ORIGINAL ARTICLE

Improving Reinforcement Learning Algorithm Based on Non-Negative Matrix Factorization Method for Controlling an Arm Model

Elham Farzaneh Bahalgerdy, Fereidoun Nowshiravan Rahatabad *

Department of Biomedical Engineering, SR.C., Islamic Azad University, Tehran, Iran

*Corresponding Author: Fereidoun Nowshiravan Rahatabad

Email: nooshiravan@gmail.com

Received: 28 June 2024 / Accepted: 21 July 2025

Abstract

Purpose: Reinforcement Learning (RL) is attracting great interest because it enables systems to learn by interacting with the environment. This study aims to enhance the RL algorithm to become more similar to human motor control by combining it with the Non-negative matrix factorization (NMF) method.

Materials and Methods: In the study, the signals recorded from six muscles involved in arm-reaching movement without carryinga certain weight.were pre-processed, and the optimal number of synergy patterns was extracted using NMF and the Variance Account For (VAF) methods. This, in turn, contributes to reducing the calculations. Subsequently, the robustness of the two-link arm model with six muscles was evaluated under various noise levels applied to the action coefficient matrix. Finally, the average synergy pattern was done on the mentioned arm model, and the RL algorithm controlled it by producing the action coefficient matrix.

Results: The average VAF% was 97.25±0.45%, and the number of synergies was four. The tip-of-the-arm model was able to reach the target after an average of 100 episodes.

Conclusion: The results indicated that the similarity in the extracted synergy patterns helps to model a system that is more similar to motor control. Additionally, the results of the synergistic patterns revealed that the two-link arm model with six muscles was suitable for the model. While controlling the model with the RL algorithm, the desired end-point position and path were achieved.

Keywords: Reinforcement Learning Algorithm; Non-Negative Matrix Factorization; Muscle Synergy; Action Coefficient Matrix; Optimization; Two-Link Arm Model.

DOI: https://doi.org/10.18502/fbt.v13i1.20781

1. Introduction

In early infancy, humans find initiation of movement control difficult, but over time and through training, they acquire knowledge and information about how to control their movements, eventually finding the best solution for the desired action. Reward and punishment can result in learning by humans.

Reinforcement Learning (RL) controllers are aligned with human motor control, whereas a controller such as the Proportional Integral Derivative (PID) controller cannot achieve this feature. The RL controller can be defined as the process of active learning while interacting with a constantly changing environment [1].

Thanks to the RL agent that generates suitable actions, the system can learn by interacting with the environment, taking actions, and gaining knowledge to reach the target through trial and error. The RL controller is based on the idea of learning from experience, which is accompanied by rewards and punishments, representing positive reinforcement (desired behaviors) and negative reinforcement (undesired behaviors), respectively [2]. An agent strives to maximize its future rewards by minimizing control costs. Systems that provide individual suggestions based on user behavior have been created through RL controllers [3]. In addition, adaptive shifting settings for survival and growth principles enable the provision of solutions to various issues in industries [4]. RL controllers have a significant impact on controlling upper extremity areas. Examples include combining the RL controller with the Functional Electrical Stimulation System (FES) [5], using the RL controller to learn how to predict the paddle target [6], and modeling multiactuator musculoskeletal systems using the RL algorithm [7]. Other examples include controlling the learnable parametrized model and a conventional feedback controller [8], as well as controlling the fuzzy neural network [9] by combining the RL algorithm. Humans often create effective and coordinated movements by ingeniously using the dynamics of their intricate musculoskeletal system. Through the Central Nervous System (CNS), various muscles that contain many motor units are activated and coordinated [10]. To handle the numerous Degrees of Freedom (DoF), humans do not have control over basic degrees of freedom; instead, they manage this issue using muscle synergistic patterns and the activation coefficient [9-12]. These synergistic patterns of muscles are similar in many cases [13, 14], and humans control movements by adjusting the activation coefficient, which in turn results in movement performance.

The Multi-Agent State Action Reward State Action (MA-SARSA) algorithm is an improved RL algorithm introduced by Martin *et al.* [15], which is based on the SARSA algorithm. This algorithm employs multiple agents to control complex systems, such as the multi-link arm model [16]. It can reduce the complexity of the agent, lower the learning speed, and minimize interference errors.

Jun Izawa *et al.* [17] discussed optimal learning control methods utilizing the RL controller for biological systems with redundant actuators.

Albers *et al.* [18] utilized the RL algorithm to control the two-link arm robot model, where the torque generated by the RL algorithm was applied as input to the robot. In a related study [16], the MA-SARSA algorithm was combined with the bee algorithm to control the two-link arm model with six muscles. Wannawas *et al.* [5] emphasized that complex environments are challenging to control using hand-crafted control policies, but the RL algorithm can learn to control them. Therefore, in FES control, RL is an essential component for governing the policies to control settings.

