two categories, policy-based and value-based. In policy-based methods, the algorithm optimizes the parameters, θ, of the policy $\pi_\theta(a|s)$. Often, these are on-policy methods, that is, each update uses data that is collected by acting according to the most recent version of the policy. This method provides stability because policy updates use the most recent data. Common policy-based methods include trust region policy optimization (TRPO) [103] and PPO [104].

Value-based methods learn an approximator for the optimal action-value function, $Q(s, a)$. Typically, these methods optimize based on the Bellman equation and the optimization is performed off-policy. In off-policy methods, each update uses data that can be collected at any point during training, regardless of the policy used to collect that data. The policy is obtained by the relationship between Q and π. Value-based methods are

more sample efficient than policy-optimization methods since data can be reused, however this benefit comes at the cost of stability [117]. In Q-learning, the policy can be obtained from the relationship between Q^* and π^*. The Q-learning agent selects actions using the following equation:

$$a(s) = \arg\max_a Q_\theta(s, a) \tag{2.6}$$

The update for Q-learning is given by:

$$Q(S_t, A_t) = Q(S_t, A_t) + \alpha[R_t + \gamma \max_a Q(S_{t+1}, A_{t+1} - Q(S_t, A_t)] \tag{2.7}$$

Common examples of value-based methods include deep Q-networks (DQN) [82] and categorical DQN (C51) [7].

Policy-optimization methods and Q-learning methods can sometimes overlap. These methods can combine the benefits of value-based and policy-based methods by using the value function to improve sample efficiency and using policy-based methods to learn the policy function. Examples include deep deterministic policy gradient (DDPG) [71] and soft actor-critic (SAC) [44].

Vanilla policy gradient (VPG) [116] is an RL algorithm that is focused on increasing the probability of actions that increase returns and decreasing the probability of actions that decrease returns until convergence to an optimal policy. In other words, VPG performing gradient descent on policy parameters θ to improve policy performance. VPG trains a stochastic policy on-policy. The randomness of action selection decreases over time since the update promotes the agent to exploit previously found rewards. This behavior can cause the policy to get converge to local optima.

TRPO builds upon VPG by taking the largest possible step to improve performance keeping the old and new policies sufficiently close. The distance between the old and new policies is measured by the KL-divergence, which can be considered as a measure of the distance between probability distributions. VPG is sensitive to step size since the gradient provides only first-order information. If the step size is large, the algorithm may suffer performance collapse, while if the step size is small learning suffers from poor sample efficiency and convergence is slow. TRPO avoids this collapse and convergence is faster than VPG.

To compensate for the high computational cost of TRPO, proximal policy optimization (PPO) takes the same approach as TRPO, at a high level, but is less computationally

expensive. PPO clips the policy update to keep the step size small so that training is stable. PPO samples data through agent-environment interaction using the most recent version of the policy. It is a model-free, policy-optimization algorithm. PPO optimizes a clipped surrogated objective function with stochastic gradient descent. At each time step, the agent observes the state and estimates the probabilities of selecting each action. Action selection is based on this probability distribution. The agent interacts with the environment at each time step and after a set number of time steps, determined by the experience horizon, the agent updates the policy with the clipped loss function equation below [104]:

$$L_t^{CLIP} = \mathbb{E}_{min}\Big[rA_t^{GAE}, \text{clip}\Big(r, 1-\epsilon, 1+\epsilon\Big)A_t^{GAE}\Big] \tag{2.8}$$

where A_t^{GAE} is the advantage function, where calculates how much better taking that action at a state is compared to the average value of the state over all actions. ϵ is a small hyperparameter that clips the update referred to as the clip factor. $r = \frac{\pi(a_t|s_t)}{\pi_{old}(a_t|s_t)}$ and it is the ratio of the action probability for the current policy and the old policy, which was used to generate the data. Consequently, if the advantage, A_t^{GAE}, is positive, the probability of selecting the given action increases. The opposite is true if the advantage is negative. In both cases, the clip factor limits the size of the policy update to ensure stability. The Value ANN is trained with supervised learning using with the update function where $(V_\theta(s_t) - V_t^{targ})^2$. The target value is computed using the discounted cumulative reward over the experience horizon.

There are many hyperparameters that need to be set for a PPO agent. The experience horizon, is the number of steps that an agent takes before it updates the policy. The discount factor, $\gamma \in [0, 1]$ is a measure of how much the agent cares about future rewards. The higher the γ value, the more the agent cares. The clip factor, $\epsilon \in [0, 1)$ is how much the policy updates are limited by. The entropy loss weight is a scalar value between 0 and 1 that promotes agent exploration by applying a penalty for being too certain about action selection.

2.4.2 Reinforcement Learning Applied to Simulated Physiological Systems

In previous work, DRL has been applied to simulated neuro-musculoskeletal models to generate natural locomotion [130], and in physiological systems for diabetes management [88, 67], and anemia treatment [24].

To achieve success simulated physiological systems, which are often highly complex, DRL algorithms require modifications [5, 121, 31]. Weng et al. [130] used curriculum learning, breaking down the goal into learning tasks for the agent to learn to walk. Lee et al. [67] modified the PPO algorithm by introducing two value functions since the long term control for the fasting state and the short-term control for moderate changes in the physiological system required the assumption of different levels of insulin administration. Both Escandell-Montero et al. [24] and Ngo et al. [88] found success with off-the-shelf DRL approaches.

Each of the previously mentioned works assume fixed dynamics in their learning environments. Of the four works mentioned above, Lee et al. [67] is the only work that explicitly includes time-dependency in their physiological system. However, there is still a clear reality gap in this work since they do not account for the effect of exercise on blood glucose. While Weng et al. [130] do not explicitly introduce variability via time-dependency into their physiological system, gait symmetry and stride frequency are not strictly constrained which introduces variability to the physiological system. The difference in simulated versus real-world variability is a common limitation across different DRL approaches to simulated physiological systems.

2.4.3 Reinforcement Learning: Sim-to-Real Transfer

Applying DRL systems from simulation to physical systems in the real world is non-trivial. Despite best efforts, simulations often do not capture the variability of the real world and optimizing with real world systems is often resource intensive. There exist two main approaches to facilitate the process of sim-to-real transfer.

The first is to develop a simulated model of the physical system and to pre-train a policy in this physical system. After, the policy continues training in the physical system. Pre-training in simulation can help the agent learn a mapping of states to actions and is more efficient than completing all training in the real world. This approach has been used previously in autonomous navigation of stratospheric balloons [8] and robotic table tennis [29]. One limitation of this approach is that the policy may not observe all states during simulation and when deployed in the real world can select sub-optimal actions in these new states. Another drawback of this approach is that the pre-trained policy is limited by the accuracy of the model. Depending on sensor availability and the complexity of the physical system, this simulated environment may not be similar enough to the ground truth. Consequently, additional steps that introduce more variability to the simulation are required to bridge this simulation to reality gap [49]. However, to introduce variability

to the simulation in a manner that mimics the physical environment requires domain knowledge and still may not replicate the real world system.

The second approach is to develop a simulated model of the physical system and use it as a learning environment to learn the best hyperparameters for real-world DRL training [128]. A limitation of this approach is that starting with a randomly initialized policy may result in long training times in the real world environment.

Chapter 3

Modelling Blood Velocity[1]

This chapter develops BV estimators and evaluates their accuracy over a longer time-scale. Six participants the CGC device, developed at the University of Waterloo, and experienced random cardiac-gated compression timings for 1.5 hours per day for two days. A data-driven model was trained on ECG and applied pressure data to predict femoral BV one heartbeat ahead in a closed loop manner. The mean R^2 for this model across participants on the second session was 0.74 ± 0.09 and the mean absolute error was approximately 3%, which is a reduction of only 11% compared to the first sessions, for both metrics. This study is the first to show that BV across IPC sessions can be predicted using a pre-trained model. This work may lead to a significant improvement in IPC performance with only an initial model training session.

3.1 Methods

3.1.1 Experimental Procedure

Hemodynamic responses to compression were assessed in six individuals (mean ± SD age 28 ± 3 yrs) with no history of cardiovascular or peripheral vascular disease. Participants were in an upright seated position and were asked to remain still to reduce motion artefacts.

[1]Material in this chapter was previously published in:
I. Santelices, C. Landry, A. Arami, and S.D. Peterson. Investigating Optimal Intermittent Pneumatic Compression Timing Across Two Days. In Proc. Int. Conf. Body Sensor Netw, Oct. 2023.

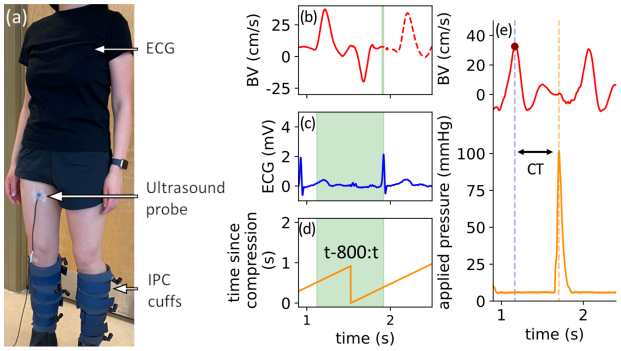


Figure 3.1: (a) Experimental setup. Participant is shown standing for visual clarity, however during data collection participants were seated, (b-d) sample BV, ECG, and computed CT signals that served as inputs for the neural network, respectively. The shaded region shows the time steps fed in as inputs. (e) A typical BV and applied pressure signal. Double-ended arrow shows the CT from the local pulse wave peak (dark red point) to peak inflation of the first cuff.

The experiment consisted of a 5-min baseline prior to the application of compression followed by two windows of IPC therapy separated by a minimum of 5 min, referred to as parts A and B. Part A lasted for 45 minutes and part B for 21 minutes. The phase lag between the ECG R-wave and the local blood pulse was recorded and fixed prior to data collection for each subject and used to ensure that compression occurred in local diastole. Herein, CT was defined as the time delay from the local blood pulse peak to the first cuff's peak inflation, see Fig. 3.1(e), and was changed every three minutes. In part A, the CTs considered were in increments of 50 ms from 200 ms to 850 ms. In part B, the CTs were in increments of 100 ms from 200 ms to 800 ms. The order of CTs in A and B was randomized between participants. Study protocols and procedures were approved by the University of Waterloo (ORE#44494) Clinical Research Ethics Committee and conformed with the Declaration of Helsinki.

The CGC device used in this study consists of five cuffs, each encasing an air bladder (see Fig. 3.1(a)). The air bladders inflate sequentially from ankle to knee. The compression cycle lasts 199 ms with a constant delay of 25 ms between sequential cuff inflation. Each cuff is inflated over 33 ms, held constant for 33 ms, then deflated over 33 ms. Because of the length of the each compression cycle and the variability in heart rate across participants and during a session, the compression did occur in systole for some heartbeats.

Pressure sensors were placed between the participant's anterior lower leg and each of the three bottom cuffs. The sensors comprised of Picopress bladders (Microlab Electronica,

Canada) connected to Omega PX390 pressure transducers (Omega Engineering, Canada). Before testing, the system was adjusted to apply approximately 100 mmHg of pressure to the limb. SFA BV of the right leg was measured with a 4-MHz Doppler probe (Multigon Industries, USA), see Fig. 3.1(a). Arterial blood flow is assumed to be proportional to BV, as previous studies have shown that arterial diameter is not impacted by IPC [141]. A standard 3-lead chest ECG (Model 3000 AC/DC differential amplifier, A-M Systems, USA) was used to time the CGC device to the cardiac cycle. These signals were collected using a data acquisition system (PowerLab, AD Instruments, USA) connected to LabChart software (version 7, AD Instruments, USA).

3.1.2 Model Architecture and Training

A NARX model, $NARX_{bv}$, was used to predict BV [65]. The model takes the previous time steps of measured BV, ECG, and time since compression occurred (sampled at 100 Hz) to predict the future BV, as shown in Fig. 3.1(b-d). We modified the model presented in [63] by increasing the network depth from 1 to 3 layers (of 200, 120, and 40 neurons), with ReLU activation functions in the hidden layers and linear activation for the output layer. The model was trained as a feedforward ANN with a mean squared error loss function with Adam optimizer. The model was trained on part A from day 1. Part B, from either day, was used for model testing. The $NARX_{bv}$ model estimated BV one heartbeat ahead on the test set in a closed-loop fashion.

3.1.3 Performance Evaluation

The mean BV between two ECG R-waves, herein referred to as BBMV, was used to define the optimal CT. HR was computed from the ECG signal. To evaluate the $NARX_{bv}$ model performance, BV mean absolute error (MAE), and BBMV R^2 were computed for the test sets of each participant. The Spearman correlation coefficient between the measured mean BBMV for each CT and the estimated mean BBMV for each CT is reported. The CT is presented as a percentage of the local BV period in which the compression occurred, as displayed in Fig. 3.1(e). Wilcoxon signed-rank tests ($\alpha = 0.5$) with Bonferroni correction were used to assess statistical significance.

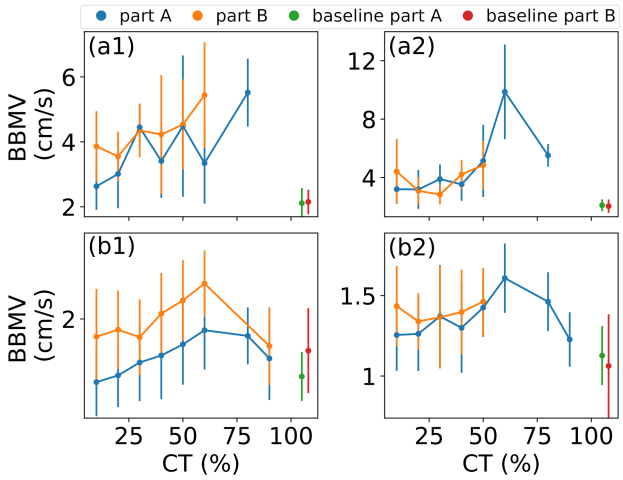


Figure 3.2: BBMV as a function of the CT. (a1) Participant 1, day 1; (a2) Participant 1, day 2; (b1) Participant 6, day 1; and (b2) Participant 6, day 2.

3.2 Results

Figure 3.2 shows the BBMV response for different CTs for 2 participants. These participants were chosen to show a range of BBMV responses over time. The applied CT range is different across participants since CTs were applied as fixed time delays from the systolic BV peak instead of as a percentage of time between R-waves. Because of variability between the fixed time delay and the CT as a percentage of the local blood pulse, there are a different number of CTs plotted for each participant while the protocol applied the same number of CTs for each participant. The BBMV response to CTs is shown in CT increments of 10% in Fig. 3.2. Increments of 10% allowed for each BBMV response to be computed over a minimum of 90 heartbeats while smaller increments would show the BBMV response across more CTs, each data point would correspond to a fewer number of CTs. This figure highlights that the CT leading to the greatest increase in BBMV is generally not constant over time, with optimal CT changing between parts A to B and from day 1 to 2, with the exception of Fig. 3.2(b1), the only instance in this study where the optimal CT remained constant.

The $NARX_{bv}$ model was used to estimate BV one heartbeat ahead in a closed loop manner on the test sets. Representative samples of estimated BV waveforms compared to the measured waveforms from same-day and different-day testing are shown in Fig. 3.3. The estimated BV waveform is in good agreement with the measured BV waveform for all CTs across participants when trained and tested on the same day. The average BBMV R^2

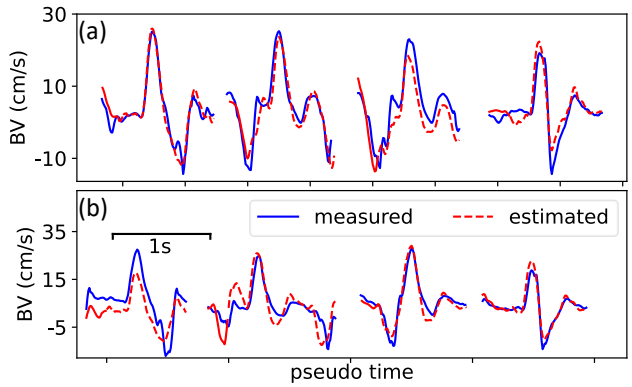


Figure 3.3: Measured (blue) and estimated (red) BV for different CTs of, from left to right, 20%, 40%, 60%, and 80% for Participant 3 on (a) same-day training and testing; and (b) different-day training.

was 0.83±0.08 (mean ± SD) across participants. BBMV estimation performance decreases when the model is trained and tested on separate days, with the average R^2 reducing to 0.74 ± 0.09.

The pooled estimated BBMV (mean ± SD) is compared against the measured BBMV for 2 participants as a function of CT in Fig. 3.4. The mean Spearman correlation coefficients across participants for same-day and different-day training and testing were 0.80 ± 0.08 and 0.72 ± 0.14, respectively.

Table 3.1 shows the large variability in results across participants for the different metrics. Across participants, the BBMV increased or stayed the same when selecting the CT based on the model compared to applying IPC in the start of diastole, 10% CT (the typical fixed timing approach) [26, 140, 141, 139, 34].

3.3 Discussion

This paper presents a NARX model architecture to predict the SFA BV response to lower limb IPC. The NARX model predicts BV one heartbeat ahead accurately over time as shown with different-day IPC session data, with a mean BBMV R^2 of 0.74 ± 09, and a mean Spearman correlation coefficient of 0.72 ± 0.14 between measured and predicted BBMV. This study's data supports the hypothesis that the optimal CT varies over time, but can still be predicted with fixed model dynamics as shown in Fig. 3.2.