Analysis of arm-reaching movements has provided information about a limited number of fundamental training signals known as muscle synergy patterns, which govern diverse activities instead of separate commands to each muscle in the CNS. One of the methods to capture muscular synergies is NMF, which is more consistent than other methods such as Principal Component Analysis (PCA) and Independent Component Analysis (ICA).

Decomposing signals can be done by the NMF method which has greater robustness [19], as well as, NMF is a matrix factorization method. PCA generates factors that can be both positive and negative, whereas NMF exclusively produces positive factors.

PCA is useful when transforming a high-dimensional dataset into lower dimensions, provided that some loss of the original features is acceptable.

ICA is generally more computationally intensive than PCA, requiring more time to execute. NMF can identify more complex patterns in non-negative data. Additionally, NMF can effectively extract new features suitable for various applications.

The selection of an appropriate number of synergies relies on the VAF criterion, and the threshold value of the VAF should be chosen to describe the arm-reaching space more clearly while minimizing calculations. Great number of studies [13, 17, 19] have utilized the NMF method for extracting synergistic patterns [20-24].

It is believed that, under the same movement [25], highly modular muscular synergetic patterns can be achieved [12, 13, 25]. Therefore, in the present investigation, we used these similarities to reduce the computational burden of the MA-SARSA controller. Consequently, the NMF-reinforcement algorithm was able to achieve the desired end-point position and path. The proposed method in this study is to utilize methods such as the NMF and VAF methods to calculate the W.C matrix (where W was the weight matrix and C was the active coefficient matrix) after it was applied to the two-link arm model with six muscles. The W matrix was then applied to the two-link arm model, and the MA-SARSA algorithm was used to control it (see Figure 1).

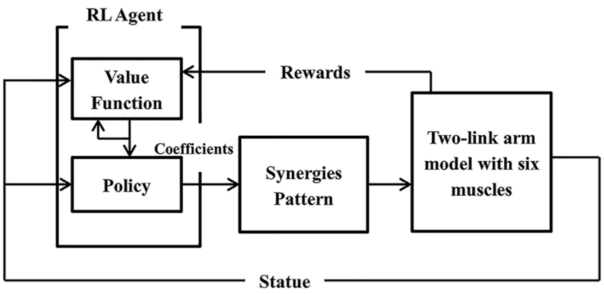

Figure 1. The diagram presents the NMF-MA- SARSA algorithm. The schematic of this paper suggests that combining the two-link arm model with the six-muscle and NMF algorithm, controlled by the MA-SARSA algorithm, can help reach the target

2. Materials and Methods

2.1. Study Population and Experimental Procedure

The study included twenty healthy, right-handed men with no neuromuscular disorders. Male' age averaged at 28 years (SD = 4.22 years), their average weight was 80 kg (SD = 11 kg), and their average height was 170 cm (SD = 8 cm). In this study, the participants were seated behind a desk while their shoulders and bodies were at a 90-degree angle (see Figure 2). Additionally, they were introduced to perform arm-reaching movements without carrying a specific weight along a specified path [29]. For each subject, the arm-reaching movement was performed up to 10 times.

It is essential to prepare the skin properly to reduce the skin's impedance. So, in this study firstly, the electrode area on the muscle should be cleaned with alcohol-soaked cotton to remove fat and perfume. Dead skin cells, which have high electrical resistance, should also be removed using a very fine sandpaper. During this process, continuous cleaning with alcohol-soaked cotton is necessary, and care must be taken to avoid damaging the skin. Surface hairs at the electrode locations should be removed. Finally, the skin's electrical impedance was measured with a multimeter, which should be less than ten kilo-ohms. EMG signals of six muscles, such as the biceps short head (BSH), biceps long head (BLH), pectoralis major (PMJ),

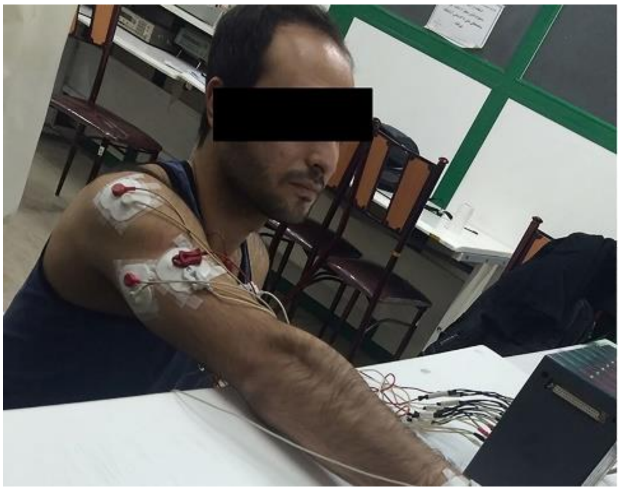
[redacted]

Figure 2. Experimental setup. The subjects sit at the table whose shoulders and bodies were at an angle of 90 degrees. Protocol was done at a certain endpoint position. Movements were recorded from each subject (20 arm-reaching movements in each person's protocol)

deltoids (DEL), triceps long head (TRIO), and triceps lateral head (TRIA) involved in arm-reaching movement [26], were recorded.