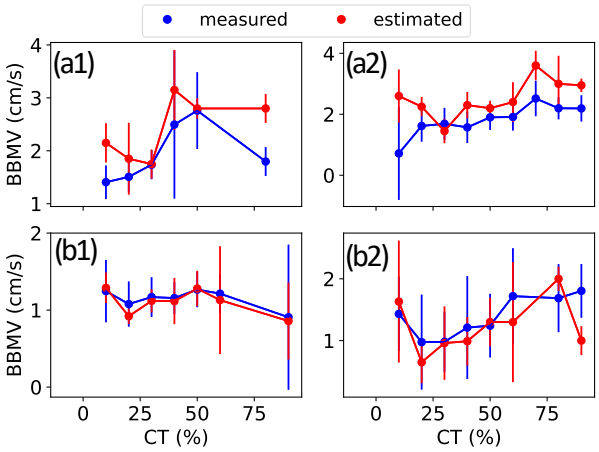


Figure 3.4: Measured (blue) and estimated (red) BV (mean ± SD). (a1) Participant 2, day 1, Spearman correlation coefficient = 0.68; (a2) Participant 2, day 2, Spearman correlation coefficient = 0.60; (b1) Participant 4, day 1, Spearman correlation coefficient = 0.94; and (b2) Participant 4, day 2, Spearman correlation coefficient = 0.62.

Landry et al. [63] suggested that the optimal CT varies over the course of a day; we observed similar results with changes in BBMV response from part A to part B and across participants. Our BBMV R^2 of 0.83 ± 0.08 on same-day data is higher than their reported $R^2 = 0.76 \pm 0.11$ when predicting BV one heartbeat ahead, which might be caused by a different testing protocol (fixed timing for 3 min vs 1 min) and different participants. Moreover, their model first predicts ECG in a closed loop fashion and feeds this signal to the NARX BV model; the two models in series might have added error to the predictions. In this work, we focused on predicting only BV; the ECG model is left for future work. Also, this work trained the NARX model on 45 minutes of data, as opposed to 20 minutes, which can help the model to better learn the BV dynamics.

The intended application of the NARX model is to be part of a controller used to select the CT every heartbeat. For instance, the BV responses to different CT can be simulated before the compression occurs and the optimal one is selected. Our work indicates that the NARX model could effectively be used as it is intended since it can reasonably rank the CTs, as observed with the mean Spearman correlation coefficient of 0.72 on different-day IPC session.

According to this study's data, using the model to select the optimal CT as opposed to 10% CT would result in an average 30% and 62% increase in BBMV, shown by Table 3.1(F), on same-day and different-day data, respectively. The larger BBMV increase for a

Table 3.1: Results of $NARX_{bv}$ model for training and testing on same and different days. Results shown are for each participant for $NARX_{bv}$ model trained on day 1. (A) BBMV R^2; (B) Spearman correlation coefficient; (C) BV MAE (cm/s); (D) Absolute change in BBMV from part A to B; (E) HR and BBMV Pearson correlation coefficient (measured data); (F) % change in BBMV from 10% CT to model optimal CT. The differences in results from same-day to next-day testing were not significant.

	A		B		C		D		E		F	
participant	same	next	same	next	same	next	same	next	same	next	same	next
1	0.81	0.78	0.75	0.70	1.98	2.14	5	8	0.41	0.43	41	10
2	0.69	0.58	0.68	0.60	3.11	3.56	29	50	0.30	0.12	77	205
3	0.93	0.77	0.82	0.77	1.78	1.84	9	19	0.51	0.51	6	147
4	0.85	0.75	0.94	0.62	1.54	1.77	21	23	0.48	0.44	42	7
5	0.85	0.83	0.80	0.80	0.89	1.42	17	19	0.60	0.62	0	0
6	0.84	0.75	0.86	0.82	1.25	1.33	13	24	0.44	0.43	12	2
$\mu \pm \sigma$	0.83±0.08	0.74±0.09	0.80±0.08	0.72±0.14	1.8±0.8	2.0±0.9	16±9	24±14	0.46±0.10	0.43±0.17	30±30	62±90

different day is due to variability in BBMV response to CT. Retraining the initial model with a small amount of data at the beginning of each session might improve the model performance over time; this is left for future work.

The mean percent change in BBMV from same-day to different-day testing, 16% to 24%, reflects, to some extent, the decrease in performance across participants across days, see Table 3.1(D). Indeed, the greater the difference between training and testing data, the more challenging it is for the NARX model to accurately predict BV. This impact is mainly seen in participant 2, who had the largest increase in BBMV percent change from same-day to different-day testing.

Correlation between HR and BBMV is also lowest for participant 2, who had the poorest NARX model performance. The Pearson correlation is highest for participants 3 and 5, who have the highest R^2 values as well. This might indicate that the NARX model is extracting heart rate as a feature to predict BV. An explainable model, with the use of physiological features, is left for future studies.

Short-term variations in blood pressure, which can result in variation in local blood flow, are not uniform and have shown large differences in magnitude over multiple half hour periods [79]. Mancial et al. suggest that central nervous mechanisms play a primary role in heart rate and blood pressure variability. The limited data collected related to central hemodynamics and IPC BV response limits our understanding of their interaction, but can potentially explain the errors associated with same-day and different-day BV predictions; the model may not have the necessary signals to estimate BV and its variation over time. In future studies, we plan to include volume clamp blood pressure measurements to learn

more about the central hemodynamic response to CT (blood pressure, stroke volume, total peripheral resistance, etc). In turn, these signals may improve the prediction accuracy of the model when used during different treatment sessions without retraining.

A limitation of this study is the measurement of BV using an ultrasound probe. A change in the probe insonation angle would result in a change in the BV measurement sensitivity, as per the Doppler equation [35], which might partially explain the change in BBMV response from part A to part B, and from day to day within a participant. This may be a factor in the data collection for participant 2, who had a large decrease from same-day to next-day BBMV, see Table 3.1. Another limitation is that the BV measurement contained artefacts because of the sensitivity of the ultrasound probe with soft tissue measurements and leg movement due to the IPC. Small differences in the pressure profile from one day to another is another limitation. Participant stress levels during data collection were not measured, which can impact BBMV.

3.4 Conclusion

It was shown in this study that the BV response to IPC CT for an individual changed over time and across days, but can still be predicted in a one heartbeat ahead fashion from the ECG and applied pressure signals with a pre-trained NARX model. This NARX model has the potential to be used to optimally select the CT for each individual, which might lead to a significant increase in IPC performance without any hardware modifications.

Chapter 4

Evaluating Deep Reinforcement Learning to Optimize Intermittent Pneumatic Compression Timing

In this chapter, we investigate the optimal CT for an individual in simulation. We explore how different DRL agents can learn the optimal CT in simulation.

4.1 Methods

4.1.1 Environment Formulation

The learning environment is a user-specific predictive simulator of lower limb physiology created in MATLAB 2022a. The process of collecting data and training this predictive simulator, that consists of a BV estimation model, is described in chapter 3.

In the learning environment the DRL agent has a time step of 5 heartbeats so that metrics can be computed on a heartbeat-to-heartbeat basis and so that an agent time step is less impacted by BV variability over successive heartbeats. We formulate this problem as an MDP, defined as a tuple (S, A, T, R, γ) [116]. S, is the set of states, where each state is a four dimensional vector comprising of the average value and standard deviation (SD) of HR and the BBMV, $[HR_\mu, HR_\sigma, BBMV_\mu, BBMV_\sigma]$. These values are computed over the last 4 heartbeats in each agent time step to avoid residual effects from the previous time step. Typical state values for a participant are shown in Fig. 4.1 (a)-(b). The set

of actions, A, is a discrete action space of 19 different CTs from 5% to 95% in steps of 5%, see Fig. 4.1(d). These 19 CTs were chosen to be similar to [63]. While this is a discrete optimization problem, we refer to it generally as an optimization problem in this document. Though each agent time step is 5 heartbeats, compression is applied at each heartbeat, see Fig. 4.1(c). A discrete action space was chosen because of the variability between command CT and actual CT. The BBMV for different CTs is simulated using the predictive simulator. State transition probabilities (T) depends on an individual's central hemodynamics and vasculature. The reward function, R, is computed using BBMV, since we use BBMV to define the optimal CT for the DRL agent, and it is described in more detail later in this section. $\gamma \in [0,1]$ is the discount rate and is related to the time horizon of the agent's optimization. In this physiological system the transition probabilities are complex and generally unknown because they are a result of an individual's hemodynamic response to IPC CT which is stochastic and may vary with time. DRL allows us to solve this MDP without knowing the probability of these state transitions.

The agent-environment interaction is shown in Fig. 4.2. The BBMV response to a selected action A_t, is estimated with $NARX_{bv}$, details about $NARX_{bv}$ structure can be in found in [101]. The action, A_t, is defined as the CT for the given time step. Successive estimates are required to compute BBMV and the time step for the $NARX_{bv}$ model is $T_N = 0.01s$. The environment, with a time step of $T_E = 1$ heartbeat, repeats this process for 5 heartbeats. The environment outputs the updated state, S_t, and reward, R_t. The agent receives the updated state and reward, performs the updates, and outputs the next action A_{t+1} (see Fig. 4.2). This sequence is repeated until the end of the ECG signal, which spans approximately 1 hour.

Three different reward functions are explored:

$$R_1 = BBMV_t \tag{4.1}$$

$$R_2 = BBMV_t - BBMV_b \tag{4.2}$$

$$R_3 = BBMV_t - (0.75BBMV_b + 0.25BBMV_{t-1}) \tag{4.3}$$

where $BBMV_t$ is defined as the mean BBMV over the previous four heartbeats. $BBMV_b$ is defined as mean BV during no compression, before training occurs. This baseline is included to explore the impact of a shift in the mean reward. $BBMV_{t-1}$ is defined as the mean BBMV over the last four heartbeats in the agent's previous time step. This is

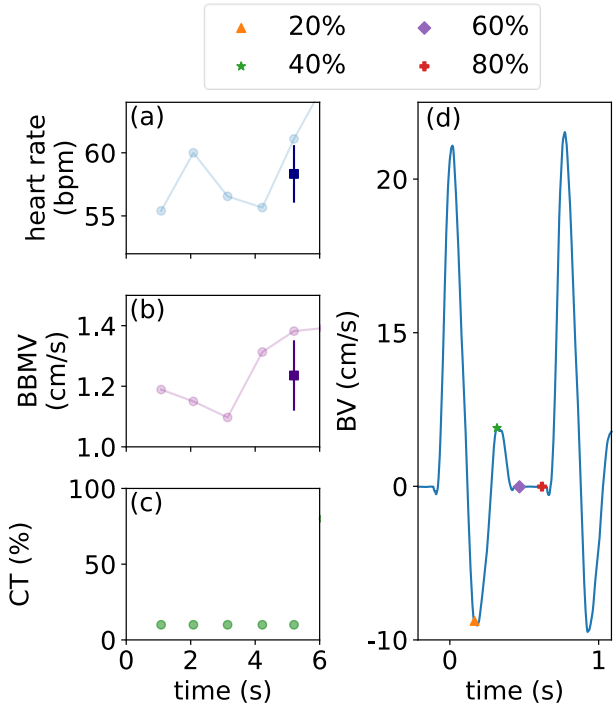


Figure 4.1: State space for agent. (a) Mean and SD of HR is shown in the dark blue square. The faded blue circles show the HR for each heartbeat. (b) Mean and SD for BBMV is shown in dark purple square. The faded purple circles show the BBMV for each heartbeat. (c) CT applied. Compression is applied at each heartbeat and the CT is held constant over one agent time step of 5 heartbeats. (d) Agent action space. Delays from 20% to 80% at 20% intervals are shown for clarity.

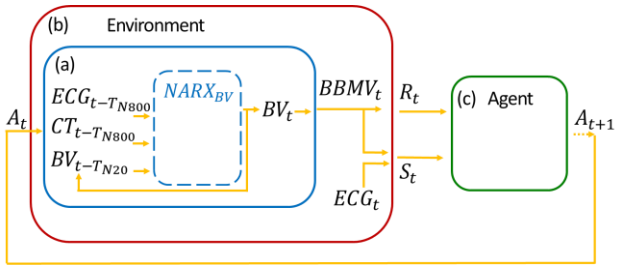


Figure 4.2: Agent environment interaction. The BV estimation model takes inputs of previous ECG, CT, and BV to predict future BV in a closed loop format. T_{N800} is the defined as t-800 ms to t-10 ms, T_{N20} is the previous t-20 ms to t-10 ms. These time steps form the input for the $NARX_{bv}$ model.

included so that the DRL agent is incentivized to continually increase BBMV. R_3 asymptotically converges to $0.75R_2$.

4.1.2 Training Strategy

To learn the optimal CT for an individual, we use PPO [105], which requires 2 ANN. This algorithm includes an ANN for stochastic policy, π_θ, that maps physiological states to the probability of selecting CTs, and a Value ANN, $V(s)$, that estimates the expected return for a given physiological state, S_t. 2-layer-64-neuron ANNs with tanh activation functions are used for both networks. The Policy and Value networks use softmax activation and linear activation for their output layers, respectively. The network layer depth and width were selected based on our empirical findings to ensure sufficient network representation capability and learning efficiency.

During training, the ECG and CT are inputs to the predictive simulator to predict the BV response, which is used in both the DRL state representation, S_t, and the reward signal R_t. For each participant, we train 5 DRL agents on a 60 minute ECG signal.

4.1.3 Hyperparameter Tuning

In the final results presented for R_1, R_2, and R_3 the entropy, w, is 0.01 and the experience horizon is 2 steps (10 heartbeats). The batch size must be equal to or lower than the experience horizon and is consequently set at 2 steps (10 heartbeats). The (actor and

critic learning rates) for R_1, R_2, and R_3 are (0.003, 0.005), (0.007, 0.03), and (0.008, 0.05), respectively.

These hyperparameters are tuned with grid search. Entropy is explored from $w = 0.01$ to 0.2 in steps of 0.01. The horizon determines how many agent time steps occur before each Policy and Value function update. We explore values from 10 heart beats to 210 heartbeats in steps of 50 heartbeats. The actor learning rates explored are from 0.001 to 0.01 in steps of 0.001. The critic learning rates explored are from 0.005 to 0.09 in steps of 0.005.

4.1.4 Policy Validation

The BBMV response to a constant CT is estimated using the predictive simulator, for each of the possible CTs with a 5 minute ECG signal. After the DRL agent is trained on the 60 minute signal, the DRL continues to train on this 5 minute ECG signal, selecting CTs and observing the estimated resultant BBMV; this is referred to as the test performance. The optimal CT sequence that the DRL agent selects is referred to as CT^*_{PPO}. The optimal CT sequence selected by the predictive simulator, where it selects the CT with the highest BBMV at each heartbeat, is CT^*_N. The fixed CT, CT_f, is a 10% CT that occurs in the beginning of diastole. These CTs are used to evaluate the DRL agent; this value was selected based on previous work [26, 140, 34, 141, 139].

To represent a change in BV response over time, we change the predictive simulator of the participant during DRL agent training. The BV estimation model in the first 30 minutes is trained on one day of BBMV data, predictive simulator A (PS_A), and the BV estimation model in the remainder is trained on a separate day of BBMV data, predictive simulator B (PS_B). We explore the effect of changing the predictive simulator on agents with the following entropy values: $w = 0.01$, 0.05, 0.1, 0.2, 0.3, and 0.4.

4.2 Results

4.2.1 Training

Fig. 4.3 shows the moving average (3 minute window) of the BBMV obtained by the agent during training across the different reward functions for each participant. The mean and SD shown are across 5 trials.

Using R_1, the agent does not consistently learn the optimal CT in 60 minutes of simulation time across participants. Participants 1 and 4, (Fig. 4.3(a),(d)), showed signs of learning closer to the end of the therapy session, however, the other participants showed only modest increases in BBMV during training. Employing R_1 also leads to more variability during training, seen especially with participants 3 and 4, see Fig. 4.3(c),(d). This is likely explained by more CT exploration due to the smaller actor and critic learning rates. R_2 resulted in consistent learning across participants. For participant 2, training with R_2. the agent learned the optimal CT for each trial, see Fig. 4.3(b). When training with R_2, the other participants had at least one trial where the second most optimal CT was selected; the agents did not converge to the same action across trials. The DRL agent using R_3 consistently increased BBMV in 20-minute simulations across participants. For participants 1, 2, and 4, R_3 resulted in learning the optimal CT for each trial. For participants 3, 5, and 6 there were some trials where the agent did not learn the optimal CT, but instead the second most optimal CT.

The BBMV SD across trials when R_1 is employed is larger than for R_2 and R_3, across all participants. Due to the smaller actor learning rate, the R_1 agent was exploring more actions for longer during training. This behavior led to greater reward variability across trials. The SD for R_2 is larger than the SD for R_3. In some cases, R_2 converged to the same behavior as R_3, but required more trials, see Fig. 4.3(b),(c),(f). In other cases, employing R_2 resulted in more variability across trials than R_3, see Fig. 4.3(a),(d),(e)). The SD for R_3 is 0, after approximately 20 minutes of training for participants 1, 2, 4, and 6. Across the 5 trials, these agents converged to the same CT after 20 minutes of training and each agent receives the same reward at each time step. The SD for R_3 is largest at the beginning of training, when the agent is exploring the action space and it decreases with time.

4.2.2 Test

The left column in Fig. 4.4 shows the BBMV response for different CT on the test ECG signal. Participants 1, 2, and 4 have the largest SD in BBMV response to different CTs. Participant 4 in particular has a larger difference in BBMV between its optimal CT (90%) and the other CTs. Across participants, we observe that the BBMV response is not constant and also that the response varies from 1.2 cm/s to 6.0 cm/s.

The right column in Fig. 4.4 shows the mean performance of the DRL agent for three reward functions over a 5 minute period compared to CT_f and CT_N^*, as determined by the predictive simulator. Across participants, CT_f is not equal to CT_N^*. The DRL agent performs, on average, better than CT_f by 1.11 cm/s $\pm$ 1.00 (44% $\pm$ 36), p = 0.03 for R_3,

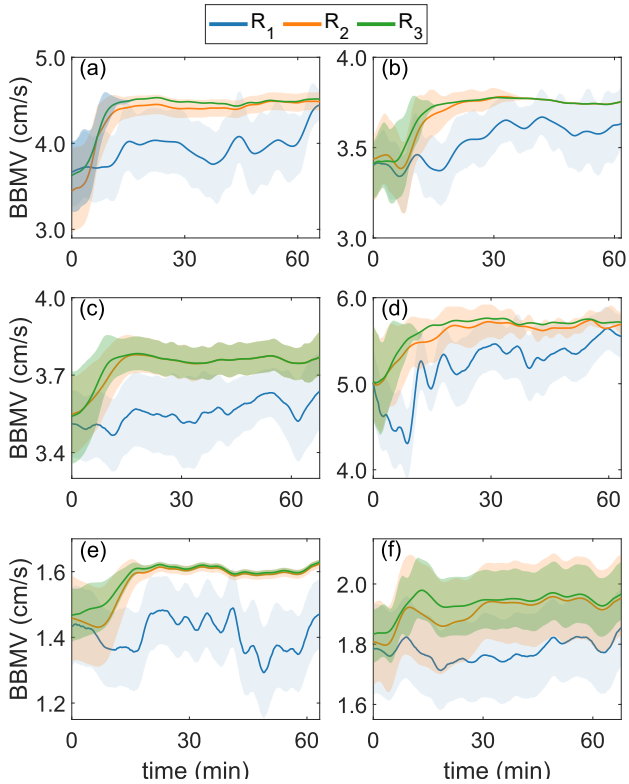


Figure 4.3: Mean ± SD of BBMV during training for agents with different reward functions across 5 trials. Results are shown for R_1, R_2, and R_3 for different participants. (a) participant 1; (b) participant 2; (c) participant 3; (d) participant 4; (e) participant 5; (f) participant 6.

respectively and lower than CT_N^* by 0.05 cm/s $\pm$ 0.05 (2% $\pm$ 1), respectively. The error bars on R_1, R_2, and R_3 show the BBMV SD across trials. The error bars are largest for R_1, similar to the training SDs. The SD bars for participants 1, 2, and 4 are too small to be visible. For these participants, the R_3 agents selected a different CT for a maximum of 5 heartbeats (2%) across the ECG signal. CT_f and CT_N^* do not have SD bars because these values are computed using one predictive simulator.

4.2.3 Changing Physiological State

Fig. 4.5 shows the BBMV during training for agents with different entropy, along with the BBMV given by the CT_N^*, across participants and for 5 trials, when the environment changes from PS_A to PS_B after 30 minutes. The BBMV during training in PS_A is similar for different entropy values, though larger entropy values generally exhibits slower learning. Conversely, in PS_B, the difference in BBMV across entropy values is larger. In the last 30 minutes of training, agents with an entropy of $w = 0.01$ achieve 89% $\pm$ 8 of the BBMV achieved by selecting CT according to PS_B; with $w = 0.4$ the agents achieve 92% $\pm$ 5 of the BBMV achieved by selecting CT according to PS_B.

4.3 Discussion

In this work, we demonstrate that a DRL agent can learn the optimal participant-specific CT in a simulated IPC therapy session. To the best of our knowledge this is the first study that optimizes IPC timing for an individual using DRL. The inherent continuous learning and adaptability of DRL enables it to deal with changing dynamics, as the DRL agent can learn the new optimal CT when the physiology changes. Our DRL approach requires less data to interact with a user to learn the optimal CT compared to the duration of data required to train a predictive simulator that can perform a CT search to optimize IPC timing (15 minutes $\pm$ 2 versus 19 minutes with other methods [63]). Our proposed DRL approach can be further tested with a human-in-the-loop to optimize IPC timing.

Reward functions with baselines (R_2 and R_3) perform better than the reward function without (R_1), as seen in Fig. 4.3. Without a baseline, the agent receives consistently positive rewards from suboptimal CTs, as shown by the left column of Fig. 4.4. Consequently, this positive reward increases the probability of the agent selecting the same suboptimal CT for the given observation, such that the agent is less likely to explore the other CTs. This behavior is observed for participants 3, 5, and 6, Fig. 4.3(c),(e),(f), where the agent

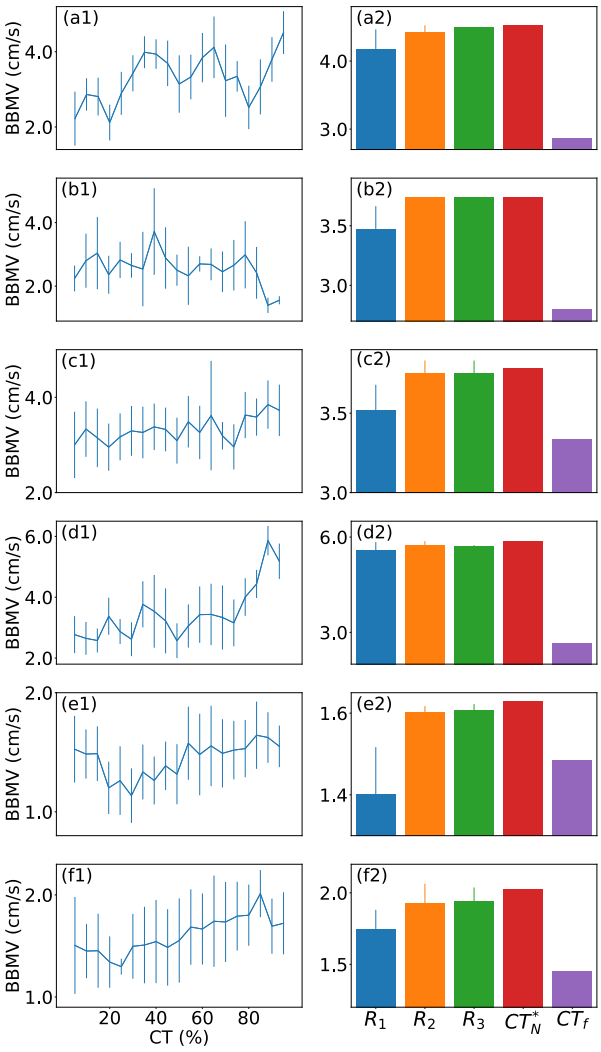


Figure 4.4: Plots (a1, b1, c1, d1, e1, and f1) show mean ± SD BV model BBMV response for different CTs. Plots (a2, b2, c2, d2, e2, and f2) show mean ± SD BBMV over 1 hour of compression for trained agents with different reward equations. CT_f and optimal CT are shown for comparison. (a1, a2) participant 1; (b1, b2) participant 2; (c1, c2) participant 3; (d1, d2) participant 4; (e1, e2) participant 5; (f1, f2) participant 6.

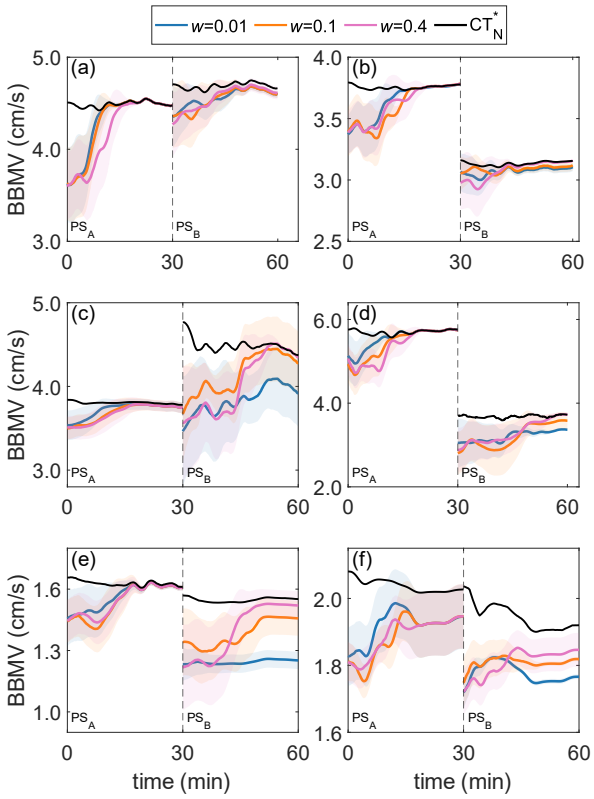


Figure 4.5: Mean ± SD of BBMV during training for agents with different entropy values across 5 trials shown for R_3. The vertical dotted line indicates when the predictive simulator changed. The black line in PS_B shows the BBMV achieved by selecting the CT with the highest BBMV. (a) participant 1; (b) participant 2; (c) participant 3; (d) participant 4; (e) participant 5; (f) participant 6.

learns a suboptimal policy. On the other hand, R_3 incentivizes the agent to perform better than the previous time step and consequently converges slightly faster compared to R_2, across participants.

The optimal hyperparameters for each reward function are outlined in the methods section. The actor and critic learning rates that led to the optimal learning were not the same for different reward functions. Because actor and critic learning rates, along with the reward range and clip factor, determine the policy update size, this is a hyperparameter that must be carefully tuned to achieve fast learning. Empirically, this work showed that larger actor and critic learning rates, of the explored set, along with a well formulated reward function led to the convergence of an optimal policy within 15 minutes $\pm$ 2 of training. However, without a reward baseline, a smaller learning rate was required to avoid suboptimal BBMV. This behavior is seen in Fig. 4.3 where the training with R_1 does not learn the optimal policy and the convergence is slower compared to the reward equations with a baseline. DRL policy uses the Adam optimization algorithm, that employs per-parameter learning rates that adapt over time. However because of the limited training time, the initial actor and critic learning rates impact convergence time. From this simulator, a reward range and learning rates can be determined and deployed for human-in-the-loop experiments. The experience horizon is another important hyperparameter. Since the simulation is limited to 60 minutes, the agent requires more updates to learn within the alloted time. We observed that even with a horizon time of 20 seconds there was a decrease in the agent convergence because there are fewer time steps for the agent to explore the environment within a 60 minute therapy session. A shorter experience horizon does, however, make the agent more sensitive to variability.

The BBMV s during DRL agent training (Fig. 4.5), exhibit an initial decrease when the learning environment changes at 30 minutes. The BBMV then increases over time because the DRL agent adapts its policy to the different learning environment. Agents with smaller entropy values show less adaptation to PS_B compared to agents with larger entropy values. With larger entropy, when the predictive simulator changed, the agent was not as biased towards a particular action and was more likely to explore. For some participants, 1 and 2, selecting the same CT from PS_A in PS_B does not result in a large decrease in BBMV , see Fig. 4.5(a),(b). This is because the optimal CT from PS_A still achieves relatively high BBMV in PS_B. For others, participants 3 and 5, the BV response changes more from PS_A to PS_B and the agents adapt. Consequently, agents with larger entropy achieve greater BBMV across participants in this study.

The abrupt change in BBMV when the learning environment changes from PS_A to PS_B, seen in Fig. 4.5, is related to the change the output of the predictive simulator. For participants 2, 4, 5, and 6 there is a distinct decrease in BBMV at 30 minutes. Each

predictive simulator is trained on a separate day of data collection. Changes in sensor positioning, participant hydration, IPC pressure profile, participant stress levels, and heart rate from one day to another may impact the BBMV response. For each participant, heart rate was statistically different across days.

Our implemented DRL uses a feedforward policy NN and implicitly assumes the environment to be an MDP. However, the limited physiological signals available to the agent and the uncertainty associated with the signals from the measured ECG and the predicted BV response make the actual states of the environment only partially observable to the agent. This makes a POMDP [93], a better fit to our problem environment. While the DRL agent may still perform well in some POMDPs, where the agent cannot see the entire state, the DRL agent may be further improved with a long short-term memory network (LSTM) policy [47] instead of a feed-forward policy. An LSTM may help the agent learn relevant information across successive observable states that is not explicitly included in the state formulation. However, LSTM training can become unstable and requires further research. We also limited the action space to 19 actions, based on previous data collection and to promote fast learning. With an expanded discrete action space, or a continuous action space, a policy may emerge that further enhances BBMV.

Only time steps of 5 heartbeats were explored. We selected this range to isolate the effect of one CT to another while encouraging faster learning. Smaller time-steps may increase the agent's convergence rate, while also making the agent more susceptible to potential BBMV variability; larger time-steps may have the opposite impact. The lack of central hemodynamics and other physiological signals included in the predictive simulator makes this environment a POMDP. This may explain why the DRL agent selects the same action 98% of the time. This behavior may be attributed to the agent learning the optimal CT independent of the state. To encourage the formation of state-action mapping we used longer time steps, up to 5 heartbeats prior, as inputs for the agent in pilot simulations, but these agents did not converge to different CTs.

In a similar vein, the predictive simulator does not capture the full extent of BBMV variation over time. From previous testing of the simulator, the average BBMV R^2 is 0.83 across participants when tested for a one heartbeat predictions [101]. In this environment, we predict up to 3600 heartbeats (approximately 1 hour) ahead, while the $NARX_{bv}$ model was validated for only one heartbeat ahead predictions. Previous studies have observed variability in physiological signals related to BV over half hour intervals [79], thus, a decrease in accuracy may be expected. The simulated environment is limited to the accuracy of the BV model used in the environment and this model with fixed dynamics does not capture the full variability of a human participant. Despite this limitation, previous work has suggested that a surrogate model can be used to learn good hyperparameters [128]

which is important given the sensitivity of agents seen in previous work [52, 76]. While the goal of this paper was to show the feasibility of DRL in obtaining CT in participant specific simulations, the built predictive environment may also have implications in pre-tuning the hyperparameters of such DRL agent for the a human-in-the-loop optimization of CT. The validity of those hyperparameters will be tested in our future work with a focus on human-in-the-loop testing.

It is assumed that the predictive simulator used in this study can predict BV when the CT is changed every 5 heartbeats. However, this predictive simulator was trained and validated on data where the CT is changed every 3 minutes. The validation of this predictive model on different timing protocols is considered as one of the future steps.

The predictive simulator is trained using previously collected data and the positioning of the BV Doppler probe does impact the resultant learned policy of the DRL agent. However, since we are using this simulated environment to learn the optimal hyperparameters for a DRL agent [128], this decrease in accuracy is not critical.

In this work, we show that DRL algorithm can be used to optimize IPC timing for an individual in simulation. DRL is faster than previous methods and unlike previous methods, DRL can adapt to changes in the physiological system. DRL is a tool that may optimize IPC CT online, with a human-in-the-loop. The proposed DRL approach showed great promise in optimizing IPC timing for individuals. Unlike previous methods, DRL was shown to be efficient in finding the optimal CT and it can adapt to changes in the physiological system. These advantages suggest that our proposed DRL approach could be integrated into a controller for CT selection, where the agent is learning online during a therapy session. This ability to learn an optimal CT in real-time and adapt to changing physiology highlights the potential of our DRL agent to be used in human-in-the-loop scenarios in the future.

4.4 Conclusion

It was shown in this study that a DRL agent can optimize IPC CT in simulation in less than 19 minutes, which was obtained in previous work. The DRL agent's learning environment was 6 user-specific data-driven predictive simulators. Using a BV prediction model and an ECG signal, this agent can learn the CT to maximize blood flow per individual. Across participants the DRL agent achieved 98% $\pm$ 2 of the optimal blood flow. This DRL agent has the potential to be used online in a controller to optimally select the CT for each individual, which would lead to a significant increase in IPC performance.

Chapter 5

Optimizing Intermittent Pneumatic Compression Timing with a Human-in-the-Loop

In this chapter, we investigate if a DRL agent can optimize IPC CT with a human-in-the-loop within a 60-minute therapy session. Because the BBMV response to CT may change over time and across individuals, the optimal CT for an individual is unknown prior to data collection. It is hypothesized that a DRL agent can learn the optimal CT for an individual within one hour, based on the performance of the DRL agent in participant specific simulations, discussed in chapter 4. DRL performance is compared to performance when selecting the same CT at each heartbeat, fixed CT. Evaluation metrics include rewards received, actions selected, and BBMV over time.

5.1 Methods

5.1.1 Compression System

The CGC IPC device used in this study and the data collection protocols are described in chapter 3 (see section 3.1.1). The DRL agent was deployed in Simulink, MATLAB 2022a (MathWorks, US).

5.1.2 Participants

The study participants included 10 healthy volunteers (6 males, 4 females) of age 26 ± 4 years (mean ± SD). The average height and weight of the participants were 173 ± 11 cm and 73 ± 13 kg, respectively. Participant inclusion and exclusion criteria are summarized in chapter 3 (see section 3.1.1). Study protocols and procedures were approved by the University of Waterloo (ORE#44494) Clinical Research Ethics Committee and conformed with the Declaration of Helsinki.