2.2. Data Acquisition Pre-Processing

The EMG signals were recorded at a sampling rate of 1 kHz through 5000 gain factor amplifiers (BIOPAC EMG 100A system). The electrode position was chosen according to the SENIAM standard [27].

The recorded EMG signals were passed through the high-pass filter at 1 Hz and the low-pass filter at 500 Hz outage frequencies. Subsequently, rectification, baseline correction, normalization, and activity level estimation were performed on the recorded signals (see Figure 3).

2.3. Non-Negative Matrix Factorization and Variability Accounted For

In the present study, to calculate the signal values, the NMF method and the required number of synergies were extracted by applying the VAF criterion [28]. The Non-Negative Matrix Factorization (NMF) method can be represented as Equation 1. Where W is the weight matrix, C denotes the coefficient matrix and e represents the residual error matrix:

$$M_j(t) = \sum_{i=1}^{r} W_{ij} C_i(t) + e(t), j = 1, \dots, m \qquad (1)$$

Figure 4 depicts information about an example of NMF method for two matrices (W and C matrices) with an inner dimension of K. Each column of the data (EMG signals) is expressed as a linear combination of basis vectors, specifically the columns of the matrix along with their corresponding weights.

The VAF criterion calculates to what extent W*H can be reconstructed from the original EMG data. The VAF criteria can be measured by Equation 2, where EMGori is the signal recorded from six muscles from 20 participants, and ||.|| is the Euclidean norm.

$$VAF = 1 - \frac{\|EMG_{ori} - (W \times C)\|^2}{\|EMG_{ori}\|^2} \qquad (2)$$

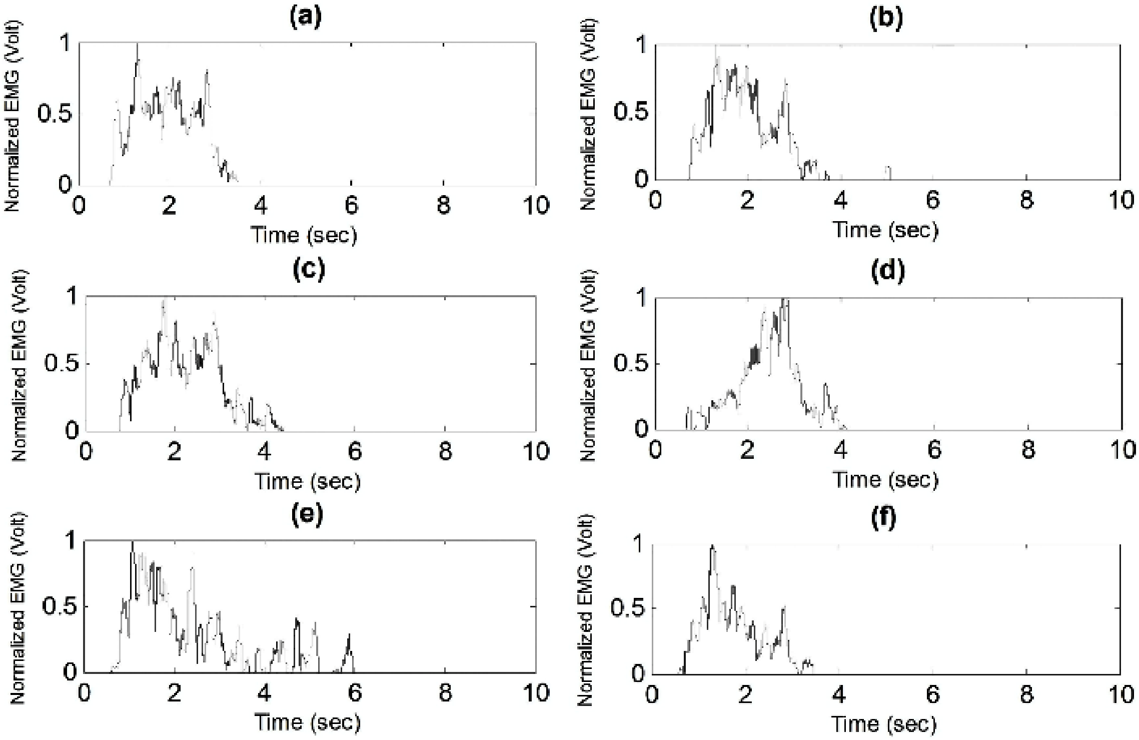

Figure 3. The entire EMG signal preprocessing for each muscle. The vertical and horizontal axis depicts normalized EMG (Volt) and time (sec), respectively