5.1.3 Experimental Procedure

The objective of this work is to investigate if a DRL agent can optimize CGC IPC with a human-in-the-loop. To that end, the experimental procedure was designed to meet certain objectives (1) to compare the performance of a DRL agent selected CT and a constant CT at each heart beat (fixed CT) with a human-in-the-loop, and (2) to observe the DRL agent's learning within a 60-minute therapy session.

The experiment consisted of a 5-minute baseline prior to the application of compression, 5 minutes of CT in the local diastolic phase of the blood pulse, (CT_d), and then approximately 60 minutes of compression where the timing alternated CT modes every 20 heartbeats, see Fig. 5.1 (a). This period was not exactly 60 minutes because it was based on a set number of heartbeats, which vary from one individual to the next, instead of time. We applied 5-min of CT_d before beginning DRL training to avoid venous pooling influencing the DRL agent's learning. The alternating CT modes were (1) control, CT_d at each heartbeat, and (2) exposure, DRL agent timing, see Fig. 5.1 (a). During exposure the DRL agent could select either CT_s, or CT_d, see Fig. 5.1. (b). CT_d applied compression in late diastole and CT_s applied compression in the middle of systole, see Fig. 5.2. The phase lag between the ECG R-wave and the local blood pulse was recorded and fixed prior to data collection for each subject. CT was applied as a percentage of the previous R-R interval of the ECG signal.

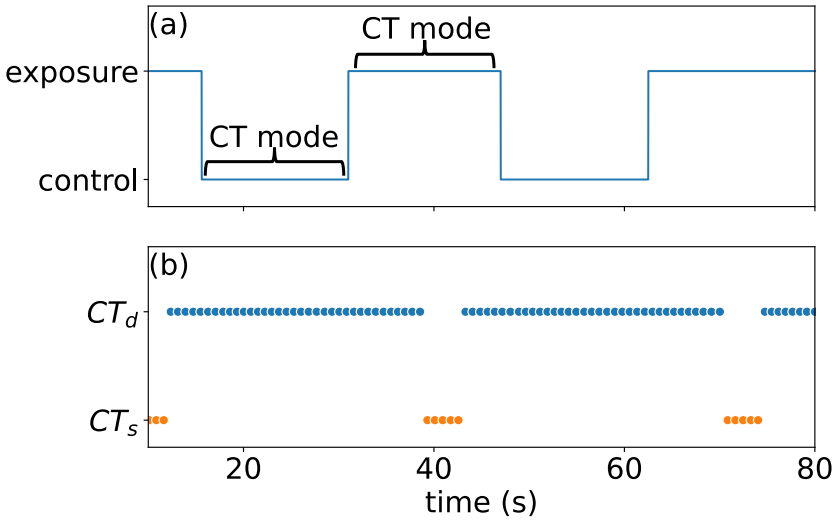


Figure 5.1: (a) Typical mode control signal that alternates every 20 heartbeats. (b) Action selected at each heartbeat.

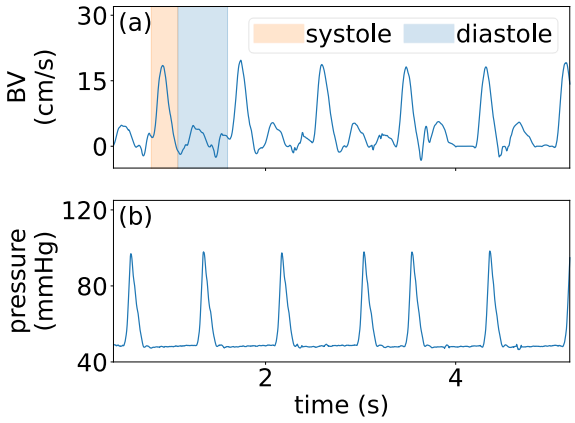


Figure 5.2: (a) Typical BV signal. Orange highlighted section shows systole and blue highlighted section shows diastole. (b) Sample pressure applied. First 3 heartbeats shows compression occurring in diastole, CT_d, and remaining heartbeats show compression occurring in systole, CT_s.

To meet the first objective, comparing the performance of the DRL agent's CT and fixed CT, the CT mode alternated during the experiment. Pilot studies showed that the BV response to a fixed CT varies with time, see Appendix (Fig. A.3). Consequently, we chose to alternate between DRL CT and control CT on a shorter timescale, 20 heartbeats, where each CT mode is applied for a set number of heartbeats to better compare the exposure and control performance. This timescale was chosen to avoid short-term physiological variability produced by respiration that appears in blocks of 5 to 10 heartbeats. As well, in pilot studies we observed that CT mode blocks of 30 heartbeats and greater are influenced by changes in BV response to CT over time. In Fig. 5.3 (d), 20 heartbeats has the smallest MAE between consecutive blocks at 0.04 cm/s $\pm$ 0.03 and 25 heartbeats, Fig. 5.3 (e), has the largest MAE at 0.14 cm/s $\pm$ 0.13. The MAE evaluates the difference between consecutive blocks, where the ideal MAE is 0, such that the BBMV does not change across consecutive blocks.

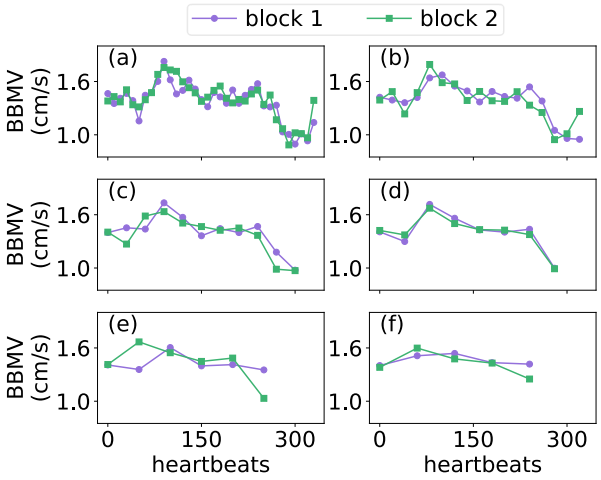


Figure 5.3: BBMV during 5 minutes IPC applied at CT_d at each heartbeat. The plot shows mean BBMV for blocks of different heartbeats, where block 1 and block 2 are consecutive blocks. (a) 5 heartbeats; (b) 10 heartbeats; (c) 15 heartbeats; (d) 20 heartbeats; (e) 25 heartbeats; (f) 30 heartbeats.

To meet the second objective, to observe learning from the DRL agent within a 60 minute therapy session, the DRL agent explored in chapter 4 was modified. The DRL agent discussed in this chapter has 2 actions: CT_d and CT_s while chapter 4's DRL agent had an action space of 19. Because the CT mode alternated between exposure and control timing, the DRL agent had only half the session to learn the optimal CT in this chapter versus the full session in simulation in chapter 4. Additionally, pilot testing with the 19-action DRL agent suggested that the the variability associated with the BV response to CT made the problem too challenging for the agent to optimize across 19 CT within 30 minutes. Part of this variability can be attributed to limitations of the experimental set-up where there is variability between the command CT and the applied CT. Consequently, the action space was reduced to encourage faster and consistent learning.

5.1.4 Reward Function

The effect of CT is best observed in the same heartbeat as when the compression was applied, see Fig. 5.2 (a). However, to minimize serial correlation from one agent time step to another, we remove the first heartbeat in an agent time step from the reward function calculations. The DRL agent's reward function is shown below:

$$R = BBMV_t + PSV_t - (0.75BBMV_c + 0.25BBMV_{t-1} + PSV_c) \tag{5.1}$$

where the subscript t refers to each agent time step and $BBMV_t$ is defined as the mean BBMV over the previous agent time step, after removing the first heartbeat. Peak systolic blood velocity PSV can be correlated with BBMV and changes with CT. The baselines are included in the equation are $BBMV_c$, $BBMV_{t-1}$, and PSV_c. The mean BBMV and PSV during the preceding control block are $BBMV_c$ and PSV_c, respectively. These terms were included as baselines for the agent. $BBMV_{t-1}$ is defined as the mean BBMV over the last four heartbeats in the previous time step and it was included so that the DRL agent was incentivized to continually increase BBMV. The inclusion of $BBMV_c$, $BBMV_{t-1}$, and PSV_c make the reward function centered around 0. Consequently, it is less sensitive to the variability of absolute BBMV response across participants. However if the BBMV is decreasing over time, independently of CT, the reward will decrease over time too. This trend is opposite if the BBMV increases over time.

5.1.5 Training

The DRL agent has the same state formulation as the agent described in chapter 4 (see Section 4.1.1). The DRL agent has an action space of 2, CT_s and CT_d. The DRL Policy and Value function use the same architecture outlined in Chapter 4 (see section 4.1.2). The DRL agent hyperparameters are summarized in Table 5.1. The DRL agent was trained for 1 episode which included 380 agent time steps, corresponding to approximately 60 minutes.

Table 5.1: Hyperparameters used for DRL agent.

	Hyperparameter
Actor learning rate	0.06
Critic learning rate	0.09
Entropy	0.01
Horizon	2 heartbeats
Batch size	2 heartbeats

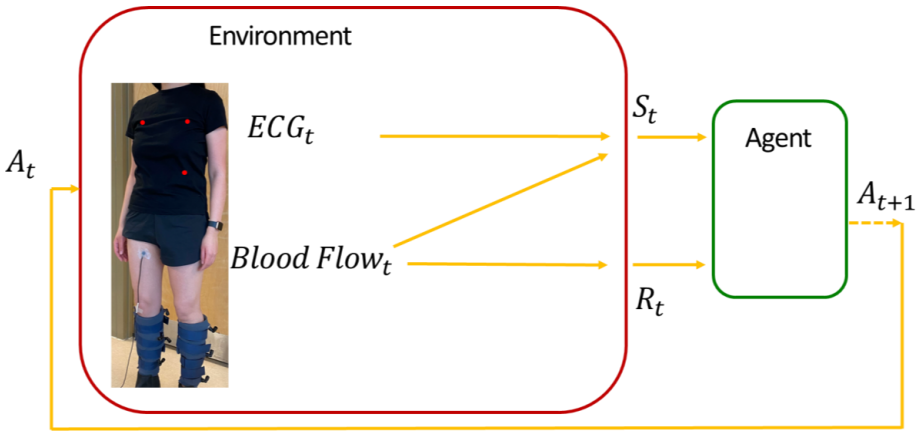


Figure 5.4: Agent-human interaction. The environment receives an action, A_t from the agent and applies compression with the desired CT. The measured BBMV and ECG signals compose the state and reward output, S_t and R_t, respectively. These values serve as inputs to the agent and the agent outputs the next action, A_{t+1}, to continue the cycle.

The agent-environment interaction loop is shown in Fig. 5.4. To complete an agent time step, the agent applies an action, A_t, which is a CT, for the length of 1 agent time step (5 heartbeats). The environment outputs the updated state, S_t, and reward, R_t. The agent receives the state and reward, to update its Policy and Value functions, and then outputs the next state, A_{t+1}.

5.1.6 Agent Evaluation

To evaluate DRL timing, the mean BBMV during exposure is compared to BBMV during control. This metric, Δ_{BBMV}, is calculated as shown below:

$$\Delta_{BBMV} = BBMV_{t,exposure} - BBMV_{t-1,control} \tag{5.2}$$

where $BBMV_{t,exposure}$ is the mean BBMV during a single exposure CT mode block and $BBMV_{t-1,control}$ is the mean BBMV during preceding control CT mode block. Rewards over time and action selection are also analyzed. Trends in rewards over time and Δ_{BBMV} over time are assessed by a Mann-Kendall (MK) test [80, 59] with the Hamed-Rao modification [45] to account for the serial correlation present in the data.

5.2 Results

The rewards over time for each participant, along with the moving average, are shown in Fig. 5.5. Based on the reward definition, equation 5.1, a positive reward indicates that there is a consistent improvement in BBMV. Participants 1, 2, 3, 4, and 8 had positive mean reward values in the last 10 minutes of the therapy session. Each participant has a different reward range. For example, participants 3 and 8 have larger reward ranges compared to the other participants as determined by the reward interquartile ranges, see Fig. 5.5 (c) and (f). Each participant has overlap in the reward values for the 2 CTs. This overlap makes this problem more difficult for the agent and suggests that CT selection is not the only factor that influences BBMV. The overlap between CT_s and CT_d, using interquartile ranges, is largest for participants 9 and 10. Participants 1, 2, 9, and 10 exhibit the slowest convergence to CT_d during training. While participants 1 and 2 have smaller reward ranges than participants 9 and 10, these participants had a larger difference in the distribution of reward values for the two actions, using interquartile ranges, compared to participants 9 and 10, which led to similar convergence rates overall. Participant 7 shows a decreasing trend in reward during training, this behavior is consistent with the decrease in BBMV during training.

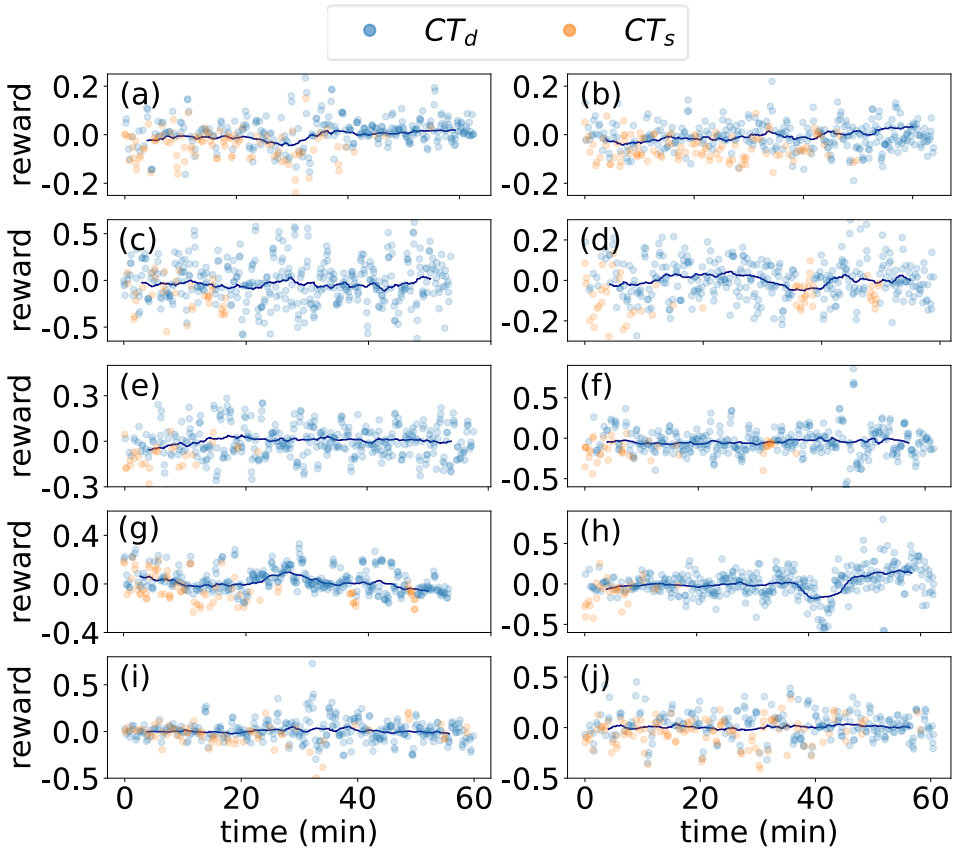


Figure 5.5: Rewards during DRL training for actions CT_d (diastole) and CT_s (systole). The moving average of the reward is shown in dark blue. (a) Participant 1; (b) Participant 2; (c) Participant 3; (d) Participant 4; (e) Participant 5; (f) Participant 6; (g) Participant 7; (h) Participant 8; (i) Participant 9; (j) Participant 10.

The actions selected during training are shown in Fig. 5.6. The moving average (5 minute window) is shown in dark blue. While the rate of convergence varies across participants, the plot shows that the agent converges to selecting CT_d the majority of the time for each participant. Participants 1, 2, 9, and 10 had relatively slower convergence compared to other participants. However, once these agents converge, their behavior is consistent. Alternatively, participants 4, 6, and 7 had faster convergence to CT_d, but later during training the agents selected CT_s more often, suggesting that slower learning can be more consistent than faster learning. CT_d was selected 83% $\pm$ 9 of the time during training, across participants.

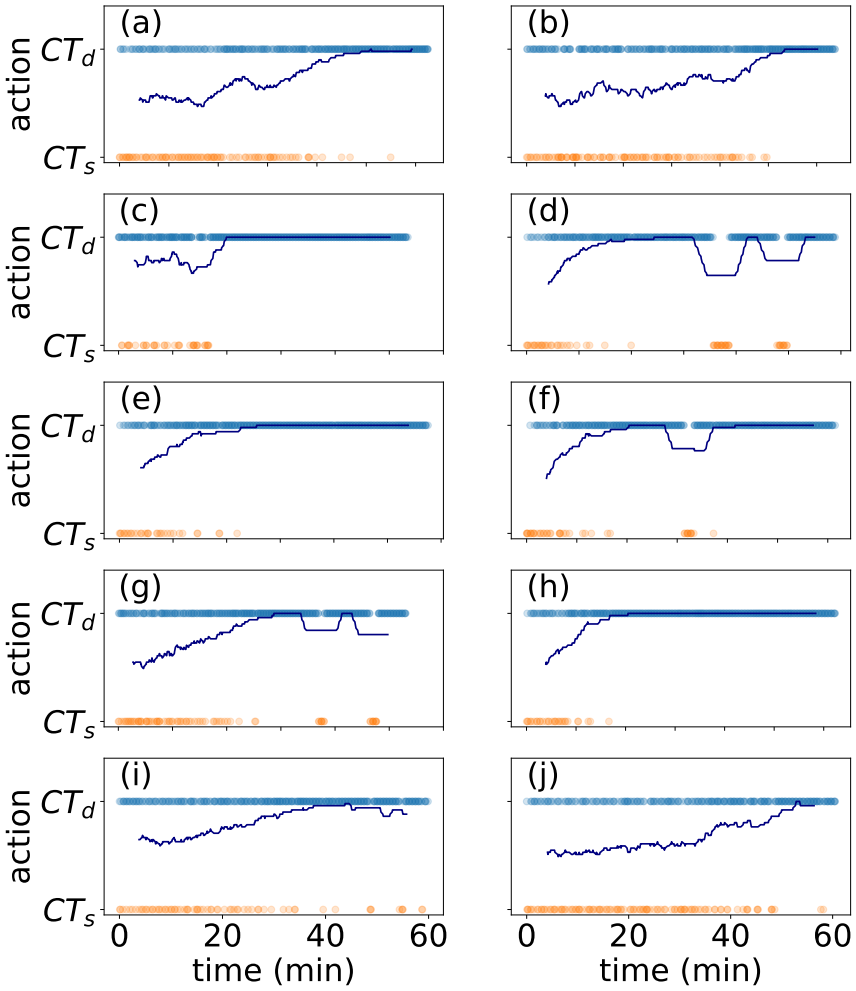


Figure 5.6: Action selection during training for DRL agents. (a) Participant 1; (b) Participant 2; (c) Participant 3; (d) Participant 4; (e) Participant 5; (f) Participant 6; (g) Participant 7; (h) Participant 8; (i) Participant 9; (j) Participant 10.

Mean BBMV and HR during the therapy session for 2 participants, along with the moving averages of both signals, is shown in Fig. 5.7. This plot shows that BBMV is noisy during a therapy session and that the BBMV response to CT is is not the same across participants, seen by the variability for these two participants. For participant 3 from 30 to 60 minutes, only CT_d was applied in this time interval, however, the BBMV ranges from 1.99 cm/s to 3.80 cm/s, a 91% increase. The decrease in BBMV for participant 7 from 30 to 60 minutes, see Fig. 5.7 (d) occurs while the CT selection was CT_d for the majority of the time, see. Fig. 5.6 (g). This continuous decrease in BBMV makes for a more challenging environment for the BBMV agent, because rewards decrease over time during training but this decrease cannot be attributed entirely to action selection. While this decrease in BBMV is accompanied by a decrease in HR, see Fig. 5.7 (b), BBMV is not only influenced by HR, it is also influenced by other factors, such as changes in blood pressure, which was not measured during data collection. Also, different participants can present different relationships between BBMV and HR. For example, participant 3 has a lower Pearson correlation coefficient between BBMV and HR compared to participant 7, 0.33 versus 0.57.

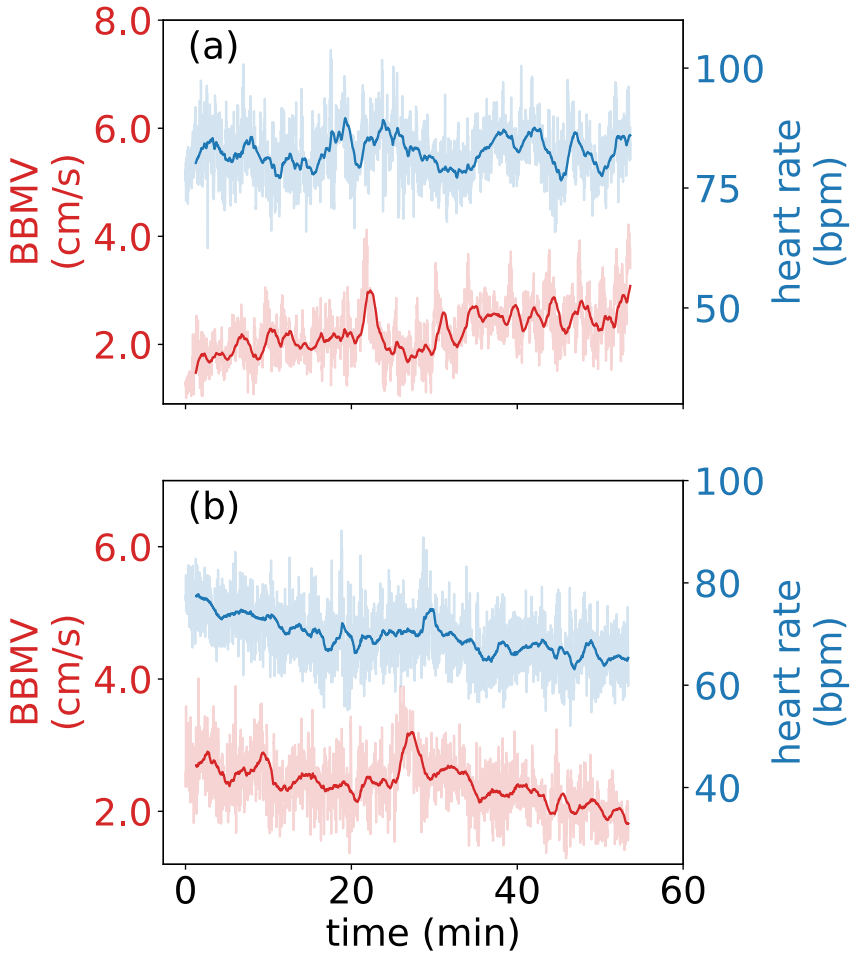


Figure 5.7: BBMV and HR over time during DRL training. The faded line shows the measured BBMV and HR and the solid dark line shows the moving average. (a) Participant 3; (b) Participant 7.

Δ_{BBMV} during training is shown in Fig. 5.8. Overlaid in red is the moving average of Δ_{BBMV}. The dashed line shows the mean Δ_{BBMV} during training. The blue shaded area shows the moving SD of the exposure timing blocks, centered about the moving average of Δ_{BBMV}. The increase in the moving average of Δ_{BBMV} is seen most prominently for participants 1 and 2; these participants also have the highest MK slope for the difference, see Table 5.2. The mean Δ_{BBMV} between the DRL CT and fixed CT was -0.09 cm/s $\pm$ 0.09 (-4% $\pm$ 4). Across participants the agents did not converge to an optimal sequence that included both CT_s and CT_d but instead to a policy that selected CT_d, the control CT, at most agent time steps. Consequently, a negative mean Δ_{BBMV} is expected and a mean Δ_{BBMV} of 0 would be the highest expected mean Δ_{BBMV}, given the agents' CT selection. We hypothesized that as the agent converged to CT_d, the SD of exposure would decrease. This trend was observed for most participants, though it was only statistically significant ($p< 0.01$) for the collected data of participants 1 and 6. The SD of BBMV with CT_d was larger than that of CT_s across participants, 5.3 cm/s $\pm$ 1.2 versus 6.1 cm/s $\pm$ 1.0, and this may explain why the decrease in SD was not significant across participants.

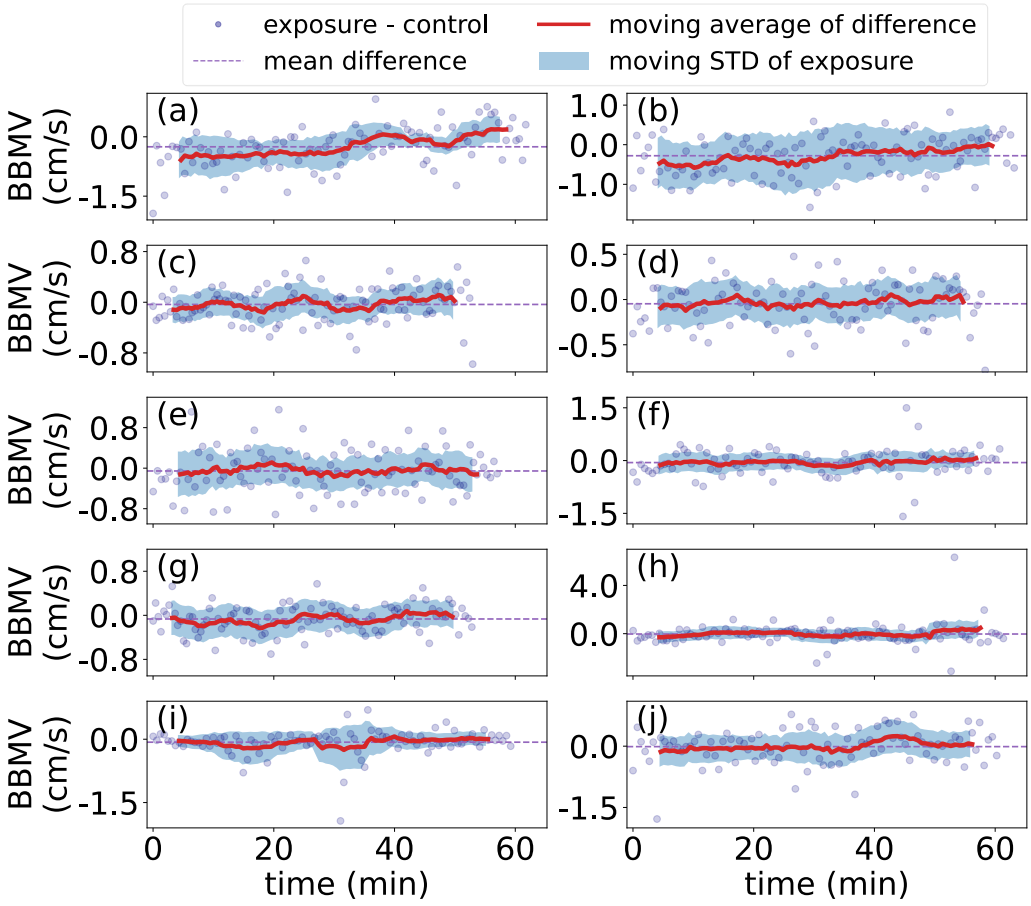


Figure 5.8: Difference between consecutive exposure and control mode blocks during DRL training (scatter points). Dashed line shows the mean difference during training. Solid line shows the moving average of the difference. The blue shaded region is the SD of the exposure blocks. (a) Participant 1; (b) Participant 2; (c) Participant 3; (d) Participant 4; (e) Participant 5; (f) Participant 6; (g) Participant 7; (h) Participant 8; (i) Participant 9; (j) Participant 10.

Table 5.2: MK test results for slope and BBMV difference. + indicates a positive slope and − indicates a negative slope. # denotes p<0.01

Participant	Reward slope	Δ_{BBMV} slope ($\frac{\frac{cm}{s}}{heartbeat}$)	Exposure standard deviation ($\frac{\frac{cm}{s}}{heartbeat}$)
1	+	2.05×10^{-4}#	-1.41×10^{-4}#
2	+ #	1.35×10^{-4}#	-8.43×10^{-5}
3	+	4.10×10^{-5}	2.78×10^{-5}
4	-	2.47×10^{-5}	-8.90×10^{-6}
5	+	2.11×10^{-5}	-5.92×10^{-6}
6	+	5.34×10^{-5}	-1.25×10^{-5}#
7	-	2.85×10^{-5}	-4.44×10^{-5}
8	+	1.84×10^{-5}	4.77×10^{-5}
9	-	1.74×10^{-5}	-1.66×10^{-5}
10	+	4.49×10^{-5}	1.07×10^{-5}

Table 5.2 summarizes the MK test results for each participant, for the rewards, Δ_{BBMV} and the SD of the exposure blocks. Since the reward value is without units, only the direction of the slope is presented. The slope is positive for 7 out of 10 participants and statistically significant for 2 of those participants. The negative reward slopes are likely because these participants selected CT_s later during training. Across participants the slope of the BBMV difference during training is positive. However, the increasing trend is only statistically significant for participants 1 and 2. This may be because participant 2 had the largest difference in BBMV between CT_d and CT_s. The moving average of the SD of the exposure blocks decreases, as determined by the MK test, for each participant except for participants 3, 8, and 10.

5.3 Discussion

In this work, we investigate if a DRL agent can optimize CT with a human-in-the-loop. This DRL method is not limited to fixed dynamics and is successful across different participants. This agent can optimize IPC CT during a therapy session to maximize blood velocity when the optimal CT is not known beforehand.

Previous work that optimizes IPC CT with a NARX model requires 19 [63] or 41 minutes (Chapter 3) of data collection. Limitations of these methods include (1) the

accuracy is limited to the next heartbeat, (2) these methods are limited to fixed dynamics, and (3) they have not been tested with a human-in-the-loop. Conversely, the DRL agent can optimize CT for each heartbeat, it can adapt to changes in physiology and, as shown here, has been shown to maximize BBMV across 2 CTs with a human-in-the-loop. The NARX models presented in chapter 4 and in Landry et al. [63] select the optimal CT across 19 and 15 CTs, respectively, whereas the agent presented in this chapter learns the optimal timing with 2 CTs. This agent converges to the optimal CT in 29 minutes ± 14. The agent's convergence time is defined as when it selects CT_d 98% of the time in the previous 5-minute window of training. This DRL agent has the potential to be expanded to optimize across a larger number of CTs and can even optimize across a continuous range of CTs from 0% to 100% with future research.

The DRL agent can interact with a human to learn the optimal CT strategy for an individual in real-time during a therapy session. The DRL agent can learn from a randomly initialized policy and does not require pre-training in simulation. We employ PPO, which is an off-the-shelf DRL algorithm that does not require additional modifications. The state and reward computations are simple to implement to an existing CGC system. Similar to enhanced external counterpulsation therapy [95], an individual can go to a clinic a few times a week to receive a 1 hour therapy session that optimizes BV.

Despite the DRL agent for each participant having the same set of hyperparameters, the variability across participants led to different behavior during the therapy session, see Fig. 5.6. For example, variability in the range of rewards across participants, for example see Fig. 5.5.(a) versus (c). This variability is likely due to changes in BBMV over time and the variability in the BBMV difference between CT_d and CT_s across participants. Another source of variability is the different states observed by the agent during training. The actor and critic learning rates, clip factor, and range of rewards impact the magnitude of the Policy and Value function updates which in turn influences the agent's convergence rate. Mahmood et al. [76] show that there is a significant spread in PPO performance for different DRL agent hyperparameters performing the same robotic task. Consequently, it is expected that for these different reward ranges and values, these agents will have different policy convergence rates across participants, for example see Fig. 5.6 (a) and (c). While hyperparameters like actor and critic learning rates and the clip factor can be decided based on simulation results, the reward range can vary more for human-in-the-loop testing and it is a source of variability.

The behavior observed for some participants in this study did not match that seen in simulation with a 19 CT agent. Some participants selected CT_s after many time steps of selecting CT_d, see Fig. 5.6 (d), (f), and (g). Whereas in simulation a 19 CT action agent converged to a single CT and did not select other CTs for more than 1 consecutive

heartbeat, see chapter 4. Factors such as state variability and reward variability can promote agent exploration during training. However, in this case the exploration does not last until the end of the therapy session, because the reward function baselines encourage the agent to return to CT_d. This exploration may also occur when the actor and critic learning rates are large and the variability in the rewards is high. In this case, large positive rewards increase the probability of the agent selecting a particular action, CT_s in this case, and then subsequent large negative rewards can revert the agent to its previous behavior where it selected CT_d. To avoid such large changes to the policy during training, PPO includes a clip factor in the policy update. However the clip factor must be tuned with respect to the reward range and the learning rate. Consequently, with human-in-the-loop experiments, it is difficult to know the best values prior to a therapy session. There is a balance between a large enough learning rate to observe learning within a typical therapy session and a small learning rate that produces more stable behavior, see participants 4, 6, and 7 versus 1 and 2, Fig. 5.6 (d), (f) and (g) versus (a) and (b).

In a typical DRL problem, the reward for an RL agent increases over time and then plateaus, as shown in the DRL agent simulation results (see chapter 4 section 4.2.1 Fig. 4.3). However, the rewards in this experiment do not exhibit such behavior, see Fig. 5.5. This behavior may be because the variability in BBMV can be larger than the difference in reward for the two CTs and the baseline terms, $BBMV_c$ and $BBMV_{t-1}$, create a reward function that is scaled to the difference between the two actions, rather than being scaled to the absolute BBMV measurement. However, this reward function also facilitated the successful use of the same hyperparameters across participants because of these baselines. Some participants may have negative reward slopes, see Table 5.2, because they selected CT_s later during training which decreased rewards. However, these participants still show learning through their action selection, see Fig. 5.6 (d), (f), and (g).

The BBMV and HR variability during a therapy session highlights the difficult environment that the DRL agent is tasked with, see Fig. 5.7. This variability is expected because of the variability of human physiology [79]. Pilot studies suggest that changes in BBMV in response to CT cannot be entirely attributed to the CT timing itself, but instead are the result of an unrelated physiological effect. For this reason, we include the SD of HR and BBMV in the state formulation so that the DRL agent can learn that when it is in states with high variability action selection has a smaller impact on resultant BBMV. High variability is relative across participants, however a SD that is above 0.5 cm/s over a 20-heartbeat CT mode block is high. This value is double the average SD from 20 sessions of 30 minutes of constant CT.