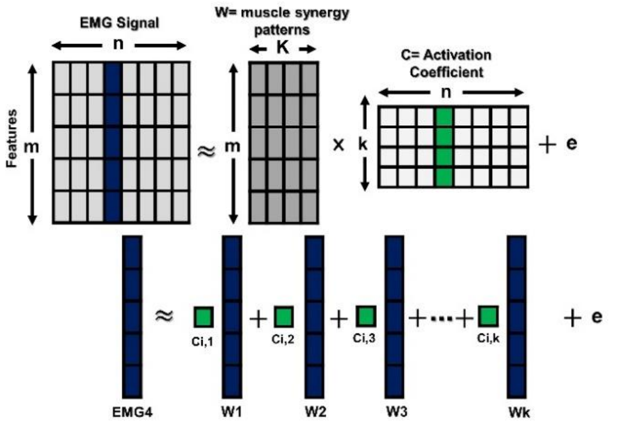

Figure 4. The process by which the NMF method functions to extract muscle synergy patterns matrix (W) and activation coefficient (C) from the EMG signals

2.4. Algorithm for Finding the most Effective VAF Threshold for Extracting the Optimum Number of Synergies

Extracting the optimum of the W.C affects the reduction of calculations and describes the arm-reaching space clearly. Therefore, in the present study, to choose the best VAF threshold, the following algorithm was used. The algorithm is repeated as long as the stop condition is satisfied. If the results (W.C matrix and desired value matrix) are in harmony, the VAF% is chosen as the best [29].

Step 1: Apply the desired average value matrix recorded from subjects as input to the two-link arm model.

Step 2: Select the initial VAF threshold.

Step 3: For each epoch step n=1, 2, ..., M:

Choose VAF threshold from 92 to 99:

Extract the W.C matrix using the NMF method.

Apply the extracted W.C matrix to the arm model as input.

Compare the output of the model results (W.C matrix and desired value matrix performed as input on the two-link arm model).

Step 4: If the results are not in harmony, increase the VAF value threshold (VAF= VAF+1) and go to Step 3.

Step 5: Stop if the stop condition is met [29].

2.4.1. Paired T-Test and Conditions for Using the Paired T-Test

The paired t-test is a crucial statistical method in order to compare two related sets of data. The Paired t-test is particularly applicable in cases where measurements are taken before and after a treatment or under similar conditions. Additionally, the paired t-test is suitable when there are two related sets of observations. For instance, it can be used for measurements taken before and after a treatment or for values produced by two methods applied to the same unit. In these cases, each value in the target group has a direct counterpart in the other group, justifying the use of the paired t-test. Since it is not assumed that one method consistently produces higher or lower values than the other, a two-tailed test is appropriate. The t-test examines any significant difference, whether positive or negative.

Sample size is a key factor in the reliability of statistical results. In this case, our dataset consists of 10,000 pairs, which represents a very large sample size. This is particularly important when the raw data may not follow a normal distribution. In such scenarios, the Central Limit Theorem applies, ensuring that the sampling distribution of the mean differences will be approximately normal.

Generally, the Paired t-test is as follows:

1. Compute the difference ($d_i = y_i - x_i$) for each pair of observations. x and y indicate test score before and after the module.

2. Determine the mean difference, denoted as d.

3. Calculate the standard deviation of these differences, s_d, and then use it to find the standard error of the mean difference, $SE(d) = s_d/\sqrt{n}$

4. Calculate the t-statistic using the formula

T=d/(SE(d)). According to the null hypothesis, this statistic follows a t-distribution with n − 1 degrees of freedom.

5. Refer to the t-distribution tables to compare your T value with the t_{n-1} distribution, which will provide the p-value for the paired t-test.

2.4.2. Cohen's d

In the realm of research and data analysis, one of the most important aspects is understanding the differences and effects of various variables on each other. In this context, Cohen's d is recognized as a key tool that helps researchers better grasp the depth and significance of observed differences. Cohen's d is a standardized measure of effect size that quantifies the magnitude of difference between two sets of values. While p-values tell us whether a difference is statistically significant, Cohen's d tells us how large or meaningful that difference is in practice. This feature becomes particularly important when sample sizes are large, as even small differences can become statistically significant but may lack practical importance.