While physiological variability can impact DRL agent's behavior, DRL is well-suited to tackle this variability. While participant 7's BBMV was decreasing over time, their HR

was decreasing too, see Fig. 5.7 (b). This decrease in BBMV made it more likely for the agent to return to CT_s, because CT_d returned negative rewards due to the reward function's baselines. However, as shown by the plot, Fig. 5.5 (g), CT_s gave rewards that were generally more negative than CT_d and consequently the agent returned to CT_d selection. The DRL agent's adaptability does produce more exploration which can have a small decrease in BBMV, 0.2 to 0.5 cm/s, during training. Over the course of a therapy session, these small periods of exploration do not have a large impact on BBMV as shown by the small difference between the mean BBMV of the control and exposure periods, see Fig. 5.7 (b).

5.3.1 Limitations and Future Work

It is assumed, in this protocol, that there are no carry-over effects from one 20-heartbeat exposure mode to the next 20-heartbeat control mode and vice versa when calculating BBMV. While the BV waveform shape changes within the heartbeat that the CT is applied and changes in the waveform shape do not last beyond that heartbeat, it is unknown if a given CT produces longer effects on BBMV. The first heartbeat in an agent timestep is dropped from the reward calculation to avoid the possibility of any carry-over effects, here it is assumed that this impact will last only one heartbeat. More research to determine the time scale of the effect of a given CT is required.

Only young and healthy participants were included in this study. Due to time limitations associated with human-in-the-loop experiments, only this state space was explored and other state formulations may improve policy convergence across participants. Other state formulations will be discussed later in this subsection.

Another limitation is that this DRL agent is tasked with optimizing CT across only two actions whereas in the previous chapter the agent optimized CT across 19 actions. In addition, based on previous work and physiology, it is known that one of the actions, CT_s, is a suboptimal CT and in pilot studies CT_s has even produced negative blood flow (not observed in this study). The other action, CT_d has been shown to enhance BBMV [26, 140, 141, 139] compared to no compression. Consequently, the agent had to optimize between an optimal and suboptimal CT which is an easier task than the 19 CT problem explored in chapter 4. Future work will focus on expanding the action space of the agent to further optimize CT. Expanding the action space is a non-trivial task because: (1) the BBMV response to CT varies across individuals and over time; and (2) Bellman's curse of dimensionality.

Some participants are more responsive to CT than others which may make it difficult for the agent to converge to a policy, that optimizes IPC across more actions, for each

participant since the reward signal will be noisy. In these cases, a 60 minute therapy session will not include enough agent time steps to optimize IPC and either more sessions will be required to train the agent or a pre-trained policy will need to be used. Learning a policy over multiple sessions introduces errors related to sensor placement and participant variability over different days. We suggest that the agent only be deployed over multiple days with a human-in-the-loop if the simulations suggest that the agent can learn a mapping of physiological states to actions that is valid for multiple days since previous work suggests that the BBMV response to CT varies over time. Using a pre-trained policy requires a high fidelity simulator because research suggests that small errors in models can result in poor policies [50]. Consequently, the predictive simulator introduced in chapter 4 could to be expanded to include more physiological signals like cardiac output, blood pressure, and oxygenated hemoglobin concentration that have shown a response to IPC [26]. Also, these signals may give more insight into the mechanism or mechanisms responsible for the changing BV response over time. This improved simulator may help the agent learn a mapping, of physiological states to action, in simulation which can be can applied to a human-in-the-loop.

Bellman's curse of dimensionality is another challenge associated with expanding the action space. As the number of features, states and actions, increases, the complexity of the problem increases too. The Policy and Value functions become larger, with more parameters to learn. Consequently, more agent time steps are required for policy convergence. The agent's task of optimizing IPC CT with the ECG and BV signals is a POMDP as discussed in chapter 4 (section 4.3). The DRL agent explored in this study is limited to two actions and a state of size four to keep the ANN parameters low. Bellman's curse of dimensionality would apply if the number of actions for the DRL agent increases and if we chose to explore a larger state space to better learn a Policy that spans more actions. However, a larger state space may also give the agent more information about the environment would which would facilitate learning. These two points should be explored further in simulation, where the cost of each trial is low. Hierarchical RL [48] can help solve this challenging problem of a POMDP by solving sub tasks. For example, one subtask could be compressing in CT_d or CT_s where the small state size used in this work could be employed. Another subtask would being smaller CT increments within diastole, such as early CT_d or late CT_d where a larger state size could be used to better distinguish between these actions. Previous work has shown the viability of hierarchical DRL and POMDP for Atari games [66] and simulated robots [118] and hierarchical DRL for autonomous driving [94].

Sim-to-real reinforcement learning is a difficult task and one of current research. Previous work has suggested different approaches for tackling this difficult problem. Some groups have found success by pre-training the DRL agent in simulation [30, 70]. However,

in simulation, we did not observe a strong relationship between physiological state and action selection. Consequently, a pre-trained DRL Policy would likely only bias the agent towards a particular action rather than giving the agent insight between a physiological state and the optimal CT that could be further explored with a human-in-the-loop. Other research suggests that using a surrogate model to learn the optimal hyperparameters can be successful as well [43], which is the approach successfully employed in this work.

5.4 Conclusion

This work demonstrated that a DRL agent can optimize IPC CT in a 60-minute therapy session for 10 participants. Each participant converged to CT_d action selection, selecting CT_d 83 $\pm$ 9% of the time during training. The mean BBMV difference between the DRL agent's CT and fixed CT 0.09 cm/s $\pm$ 0.09 (4 % $\pm$ 4). This DRL agent is the first of its kind to optimize IPC CT with a human-in-the-loop. Future steps will focus on expanding the action space for the DRL agent. This DRL agent can be used with a human-in-the-loop to maximize BV.

Chapter 6

Conclusions and Recommendations

6.1 Conclusions

In this thesis, a method to optimize intermittent pneumatic compression (IPC) compression timing (CT) was explored with a human-in-the-loop, an approach to maximize blood flow by selecting the optimal CT for an individual, which is unknown prior to the compression session. This approach requires a deep reinforcement learning (DRL) agent to select CT during the therapy session and select CT and observe the resulting physiological state to learn the optimal CT. Previous methods to optimize IPC are limited to one heartbeat ahead predictions and employ fixed dynamics. However, the blood velocity (BV) response to CT changes over time.

In this thesis, gaps in understanding optimal CT were studied to achieve the three objectives below.

1. Estimate BV response to CT over two days.
2. Evaluate a DRL agent's ability to optimize IPC CT with simulation.
3. Optimize IPC CT with a human-in-the-loop.

The main findings, as they relate to these objectives, are summarized as follows.

In Chapter 3, we applied random CT to 6 individuals over 2 days. A nonlinear autoregressive network with exogenous inputs (NARX) to estimate the BV response to CT was trained on one day of data and then deployed on the next day, referred to as $NARX_{bv}$.

Though the BV response to IPC CT for an individual changed over time and across days, the response can still be predicted in a one heartbeat ahead fashion from the electrocardiography (ECG) and applied pressure signals with this pre-trained $NARX_{bv}$ model. We trained this $NARX_{bv}$ model on one day of data and predicted the BV response on the second day (objective 1). This $NARX_{bv}$ model is the first to estimate the BV response to CT across two days of data.

In Chapter 4, we built a DRL agent that interacts with a data-driven lower limb environment that uses a $NARX_{bv}$ estimation model for 6 participants over the course of a 60-minute therapy session. We show that with a well-formulated reward function and the optimal hyperparameters, the DRL agent can learn the optimal CT within 29 $\pm$ 14 minutes across participants (objective 2). To mimic the real-life conditions, we changed the physiological system of the DRL agent and observed that the agent can adapt its behavior to the changed environment across participants. The DRL agent converges to the optimal CT faster than previous methods and the agent can adapt to changes in its learning environment.

In Chapter 5, we deploy a DRL agent to optimize IPC CT with a human-in-the-loop. The DRL agent interacts with the human-in-the-loop to observe its physiological state and optimize CT across two actions, compression in diastole (CT_d) and compression in systole (CT_s). The DRL agent converges to CT_d for each participant in 29 minutes $\pm$ 14 (objective 3). We observe that despite variability in BV response across participants, CT_d is selected the majority of the time during training. We compare the performance of the DRL agent to fixed timing, CT_d.

6.2 Limitations and Recommendations

A source of uncertainty throughout this thesis is the measurement of BV using an ultrasound probe that is held in place with medical tape. With soft tissue ultrasound measurements, the signal is sensitive to changes in the insonation angle from testing over days. Additionally, the compression causes the leg to shake which produces noise in the measured BV signal. Research has focused on a self-contained Doppler ultrasound unit that is designed for continuous, long-term ultrasound imaging [127] which may improve measurements if incorporated in future work. This sensor has been tested under different frequencies of vibration (0 to 3.5 Hz) with an amplitude of 2 cm and the resultant motion artefact, in the imaging, is minimal.

The predictive simulator, $NARX_{bv}$, has a number of limitations which bias the DRL

agent during DRL training in simulation. The $NARX_{bv}$ model can be improved by including sensors related to central hemodynamics. For example, a blood volume clamp can be used to measure mean arterial pressure on the finger and this signal is related to blood flow [23]. However, this is an expensive sensor. Alternatively, a PPG sensor can be attached to a participants finger, toe, or wrist to monitor blood flow [135]. Studies should evaluate which location is best suited as a surrogate for SFA BV and least susceptible to motion from the IPC device. These sensors can give the model more information about the individual's lower limb hemodynamics to better estimate future BV. Additionally, the PPG sensor is easier to use in comparison to the Doppler ultrasound probe. We assume that $NARX_{bv}$ can estimate BBMV response to CT up to 3600 heartbeats ahead, which corresponds to approximately 60 minutes of an ECG signal, when used as the predictive simulator. This assumption is necessary because the simulator does not have access to the true BV, such that the BV response to IPC must be estimated many heartbeats ahead to simulate a 60-minute therapy session. However, $NARX_{bv}$ has been validated for only one heartbeat ahead predictions. Since the BV response to CT changes over time and BV is not constant over time due to fluctuations in blood pressure [79], it is likely that $NARX_{bv}$ performance decreases over successive heartbeats. More sensors may also improve the timescale of the $NARX_{bv}$ predictions, which will improve the simulator. Another assumption is that the $NARX_{bv}$ can be used to estimate the BV response to CT that changes every five heartbeats. However, the data collection protocol changes the CT every three minutes. Changing the protocol of data collection to align with the intended purpose of changing the CT every 5 heartbeats in the simulator may benefit future researchers. Additionally, including more sensors, such as the previous mentioned blood volume clamp, would allow for a richer DRL agent state. This may result in an agent that learns a state-action mapping. If such a mapping is learned, future work should focus on DRL explainability techniques, such as decision trees [42] or local interpretable model-agnostic explanations (LIME) [99].

Bridging the gap between the simulated environment and human participants would improve the DRL simulator. Introducing more stochasticity to the simulator could help bridge this gap, For example, adding stochasticity to the actions to mimic the variability of the cardiac gated compression (CGC) IPC device, or including a perturbance to the BV signal to mimic the BV variability measured in humans.

The human-in-the-loop DRL agent optimizes CT across 2 actions. Future work should focus on expanding the DRL agent such that it optimizes CT across more actions. To achieve this goal, a number of different avenues exist. One is to build a higher fidelity simulator, which can be achieved by taking the steps discussed in the previous paragraph, that may allow the possibility to transfer a pretrained policy from simulation and deploy it with a human-in-the-loop. The inclusion of more sensors discussed in the previous

paragraph can also be used to enrich the DRL agent's state. Alternatively, or in addition to, this DRL agent may need to be trained over multiple days because more interactions with the environment may be required to learn the optimal CT with a larger state space due to the curse of dimensionality. A second options is to augment or modify the DRL agent to optimize across more actions. Because this is environment is partially observable Markov decision process (POMDP), a long short-term memory network (LSTM) policy may help the agent learn long-term relationships between state and action. This LSTM policy can be trained with proximal policy optimization PPO, or a different algorithm such as SAC [137] or twin delayed deep deterministic policy gradient (TD3) [29], which have shown success with sim-to-real problems.

References

[1] J. Achiam. Part 2: Kinds of rl algorithms - spinning up documentation.

[2] adh30 revised work by DanielChangMD who revised original work of DestinyQx; Redrawn as SVG by xavax. Wiggers diagram, July 2016.

[3] E. Agliari, F. Alemanno, A. Barra, Orazi Antonio Barra, Alberto Fachechi, Lorenzo Franceschi Vento, and Luciano Moretti. Analysis of temporal correlation in heart rate variability through maximum entropy principle in a minimal pairwise glassy model. *Scientific Reports*, 10(1):15353, 2020.

[4] S. Amiri, K. Chandan, and S. Zhang. Reasoning with scene graphs for robot planning under partial observability. *IEEE Robotics and Automation Letters*, 7(2):5560–5567, 2022.

[5] C. Angermueller, D. Dohan, D. Belanger, R. Deshpande, K. Murphy, and L. Colwell. Model-based reinforcement learning for biological sequence design. In *International conference on learning representations*, 2019.

[6] A. G Barto and S. Mahadevan. Recent Advances in Hierarchical Reinforcement Learning. *Discrete Event Dyn. S.*, 13(1):41–77, 2003.

[7] Marc G. Bellemare, Will Dabney, and Rémi Munos. A distributional perspective on reinforcement learning, 2017.

[8] M.G. Bellemare, S. Candido, P.S. Castro, J. Gong, M.C. Machado, S. Moitra, S. S. Ponda, and Z. Wang. Autonomous navigation of stratospheric balloons using reinforcement learning. *Nature*, 588(7836):77–82, 2020.

[9] J. Book, C. N Prince, R. Villar, R. L Hughson, and S. D Peterson. Investigating the impact of passive external lower limb compression on central and peripheral hemodynamics during exercise. *Eur. J. Appl. Physiol.*, 116(4):717–727, apr 2016.

[10] G.E.P. Box and G.M. Jenkins. *Time Series Analysis: Forecasting and Control*. Holden-Day, 1976.

[11] E. Burns and R. Buttner. The st segment, 3 2022.

[12] Darren P Casey and Emma C Hart. Cardiovascular function in humans during exercise: role of the muscle pump. *The Journal of physiology*, 586(Pt 21):5045, 2008.

[13] S. L. Chung and H. D. Alun. Graduated compression stockings. *Can. Med. Assoc. J.*, 186(10):E391–E398, 2014.

[14] OpenStax College. Blood flow through the heart, December 2013.

[15] OpenStax College. Skeletal muscle vein pump, December 2013.

[16] OpenStax College. Systemic blood pressure, December 2013.

[17] A. J. Comerota. Intermittent pneumatic compression: Physiologic and clinical basis to improve management of venous leg ulcers. *J. Vasc. Surg.*, 53(4):1121–1129, 2011.

[18] J. Cusidó, J. Comalrena, H. Alavi, and L. Llunas. Predicting hospital admissions to reduce crowding in the emergency departments. *Appl. Sci.*, 12(21), 2022.

[19] M Dennis, PA Sandercock, J Reid, C Graham, G Murray, G Venables, A Rudd, and G. Bowler. Effectiveness of thigh-length graduated compression stockings to reduce the risk of deep vein thrombosis after stroke (clots trial 1): a multicentre, randomised controlled trial. *Lancet*, 373(9679):1958–1965, 2009.

[20] P. Duarte-Mendes, R. Paulo, P. Coelho, F. Rodrigues, V. Marques, and S. Mateus. Variability of Lower Limb Artery Systolic-Diastolic Velocities in Futsal Athletes and Non-Athletes: Evaluation by Arterial Doppler Ultrasound. *Int. J. Environ. Res. Public Hea*, 17(2), jan 2020.

[21] J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *J. Mach. Learn. Res.*, 12(null):2121–2159, jul 2011.

[22] R. Eisele, L. Kinzl, and T. Koelsch. Rapid-Inflation Intermittent Pneumatic Compression for Prevention of Deep Venous Thrombosis. *J. Bone Jt. Surg.*, 89(5), 2007.

[23] Laura M Ellwein, Hien T Tran, Cheryl Zapata, Vera Novak, and Mette S Olufsen. Sensitivity analysis and model assessment: mathematical models for arterial blood flow and blood pressure. *Cardiovascular Engineering*, 8:94–108, 2008.

[24] P. Escandell-Montero, M. Chermisi, J.M. Martínez-Martínez, J. Gómez-Sanchis, Barbieri C., E. Soria-Olivas, F. Mari, J. Vila-Francés, A. Stopper, E. Gatti, and J.D. Martín-Guerrero. Optimization of anemia treatment in hemodialysis patients via reinforcement learning. *Artif. Intell. Med.*, 62(1):47–60, 2014.

[25] A R Eze, A J Comerota, P L Cisek, B S Holland, R P Kerr, R Veeramasuneni, and A J Jr Comerota. Intermittent calf and foot compression increases lower extremity blood flow. *Am. J. Surg.*, 172(2):130–4; discussion 135, aug 1996.

[26] A. Fawcett. *The Influence of Timing on the Hemodynamic Effects of Compression Devices and Development of Sensor Driven Timing Mechanisms*. Master's thesis, University of Waterloo, 2018.