To calculate Cohen's d for paired data, such as comparing a desired value to a W.C value for the same cases, the following formula is used (Equation 3):

$$D = standard\ deviation\ of\ differences \\ /\ mean\ of\ differences \tag{3}$$

In this formula, the mean of the differences represents the size of the difference between the two groups, while the standard deviation of the differences indicates the dispersion of these differences.

The value of Cohen's d can provide valuable insights into effect size. Generally, the following values indicate different effect sizes:

- Values around 0.2: small effect
- Values around 0.5: medium effect
- Values of 0.8 or more: large effect

2.5. Muscle Modeling Structure in the Horizontal Plane and MA-SARSA

The selection of the optimizer in the arm model influences the choices made by the MA-SARSA algorithm [16]. Therefore, in this study, the two-link arm model with six muscles in the horizontal plane [30] was simulated using MATLAB-2022 SimMechanics. This model had two degrees of freedom (2-DoF). The two-link arm model with six muscles was activated when the matrices (such as W.C) were performed as input. The lengths of the first link (a1) and the second link (a2) were 0.31m and 0.34 m, respectively. The arm reached the target at the specified point. Theta 1 represented the first joint angle located at the robot base, and Theta 2 represented the middle joint angle.

In this study, To control the model, the MA-SARSA method [15], which is one of the RL methods, was simulated using MATLAB code [31]. This model can be considered a better model compared to the classical RL controller, as continuous reward functions enhance the learning speed [15].

In the present study, after extracting synergy patterns and simulating the two-link arm model with six muscles, the Weight matrix (W) was applied to the model. To control the model, the MA-SARSA algorithm was simulated. Arm-reaching movement was achieved by generating coefficient matrices (C) using the MA-SARSA algorithm.

2.6. The General Method of MA-SARSA Algorithm

Generally, the MA-SARSA algorithm is as follows [15]:

Initialization:

Q-Table: Each agent i initializes a Q-table $Q_i(s,a)$ for all states s and actions a. Learning parameters: Learning rate a (e.g., 0.1), discount factor $0 \leq \gamma \leq 1$, and exploration rate ϵ (for ε- greedy policy):

```
    Start learning:
    Initialize $Q_i$ arbitararily; $\forall_i = 1, \dots, n$
    Repeat (for each episode)
    Observe $(S_1, \dots, S_n)$
    Select $a_i$ for $S_i$ by ε- greedy policy; $\forall_i = 1, \dots, n$
    Repeat (for J steps)
    Take actions $(a_1, \dots, a_n)$, observe r, $(S'_1, \dots, S'_n)$
    Select $a'_i$ for $s'_i$ by ε- greedy policy; $\forall_i = 1, \dots, n$
    $Q_i(s,a) = Q_i(s,a) + a[r + \gamma Q(s'_i, a'_i) - Q(s_i, a_i)]$
    $S_i = s'_i$; $a_i = a'_i$
  Until s is terminal
```

In the present study, the following algorithm was used in order to enhance the RL algorithm to become more similar to human motor control by combining it with the Non-negative matrix factorization (NMF) method.

Step 1: Insert weight matrix (W) as an input to the system

Step 2: Choose the coefficients matrix (C) and multiply it to the W matrix

Step 3: Apply it to the two-link arm model with six muscles

Step 4: Change the C matrix until reach the target in fewer steps.

3. Results

3.1. Synergy Extraction

In the study, synergetic patterns and the number of synergies were extracted using the NMF and VAF methods. As mentioned previously, extracting the appropriate number of synergies helps reduce computations and describes the wide space. This desired goal is achieved by choosing the best VAF criterion threshold. In this study, the number of synergies was selected as 4 to describe the movement, and the average VAF criterion extracted from all participants was 97.25±0.45%. Each number of synergy value ranges from 1 to 5, and the corresponding VAF% value is depicted.

Figure 5 (a and b) shows bar graphs that illustrate the average number of synergies extracted from 20 subjects from six major muscles involved in arm-reaching movement. Additionally, the bar graphs present information on the average number of synergies in four different categories for each of the six muscles.

Regarding the first number of synergies (W1) in the BSH and BLH muscles (see Figure 5a and b), the figures were the highest (18.31 and 12.97, respectively). The figure for the DEL muscle was also relatively high (7.06) compared to the figures for the PMJ, TRIA, and TRIO muscles, which were the lowest (4.59, 3.54, and 2.01, respectively). According to Figure 5 (a and b), looking at the second number of synergies (W2), the figures for the TRIA and TRIO muscles were the highest (8.33, 7.10, respectively). The figures for the other muscles were around 5. Regarding the third number of synergies (W3), the figures for the TRIO, TRIA, and DEL muscles indicated a greater involvement in arm-reaching movement (10.51, 9.15, and 5.78, respectively), while the figures for the Biceps and PMJ muscles were the lowest (approximately 2).

Finally, the average number of fourth synergies (W4) illustrates that the DEL and PMJ muscles showed the highest level of involvement in the arm-reaching movement (9.20 and 8.61, respectively). The figure for the TRIO muscle was also high (4.29) compared to the figures for the TRIA, BLH, and BSH muscles, which were the lowest (1.30, 2.18, and 0.87, respectively).

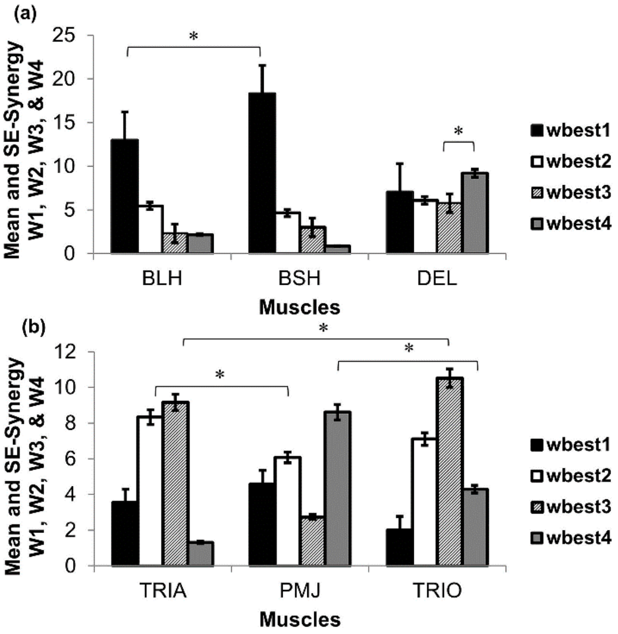

Figure 5. Plot of the average weight matrix for 20 participants with respect to the value role of six muscles, clustering the synergies into four groups. Bar graph (a) shows the value role of the BLH, BSH, and DEL muscles while bar graph (b) illustrates the value role of TRIA, PMJ, and TRIO muscles. The horizontal and vertical axes represent the number of muscles and the value role of the muscles in reaching movement, respectively

3.2. Extracting the Optimum W.C Matrix by the Best VAF% Threshold

In the present investigation, to choose the best VAF threshold, the algorithm mentioned in this paper (section 2.1.4) was utilized to find the most effective VAF threshold for extracting the optimum number of synergies in order to reducing the calculations done by MA-SARSA algorithm.

As shown in Table 1, the number of synergies (NS) can range from 1 to 5, with the optimal number determined by the VAF method. The average VAF % across 20 subjects were 97.25% (SD=±0.45), yielding four synergies. The first number of synergies achieving 96% variance in the input signal was deemed suitable for movement description. Given such an approach, NS>4 enjoyed this feature. On the opposite side, variation (representation matrix is more similar to the desired value matrix) in NS=5 decreased as compared to NS=4's variation but it can result in increasing the distance between the number of synergies and the main goal (that is dimension reduction).

Additionally, in this study, the results of the two-link arm model with six muscles when the desired value matrix (average EMG signals recorded from 20 subjects) and the W.C matrix (when the VAF% threshold of 96%) were used as inputs. The outputs included Theta 1, Theta 2, and endpoint position (EP) x and y. It is evident that the output of the W.C matrix were in harmony with the desired value matrix's outputs.

Table 1. The average VAF % criterion extracted from all 20 subjects was 97.25% (SD=±0.45). The horizontal axis shows the number of extracted synergies and the vertical axis represents the VAF% value. Four number of synergies were chosen as the appropriate number of synergies to describe the movement

Number of Synergy (NS)	VAF%	SD
1	76.2877	4.0116
2	89.7411	1.7356
3	93.7099	1.0641
4	97.2509	0.4552
5	99.0011	0.1689

3.2.1. Paired T-Test

The result obtained indicates a high p-value (see Figure 6 and Table 2). The outputs include Theta 1 (see Figure 6a), Theta 2 (see Figure 6b), and EP x and y shown in Figure 6c and d, respectively. The value reinforces that there is no strong evidence for a