[27] C. Ferraresi, D. Maffiodo, and H. Hajimirzaalian. A model-based method for the design of intermittent pneumatic compression systems acting on humans. *Proc. Inst. Mech. Eng. H.*, 228(2):118–126, feb 2014.

[28] D. L. Franklin, W. Schlegel, and R. F. Rushmer. Blood flow measured by doppler frequency shift of back-scattered ultrasound. *Science*, 134(3478):564–565, 1961.

[29] Y. Gao, J. Tebbe, and A. Zell. A model-free approach to stroke learning for robotic table tennis. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2022.

[30] Y. Gao, J. Tebbe, and A. Zell. Optimal stroke learning with policy gradient approach for robotic table tennis. *Appl. Intell.*, 53(11):13309–13322, 2023.

[31] S. Gönül, T. Namli, A. Coşar, and İ. Toroslu. A reinforcement learning based algorithm for personalization of digital, just-in-time, adaptive interventions. *Artif Intell Med*, 115:102062, 2021.

[32] S. Gelman, D. S. Warner, and M. A. Warner. Venous Function and Central Venous Pressure: A Physiologic Story. *Anesthesiology*, 108(4):735–748, 04 2008.

[33] Dhanjoo N Ghista and Foad Kabinejadian. Coronary artery bypass grafting hemodynamics and anastomosis design: a biomedical engineering review. *Biomedical engineering online*, 12:1–28, 2013.

[34] T. D. Gibbons, K. A. Zuj, C. N. Prince, D. C. Kingston, S. D. Peterson, and R. L. Hughson. Haemodynamic and cerebrovascular effects of intermittent lower-leg compression as countermeasure to orthostatic stress. *Exp. Physiol.*, 104(12):1790–1800, 2019.

[35] R. W. Gill. Measurement of blood flow by ultrasound: accuracy and sources of error. *Ultrasound Med. Biol.*, 11(4):625–641, 1985.

[36] E. Ginzburg, S. M. Cohn, J. Lopez, J. Jackowski, M. Brown, and S. M. Hameed. Randomized clinical trial of intermittent pneumatic compression and low molecular weight heparin in trauma. *Br. J. Surg.*, 90(11):1338–1344, nov 2003.

[37] D.S. Goldstein, O. Bentho, M.Y. Park, and Y. Sharabi. Low-frequency power of heart rate variability is not a measure of cardiac sympathetic tone but may be a measure of modulation of cardiac autonomic outflows by baroreflexes. *Exp. Physiol.*, 96(12):1255–1261, 2011.

[38] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

[39] Alex Graves. *Neural Networks*, pages 15–35. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[40] M. Griffin, S. K. Kakkos, G. Geroulakos, and A. N. Nicolaides. Comparison of three intermittent pneumatic compression systems in patients with varicose veins: a hemodynamic study. *Int. Angiol.*, 26(2):158–164, jun 2007.

[41] S. D. Grosse Grosse, R. E. Nelson, K. A. Nyarko, L. C. Richardson Richardson, and G. E. Raskob. The economic burden of incident venous thromboembolism in the united states: A review of estimated attributable healthcare costs. *Thromb. Res.*, 137:3–10, 2016.

[42] W. Guo and P. Wei. Explainable deep reinforcement learning for aircraft separation assurance. In *2022 IEEE/AIAA 41st Digital Avionics Systems Conference (DASC)*, pages 1–10, 2022.

[43] T. Haarnoja, S. Ha, A. Zhou, J. Tan, G. Tucker, and S. Levine. Learning to walk via deep reinforcement learning, 2019.

[44] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor, 2018.

[45] K.H. Hamed and A. Ramachandra Rao. A modified mann-kendall trend test for autocorrelated data. *J. Hydrol.*, 204(1):182–196, 1998.

[46] A.Y. Hannun, P. Rajpurkar, M. Haghpanahi, G.H. Tison, C. Bourn, M.P. Turakhia, and A.Y. Ng. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nat. Med.*, 25(1):65–69, 2019.

[47] M. Hausknecht and P. Stone. Deep recurrent q-learning for partially observable mdps. In *2015 AAAI Fall Symposium Series*, 2015.

[48] B. Hengst. *Hierarchical Reinforcement Learning*, pages 495–502. Springer US, Boston, MA, 2010.

[49] D. Ho, K. Rao, Z. Xu, E. Jang, M. Khansari, and Y. Bai. Retinagan: An object-aware approach to sim-to-real transfer. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 10920–10926, 2021.

[50] S. Höfer, K. Bekris, A Handa, J.C. Gamboa, M. Mozifian, F. Golemo, C. Atkeson, D. Fox, K. Goldberg, J. Leonard, C. Liu, J. Peters, S. Song, P. Welinder, and M. White. Sim2real in robotics and automation: Applications and challenges. *IEEE Trans. Autom. Sci. Eng.*, 18(2):398–400, 2021.

[51] P. R. Hoskins. Measurement of arterial blood flow by doppler ultrasound. *Clin. phys. physiol. meas.*, 11(1):1, 1990.

[52] J. Ibarz, J. Tan, C. Finn, M. Kalakrishnan, P. Pastor, and S. Levine. How to train your robot with deep reinforcement learning: Lessons we have learned. *Int. J. Robot. Res.*, 40(4-5):698–721, 2021.

[53] P. Jonas, S. Charlois, M. Chevalerias, D.E. Delmas, J. Kerihuel, and P. Blanchemaison. Efficacy of the Stendo Pulsating suit in patients with leg lymphedema: a pilot randomized study. *Eur. J. Dermatol.*, 26(1):82–89, 2016.

[54] L.P. Kaelbling, M.L. Littman, and A.R. Cassandra. Planning and acting in partially observable stochastic domains. *Artif Intell*, 101(1):99–134, 1998.

[55] S. K. Kakkos, A. N. Nicolaides, M. Griffin, and G. Geroulakos. Comparison of two intermittent pneumatic compression systems. A hemodynamic study. *Int. Angiol.*, 24(4):330–335, dec 2005.

[56] S. K. Kakkos, G. Szendro, M. Griffin, S. S. Daskalopoulou, and A. N. Nicolaides. The efficacy of the new SCD response compression system in the prevention of venous stasis. *J. Vasc. Surg.*, 32(5):932–940, 2000.

[57] C. Kasai, K. Namekawa, A. Koyano, and R. Omoto. Real-time two-dimensional blood flow imaging using an autocorrelation technique. *IEEE Transactions on sonics and ultrasonics*, 32(3):458–464, 1985.

[58] H.M. Keller, W.E. Meier, M. Anliker, and D.A. Kumpe. Noninvasive measurement of velocity profiles and blood flow in the common carotid artery by pulsed doppler ultrasound. *Stroke*, 7(4):370–377, 1976.

[59] M. G. Kendall. *Rank correlation methods.* Griffin, Oxford, England, 1948.

[60] E. S. Kim, A. M. Sharma, R. Scissons, D. Dawson, R. T. Eberhardt, M. Gerhard-Herman, J. P. Hughes, S. Knight, A. Marie Kupinski, G. Mahe, M. Neumyer, P. Poe, R. Shugart, P. Wennberg, D. M. Williams, and R. E. Zierler. Interpretation of peripheral arterial and venous doppler waveforms: A consensus statement from the society for vascular medicine and society for vascular ultrasound. *Vasc. Med.*, 25(5):484–506, oct 2020.

[61] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 12 2014.

[62] P. A. Kyrle and S. Eichinger. Deep vein thrombosis. *Lancet*, 365(9465):1163–1174, mar 2005.

[63] C. Landry, A. Arami, and S.D. Peterson. Intermittent Pneumatic Compression: Towards Optimal Timing for Enhancing Lower Limb Blood Flow. In *Proceedings of the Canadian Society for Mechanical Engineering International Congress 2020*, Charlottetown,PE, 2020.

[64] C. Landry, E.T. Hedge, R.L. Hughson, S.D. Peterson, and A. Arami. Accurate blood pressure estimation during activities of daily living: A wearable cuffless solution. *IEEE J. Biomed. Health Inform.*, 25(7):2510–2520, 2021.

[65] C. Landry, S. D. Peterson, and A. Arami. Nonlinear dynamic modeling of blood pressure waveform: Towards an accurate cuffless monitoring system. *IEEE Sens. J.*, 20(10):5368–5378, 2020.

[66] T. P. Le, N. A. Vien, and T. Chung. A deep hierarchical reinforcement learning algorithm in partially observable markov decision processes. *IEEE Access*, 6:49089–49102, 2018.

[67] S. Lee, J. Kim, S.W. Park, S. Jin, and S. Park. Toward a fully automated artificial pancreas system using a bioinspired reinforcement learning design: In silico validation. *IEEE J. Biomed. Health Inform.*, 25(2):536–546, 2021.

[68] P.M. Lehrer. Biofeedback training to increase heart rate variability. *Principles and practice of stress management*, 3:227–248, 2007.

[69] A. W.A. Lensing, P. Prandoni, M. H. Prins, and H. R. Büller. Deep-vein thrombosis. *Lancet*, 353(9151):479–485, feb 1999.

[70] X. Liang, Y. Liu, T. Chen, M. Liu, and Q. Yang. *Federated Transfer Reinforcement Learning for Autonomous Driving*, pages 357–371. Springer International Publishing, Cham, 2023.

[71] T.P. Lillicrap, J.J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra. Continuous control with deep reinforcement learning, 2019.

[72] L. Ljung. *System Identification*, pages 163–173. Birkhäuser Boston, Boston, MA, 1998.

[73] Karl Logason, T Bärlin, M-L Jonsson, Annika Boström, Hans-Göran Hårdemark, and Sadettin Karacagil. The importance of doppler angle of insonation on differentiation between 50–69% and 70–99% carotid artery stenosis. *European Journal of Vascular and Endovascular Surgery*, 21(4):311–313, 2001.

[74] Y. Lu, Z. Yang, and Y. Wang. A critical review on the three-dimensional finite element modelling of the compression therapy for chronic venous insufficiency. *Proc. Inst. Mech. Eng. H.*, 233(11):1089–1099, nov 2019.

[75] B. A. MacRae, J. D. Cotter, and R. M. Laing. Compression Garments and Exercise. *Sports Med.*, 41(10):815–843, 2011.

[76] A R. Mahmood, D. Korenkevych, G. Vasan, W. Ma, and J. Bergstra. Benchmarking reinforcement learning algorithms on real-world robots. In *Conference on robot learning*, pages 561–591. PMLR, 2018.

[77] D. Makowski, T. Pham, Z. J Lau, J. C. Brammer, F. Lespinasse, H. Pham, C. Schölzel, and S. H. A. Chen. NeuroKit2: A Python toolbox for neurophysiological signal processing. *Behav. Res. Methods.*, 53(4):1689–1696, 2021.

[78] M. D. Malone, P. L. Cisek, A. J. Jr. Comerota, B. Holland, I. G. Eid, and A. J. Comerota. High-pressure, rapid-inflation pneumatic compression improves venous hemodynamics in healthy volunteers and patients who are post-thrombotic. *J. Vasc. Surg.*, 29(4):593–599, apr 1999.

[79] G. Mancia, A. Ferrari, L. Gregorini, G. Parati, G. Pomidossi, G. Bertinieri, G. Grassi, M. di Rienzo, A. Pedotti, and A. Zanchetti. Blood pressure and heart rate variabilities in normotensive and hypertensive human beings. *Circ. Res.*, 53(1):96–104, jul 1983.

[80] H.B. Mann. Nonparametric tests against trend. *Econometrica*, 13(3):245–259, 1945.

[81] A. P. Mnih, V.and Badia, M. Mirza, A. Graves, T. Harley, T.P. Lillicrap, D. Silver, and K. Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, page 1928–1937. JMLR.org, 2016.

[82] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller. Playing atari with deep reinforcement learning, 2013.

[83] C. Moffatt. Variability of pressure provided by sustained compression. *Int. Wound J.*, 5(2):259–265, 2008.

[84] R. J. Morris, H. Griffiths, and J. P. Woodcock. Analysis of the operation of the scd response intermittent compression system. *J Med Eng Technol*, 26(3):111–116, 2002.

[85] R. J. Morris and J. P. Woodcock. Evidence-Based Compression: Prevention of Stasis and Deep Vein Thrombosis. *Ann. Surg.*, 239(2):162–171, feb 2004.

[86] O. Nelles. *Linear Dynamic System Identification*, pages 457–546. Springer Berlin Heidelberg, Berlin, Heidelberg, 2001.

[87] E. A. Nelson, A. Hillman, and K. Thomas. Intermittent pneumatic compression for treating venous leg ulcers. *Cochrane Database Syst. Rev.*, 239(5), 2014.

[88] P. D. Ngo, S. Wei, A. Holubová, J. Muzik, and F. Godtliebsen. Reinforcement-learning optimal control for type-1 diabetes. In *2018 IEEE EMBS Int. Conf. Biomed. Health Inform. BHI*, pages 333–336. IEEE, 2018.

[89] V. Novak, P. Novak, J. de Champlain, A. R. Le Blanc, R. Martin, and R. Nadeau. Influence of respiration on heart rate and blood pressure fluctuations. *J. Appl. Physiol.*, 74(2):617–626, 1993. PMID: 8458777.

[90] National Institute of Health. Facts about peripheral artery disease (p.a.d.), 8 2006.

[91] Task Force of the European Society of Cardiology the North American Society of Pacing Electrophysiology. Heart rate variability. *Circulation*, 93(5):1043–1065, 1996.

[92] Joni Pajarinen, Jens Lundell, and Ville Kyrki. Pomdp planning under object composition uncertainty: Application to robotic manipulation. *EEE Trans. Robot*, 39(1):41–56, 2022.

[93] P. Poupart. *Partially Observable Markov Decision Processes*, pages 754–760. Springer US, Boston, MA, 2010.

[94] Z. Qiao, K. Muelling, J. Dolan, P. Palanisamy, and P. Mudalige. Pomdp and hierarchical options mdp with continuous actions for autonomous driving at intersections. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 2377–2382, 2018.

[95] A. Raza, K. Steinberg, J. Tartaglia, W. H Frishman, and T. Gupta. Enhanced External Counterpulsation Therapy: Past, Present, and Future. *Cardiol. Rev.*, 25(2), 2017.

[96] G.A. Reyes del Paso, W. Langewitz, L.J.M. Mulder, A. Van Roon, and S. Duschek. The utility of low frequency heart rate variability as an index of sympathetic cardiac tone: a review with emphasis on a reanalysis of previous studies. *J. Psychophysiol.*, 50(5):477–487, 2013.

[97] A Rezki, B. Merioud, D. Delmas, C. Cyril, R. Scheiwiller, R. Leblé, and P. Valensi. Sequential compression/decompression by a pulsating suit increases cutaneous microcirculatory blood flow in patients with type 2 diabetes. *Diabetes Metab.*, 42(4):298, 2016.

[98] Daniel W. Rickey. Doppler ultrasonography, 1 2006.

[99] S. G. Rizzo, G. Vantini, and S. Chawla. Reinforcement learning with explainability for traffic signal control. In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, pages 3567–3572, 2019.

[100] K. Sakai, N. Takahira, K. Tsuda, and A. Akamine. Effects of intermittent pneumatic compression on femoral vein peak venous velocity during active ankle exercise. *J. Orthop. Surg. Res.*, 29(1):2309499021998105, 2021.

[101] I. Santelices, C. Landry, A. Arami, and S. Peterson. Investigating Optimal Intermittent Pneumatic Compression Timing Across Two Days. In *Proc. Int. Conf. Body Sensor Netw*, 2023.

[102] Amit Sapra, Ahmad Malik, and Priyanka Bhandari. Vital sign assessment. *Vital sign assessment*, 2020.

[103] J. Schulman, S. Levine, P. Moritz, M.I. Jordan, and P. Abbeel. Trust region policy optimization, 2017.

[104] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms, 2017.

[105] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[106] Rebecca E Scully, Dean J Arnaoutakis, Ann DeBord Smith, Marcus Semel, and Louis L Nguyen. Estimated annual health care expenditures in individuals with peripheral arterial disease. *Journal of Vascular Surgery*, 67(2):558–567, 2018.

[107] F. Shaffer and J. P. Ginsberg. An Overview of Heart Rate Variability Metrics and Norms. *Front. Public Health*, 5:258, 2017.

[108] R. D. Sheldon, B. T. Roseguini, M. H. Laughlin, and S. C. Newcomer. New insights into the physiologic basis for intermittent pneumatic limb compression as a therapeutic strategy for peripheral artery disease. *J. Vasc. Surg.*, 58(6):1688–1696, dec 2013.

[109] Ryan D Sheldon, Bruno T Roseguini, M Harold Laughlin, and Sean C Newcomer. New insights into the physiologic basis for intermittent pneumatic limb compression as a therapeutic strategy for peripheral artery disease. *Journal of vascular surgery*, 58(6):1688–1696, 2013.

[110] Don Sheriff. Point: The muscle pump raises muscle blood flow during locomotion. *Journal of Applied Physiology*, 99(1):371–375, 2005.

[111] J. K. Shoemaker, Z. I. Pozeg, and R. L. Hughson. Forearm blood flow by Doppler ultrasound during rest and exercise: tests of day-to-day repeatability. *Med. Sci. Sports. Exerc.*, 28(9), 1996.

[112] D.U. Silverthorn, B.R. Johnson, W.C. Ober, C.W. Garrison, and A. C. Silverthorn. *Human physiology : an integrated approach*. Pearson Education Boston, Boston SE -, 6th ed edition, 2013.