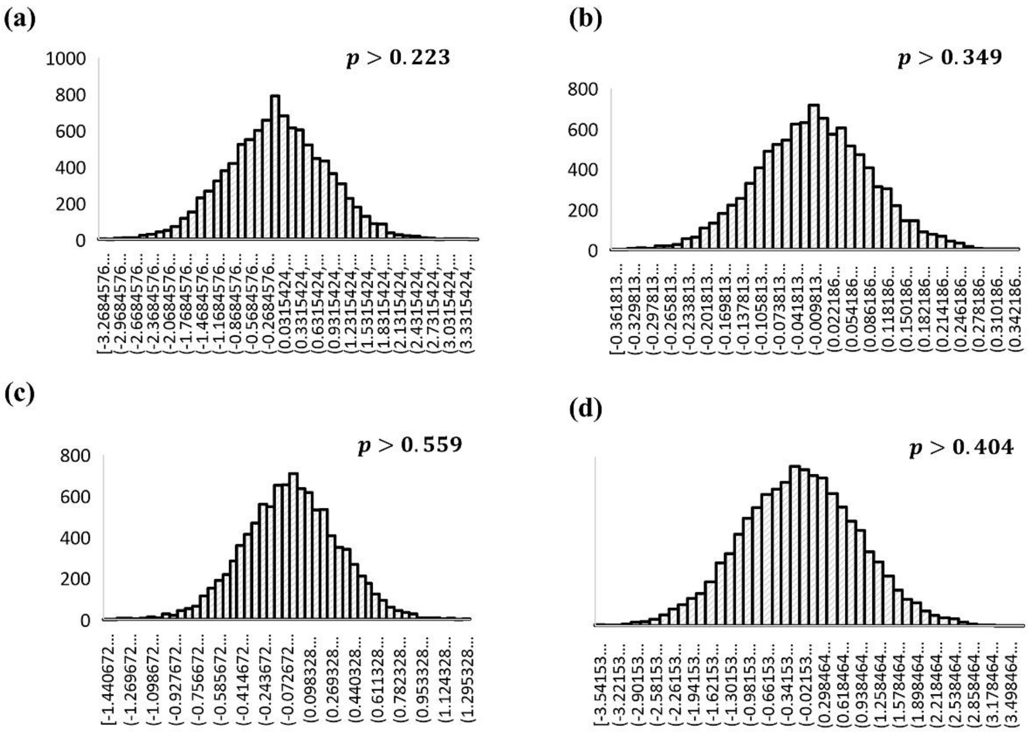

Figure 6. Histogram of outputs from the two-link arm model with six muscles. The differences of outputs include the W.C matrix (with a VAF criterion of 96%) and the desired value matrix, which comprises Theta 1 (a), Theta 2 (b), EP x (c), and EP y (d). a) The p-value for the Theta 1 desired and Theta 1 WC is 0.223. b) The p-value for the Theta 2 desired and Theta 2 WC is 0.349. c) The p-value for the end point position x desired and WC is 0.559. d) The p-value for the end point position y desired and WC is 0.404. The outputs are presented that by choosing a VAF threshold of 96% the optimum number of synergies can be achieved in this study based on section 2.4 algorithm. The results obtained from the tests indicate that there is no strong evidence for a meaningful difference

Table 2. Summary of paired t-test Results. The p-value is much greater than the alpha level (0.05) in all results

Variable	Null Hypothesis (H_0)	Alternative Hypothesis (H_1)	p-value	Conclusion	Interpretation Summary
Theta 1	No difference between desired and W.C values	There is a difference	0.223	Fail to reject H_0	No significant difference. Results consistent even with large sample size.
Theta 2	No difference between desired and WC values	There is a difference	0.349	Fail to reject H_0	No significant difference. Results consistent even with large sample size.
End Point position x	No difference between desired and WC values	There is a difference	0.5597	Fail to reject H_0	No significant difference. Results consistent even with large sample size.
End Point position y	No difference between desired and WC values	There is a difference	0.404	Fail to reject H_0	No significant difference. Results consistent even with large sample size.

meaningful difference between the two groups under consideration. In other words, we cannot confidently assert that one method or condition is significantly different from the other.

As can be seen in Table 2, the p-value is much greater than the alpha level (0.05) in all results. In Theta 1 outputs of the six muscles arm model, their differences are not significant, $p = 0.223$. Likewise, the difference between Theta 2 desired and Theta 2 WC are not significant, with a p-value of 0.349. Furthermore, the difference between the end point position y desired and the end point position y WC is not also significant, with $p = 0.404$.

Thanks to the large sample size and the paired nature of our data, the p-value produced by the t-test will be trustworthy. These results illustrate that by choosing a VAF threshold of 96%, the optimum number of synergies can be achieved.

3.2.2. Cohen’s d

Based on Table 3, all Cohen’s d values are very close to zero, indicating negligible practical differences between the desired and WC values for all variables. As can be seen, a very small difference between Theta 1 desired and WC values is observed, making the effect negligible (Cohen’s d=-0.01218669). Additionally, a very small difference is observed between Theta 2 desired and WC values (Cohen’s d = -0.00937019).