[113] J. J. Skillman, R. E. C. Collins, N. P. Coe, B. S. Goldstein, R. M. Shapiro, N. T. Zervas, M. A. Bettmann, and E. W. Salzman. Prevention of deep vein thrombosis in neurosurgical patients: A controlled, randomized trial of external pneumatic compression boots. *Surgery*, 83(3):354–358, mar 1978.

[114] J.P.F. Sum, W.K. Kan, and G.H. Young. A Note on the Equivalence of NARX and RNN. *Neural Computing and Applications*, 8(1):33–39, 1999.

[115] X. Sun, L. Cheng, S. Plein, P. Garg, M. H. Moghari, and R. J. van der Geest. Deep learning-based prediction of intra-cardiac blood flow in long-axis cine magnetic resonance imaging. *Int. J. Cardiovasc. Imaging*, 39(5):1045–1053, 2023.

[116] R.S. Sutton and A.G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, second edition, 2018.

[117] C. Szepesvári. *For Further Exploration*, pages 63–64. Springer International Publishing, Cham, 2010.

[118] G. Theocharous and S. Mahadevan. Approximate planning with hierarchical partially observable markov decision process models for robot navigation. In *Proc. 2002 IEEE Int. Conf. Robot. Autom.*, volume 2, pages 1347–1352 vol.2, 2002.

[119] C. Thorn and A. Shore. *Medical Products Inspired by Biological Oscillators: Intermittent Pneumatic Compression and the Microcirculation*, pages 385–399. Springer International Publishing, Cham, 2021.

[120] O. Tochikubo, S. Ri, and N. Kura. Effects of pulse-synchronized massage with air cuffs on peripheral blood flow and autonomic nervous system. *Circ. J.*, 70(9):1159–1163, sep 2006.

[121] H. Tseng, Y. Luo, S. Cui, J. Chien, R.K Ten Haken, and I.E. Naqa. Deep reinforcement learning for automated radiation adaptation in lung cancer. *Med Phys*, 44(12):6690–6705, 2017.

[122] J. Urbankova, R. Quiroz, N. Kucher, and S. Z. Goldhaber. Intermittent pneumatic compression and deep vein thrombosis prevention. A meta-analysis in postoperative patients. *Thromb. Haemost.*, 94(6):1181–1185, dec 2005.

[123] P. S. van Bemmelen, M. A. Mattos, W. E. Faught, M. A. Mansour, L. D. Barkmeier, K. J. Hodgson, D. E. Ramsey, and D. S. Sumner. Augmentation of blood flow in limbs with occlusive arterial disease by intermittent calf compression. *J. Vasc. Surg.*, 19(6):1052–1058, jun 1994.

[124] R. Villar and R. L. Hughson. Repeatability of popliteal blood flow and lower limb vascular conductance at rest and exercise during body tilt using doppler ultrasound. *Physiol. Meas.*, 34(3):291, 2013.

[125] G. A. Vouros. Explainable deep reinforcement learning: State of the art and challenges. *ACM Comput. Surv.*, 55(5), dec 2022.

[126] R. Wade, F. Paton, and N. Woolacott. Systematic review of patient preference and adherence to the correct use of graduated compression stockings to prevent deep vein thrombosis in surgical patients. *J. Adv. Nurs.*, 73(2):336–348, 2017.

[127] C. Wang, X. Chen, L. Wang, M. Makihata, H. Liu, T. Zhou, and X. Zhao. Bioadhesive ultrasound for long-term continuous imaging of diverse organs. *Science*, 377(6605):517–523, 2022.

[128] H. Wang, A. Sakhadeo, A.M. White, J.M. Bell, V. Liu, X. Zhao, P. Liu, T. Kozuno, A. Fyshe, and M. White. No more pesky hyperparameters: Offline hyperparameter tuning for RL. *Trans. Mach. Learn. Res.*, 2022.

[129] C. J. C. H. Watkins. *Learning from Delayed Rewards*. PhD thesis, King's College, Cambridge, UK, May 1989.

[130] J. Weng, E. Hashemi, and A. Arami. Natural walking with musculoskeletal models using deep reinforcement learning. *IEEE Robot. Autom. Lett.*, 6(2):4156–4162, 2021.

[131] M. Winke and S. Williamson. Comparison of a Pneumatic Compression Device to a Compression Garment During Recovery from DOMS. *Int. J. Exerc.*, 11(3):375–383, 2018.

[132] Z. Yang, Y. Tian, T. Zhou, Y. Zhu, P. Zhang, J. Chen, and J. Li. Optimization of dry weight assessment in hemodialysis patients via reinforcement learning. *IEEE J Biomed Health Inform.*, 26(10):4880–4891, 2022.

[133] K. You, Z. Kou, M. Long, and J. Wang. Co-tuning for transfer learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Adv. Neural Inf. Process. Syst.*, volume 33, pages 17236–17246. Curran Associates, Inc., 2020.

[134] H. Zhang and T. Yu. *Taxonomy of Reinforcement Learning Algorithms*, pages 125–133. Springer Singapore, Singapore, 2020.

[135] Qiuxia Zhang, Lars-Göran Lindberg, Roland Kadefors, and Jorma Styf. A non-invasive measure of changes in blood flow in the human anterior tibial muscle. *European Journal of Applied Physiology*, 84:448–452, 2001.

[136] S. Zhang, P. Ma, A. Muheremu, R. Sun, H. Chai, and K. Jiang. Hemodynamic analysis of intermittent pneumatic compression combined with hyperthermia after total hip arthroplasty: an experiment on male rabbits. *Am. J. Transl. Res.*, 14(5):3344–3359, 2022.

[137] T. Zhang, K. Zhang, J. Lin, W. G. Louie, and H. Huang. Sim2real learning of obstacle avoidance for robotic manipulators in uncertain environments. *IEEE Robot. Autom. Lett.*, 7(1):65–72, 2022.

[138] S. Zhao, R. Liu, X. Wu, C. Ye, and A. W. Zia. A programmable and self-adaptive dynamic pressure delivery and feedback system for efficient intermittent pneumatic compression therapy. *Sens. Actuator A Phys.*, 315:112285, nov 2020.

[139] K. A. Zuj, E. T. Hedge, J. D. Milligan, S. D. Peterson, and R. L. Hughson. Intermittent compression of the calf muscle as a countermeasure to protect blood pressure and brain blood flow in upright posture in older adults. *Eur. J. Appl. Physiol.*, 121(3):839–848, 2021.

[140] K. A. Zuj, C. N. Prince, R. L. Hughson, and S. D. Peterson. Enhanced muscle blood flow with intermittent pneumatic compression of the lower leg during plantar flexion exercise and recovery. *J. Appl. Physiol.*, 124(2):302–311, 2018. PMID: 29122964.

[141] K. A. Zuj, C. N. Prince, R. L. Hughson, and S. D. Peterson. Superficial femoral artery blood flow with intermittent pneumatic compression of the lower leg applied during walking exercise and recovery. *J. Appl. Physiol.*, 127(2):559–567, 2019.

APPENDICES

Appendix A

This section will describe pilot studies that evaluate a DRL agent's CT timing. This information is not included as part of the main thesis due to difficulties with comparison of DRL agent CT selection to fixed CT selection. Consequently, this section will focus on the measured physiological signals rather than DRL agent insights.

A.1 Methods

The CGC IPC device used in this study and the data collection protocols are described in chapter 3 (see section 3.1.1).

A.1.1 Participants

The study participants included 3 healthy volunteers (2 males, 1 females) of age 27 $\pm$ 4 years (mean $\pm$ standard deviation) with sedentary to moderate exercise lifestyles. The average height and weight of the participants were 177 $\pm$ 11 cm and 80 $\pm$ 15 kg, respectively. Participant inclusion/exclusion criteria and ethics clearance is summarized in chapter 3 (see section 3.1.1).

A.1.2 Experimental Procedure

Participants were seated. The experiment consisted of a 5-minute baseline prior to the application of compression, 5 minutes of CT_d , followed by 30 minutes of either DRL agent timing or CT_d.

A.1.3 Environment Formulation

The learning environment is a user-specific predictive simulator of lower limb physiology created in MATLAB 2022a. The process of collecting data and training this predictive simulator, that consists of a BV estimation model, is described in chapter 3.

In the learning environment the DRL agent has a time step of 5 heartbeats so that metrics can be computed on a heartbeat-to-heartbeat basis and so that an agent time step is less impacted by BV variability over successive heartbeats. We formulate this problem as an MDP, defined as a tuple (S, A, T, R, γ) [116]. The formulation is the same as that described in chapter 4 unless otherwise stated. S, is the set of states, where each state is a 9 dimensional vector comprising of ECG and BV information over the previous 3 heartbeats. BV information includes the location and amplitude of the previous 3 systolic BV peaks, ECG information includes the previous 3 ECG R-wave amplitudes and previous 3 heartbeat values. Typical state values for a participant are shown in Fig. A.1.

The set of actions, A, is a discrete action space of 14 different CTs from 10% to 75%, inclusive, in steps of 5%. Though each agent time step is 5 heartbeats, compression is applied at each heartbeat. The reward function, R, is computed using BBMV, since we use BBMV to define the optimal CT for the DRL agent. The agent-environment interaction is shown in chapter 5 (Fig. 5.4). The reward function is shown below:

$$R = BBMV_t \tag{A.1}$$

where $BBMV_t$ is defined as the mean BBMV over the previous four heartbeats.

A.2 Results and Discussion

Table A.1 shows the mean BBMV, HR, and pulse delay, mean standard deviation of BBMV, HR, and pulse delay, and mean standard deviation of BBMV as a percentage of mean BBMV. These metrics are presented for no compression (5 minutes), fixed CT (30 minutes) and DRL CT (30 minutes).

Fig. A.2 shows typical BBMV and HR for fixed CT and for DRL CT during data collection over one day. Though the same fixed CT is applied for these 30 minute data collections, the BV response is different. As well, the HR and BBMV response is different from morning to afternoon. This figure highlights the variability associated with BV response to CT and the consequent challenges associated with deploying a DRL agent in

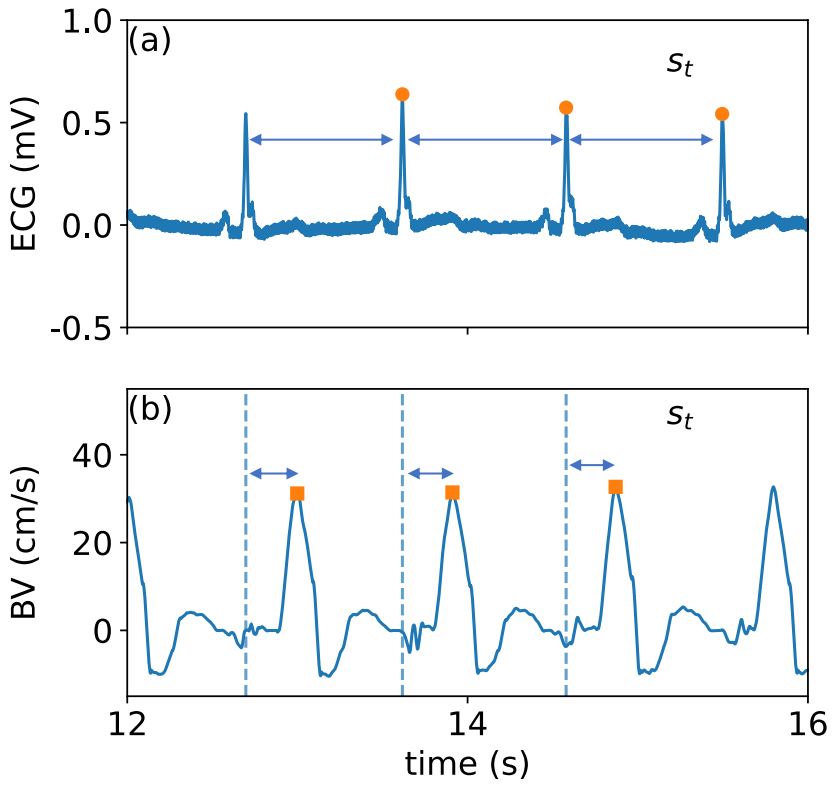


Figure A.1: (a) ECG R-peak amplitude (circles) and heartbeat values (arrows);(b) BV systolic peak amplitudes (squares) and BV systolic peak locations relative to the ECG R-wave location (arrows).

Table A.1: Mean and SD BBMV, SD of BBMV as a percentage of the mean BBMV, HR, and pulse delay for 5 minutes of no compression, 30 minutes of fixed CT, and 30 minutes of DRL CT selection at each heartbeat.

	No compression	Compression (fixed)	Compression (RL timing)
BBMV	1.62 cm/s ± 0.2	1.95 cm/s ± 0.99	1.95 cm/s ± 1.22
$BBMV_{\%\sigma}$	30 %	27 %	35 %
heart rate	66.0 bpm ± 6.1	62.0 bpm ± 3.2	61.6 bpm ± 4.2
pulse delay	282 ms ±8	288 ms ±8	286 ms ±12

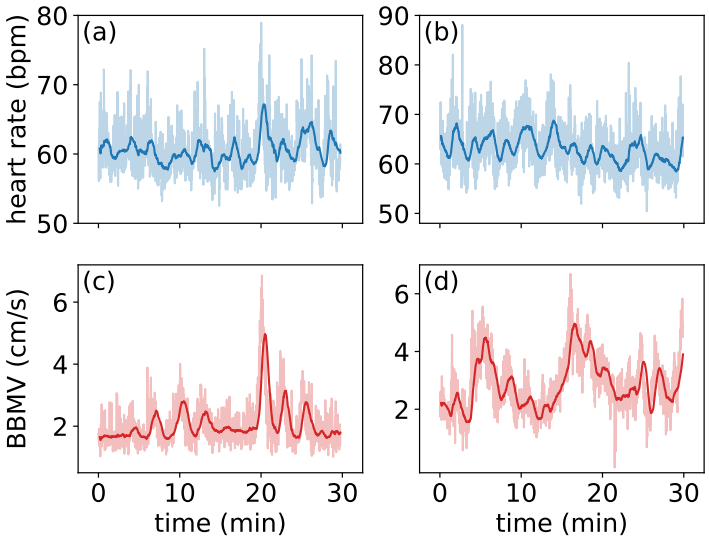


Figure A.2: BBMV and HR when fixed CT is applied for 30 minutes on the same day for the same participant: (a) morning HR; (b) afternoon HR; (c) morning BBMV; and (d) afternoon BBMV.

this complex environment. This figure also suggests that applying a CT for one minute [63], or three, as done in chapter 3, is not enough time to evaluate the BV response to a CT. The BBMV changes in Fig. A.2(d) even when the CT is held constant. A longer time period and more sensors are needed to better evaluate the BV response to different CTs.

Fig. A.3 shows the mean BBMV for different CTs. Fig. A.3(a),(b) are from the same day, morning Fig. A.3(a) and afternoon Fig. A.3(b). Fig. A.3(c),(d) were collected from the following day, morning Fig. A.3(c) and afternoon Fig. A.3(d). Only CTs that included more than 30 heartbeats were included in this plot and that is why some CTs do not appear. From one day to another, we observe that the BBMV response to CT is not constant from the morning to the afternoon and is not constant from one day to another. This plot highlights the difficulties associated with evaluating the DRL agent that had 14 CTs to select from.

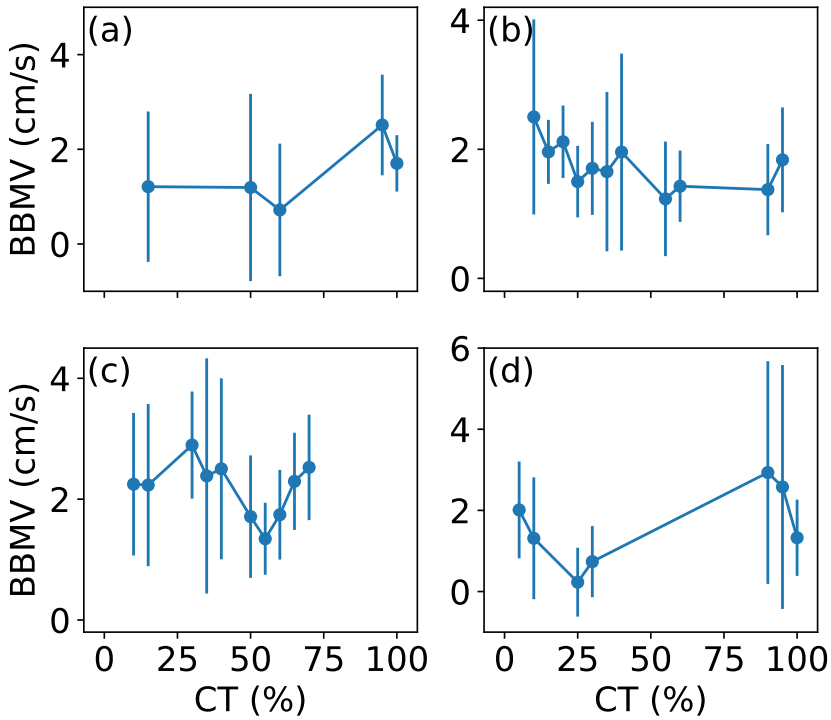


Figure A.3: mean and SD BBMV for different CTs: (a) day 1 morning; (b) day 1 afternoon; (c) day 2 morning; (d) day 2 afternoon.