3.3. Applying the Arm Model and Studying the Robustness of W.C Matrices at Various Noise Levels

In the study, before performing W (weight matrix) on the two-link arm model with six muscles and controlling model by MA-SARSA, some random noise levels were applied to the coefficients matrix (C), and then they were applied with W matrix on the two-link arm model with six muscles.

Random noise can be measured by Equation 4. Where x is equal to 0.1, 0.2, 0.3, 0.4 and the interval is defined with m as the lower bound and n as the upper bound.

$$noise = x.mean(m,n).rand(m,n) \tag{4}$$

This is mainly because the study wanted to survey the robustness of the two-link arm model with six muscles when the MA-SARSA algorithm controls it by producing a random C matrix. If the two-link arm model is robust, then by performing the W matrix on the two-link arm model and controlling the model by

Table 3. Summary of Cohen's d results. All Cohen's d values are very close to zero

Variable	Cohen's d	Effect Size	Interpretation
Theta 1	-0.01218669	Negligible	The difference between Theta 1 desired and WC values is minimal, making the effect negligible.
Theta 2	-0.00937019	Negligible	The difference between Theta 2 desired and WC values is minimal, making the effect negligible.
End Point Position x	-0.00583212	Negligible	The difference between end point position x desired and WC values is minimal, making the effect negligible.
End Point Position y	-0.008348591	Negligible	The difference between end point position y desired and WC values is minimal, making the effect negligible.

the MA-SARSA algorithm, the system remains in its robustness.

The results of the error bar analysis of the W.C matrix (mean and Standard Deviation (STD)) for the 20 subjects were obtained by applying this matrix to the arm model.

These results were then compared with the findings of the error bar matrix W.C*, where C* represents the C matrix with various noise levels (x=0.1, 0.2, 0.3, 0.4). The outputs include Theta 1 (see Figure 7a), Theta 2 (see Figure 7b), and EP x and y shown in Figure 7c and d, respectively. Figure 7 illustrates the error bars for four cases, ranging from x= 0.1 to 0.4. The large dots indicate the mean (M) of the data. The error bars on the left (representing the W.C matrix) remained unchanged, while the error bar on the right depicts the W.C* matrix, where C varied at different noise levels. As x increased, the error bars on the right (W.C*) also increased. The probability that the right error bars were captured μ varies according to x and was greater for x=0.4.

3.4. Performing W Matrix in the Two-Link Arm Model with Six Muscles and Controlling it by Reinforcement Learning

In the present investigation, the MA-SARSA algorithm was aimed at generating the coefficient matrix, performed as input to the two-link arm model. This was achieved by multiplying the C matrix by the W matrix extracted by the NMF method. As a result, the W.C matrix could be considered as the input for each of the six muscles, generating forces for each muscle. These forces were then applied to the torque model [30], resulting in the generation of torques 1 and 2. These torques were applied to the joints, ultimately causing arm movement.

3.5. The Trajectory of the Two-Link Arm Model Controlled by the MA-SARSA Algorithm Using the NMF Algorithm was Examined in This Study

The RL controller uses two techniques to reach its target: the discovery technique, which involves the best task, and the experience technique, allowing policy RL methods to reuse past experiences. The agent strives to maximize its future rewards by minimizing control costs. During each episode of MA-SARASA, the two-link arm model did not follow a predetermined path to reach the target.

In the context of reinforcement learning, an episode refers to a sequence of interactions that starts from an initial state and ends in a terminal state, while a step refers to a single interaction between the agent and the environment. In each episode, the agent interacts with the environment with the goal of maximizing the total reward.

Figures 8 and 9 illustrate the trajectory of the two-link arm model controlled by the MA-SARSA and NMF-MA-SARSA algorithms, respectively.

Figure 10a shows that the arm model was able to reach the target after an average of 100 episodes. The horizontal and vertical axes illustrate the steps and the total reward which are obtained in a best pathway episode, respectively. In Figure 10a, the total reward was achieved after approximately 27 steps in the best pathway episode. The total reward amount was 25.

Table 4 is presented a brief comparison of the results of the MA-SARSA and NMF- MA- SARSA algorithms in control of six-muscle two-link arm. According to Table 4, after running approximately 100 episodes, the average number of steps produced by the NMF-MA-SARSA algorithm was 25. On the contrary, the average steps produced by the MA-SARSA algorithm was 32.

The plot of the C matrix (action coefficient matrix) produced by the MA-SARSA algorithm is depicted in Figure 10b. The values of the C matrix ranged from 0 to 1.

algorithm is depicted in Figure 6b. The values of the C matrix ranged from 0 to 1.

4. Discussion

An important issue in human motor control is the generation of a controller that is more similar to human motor control. One type of controller that works highly similar to the human motor controller is reinforcement learning. This is mainly because the

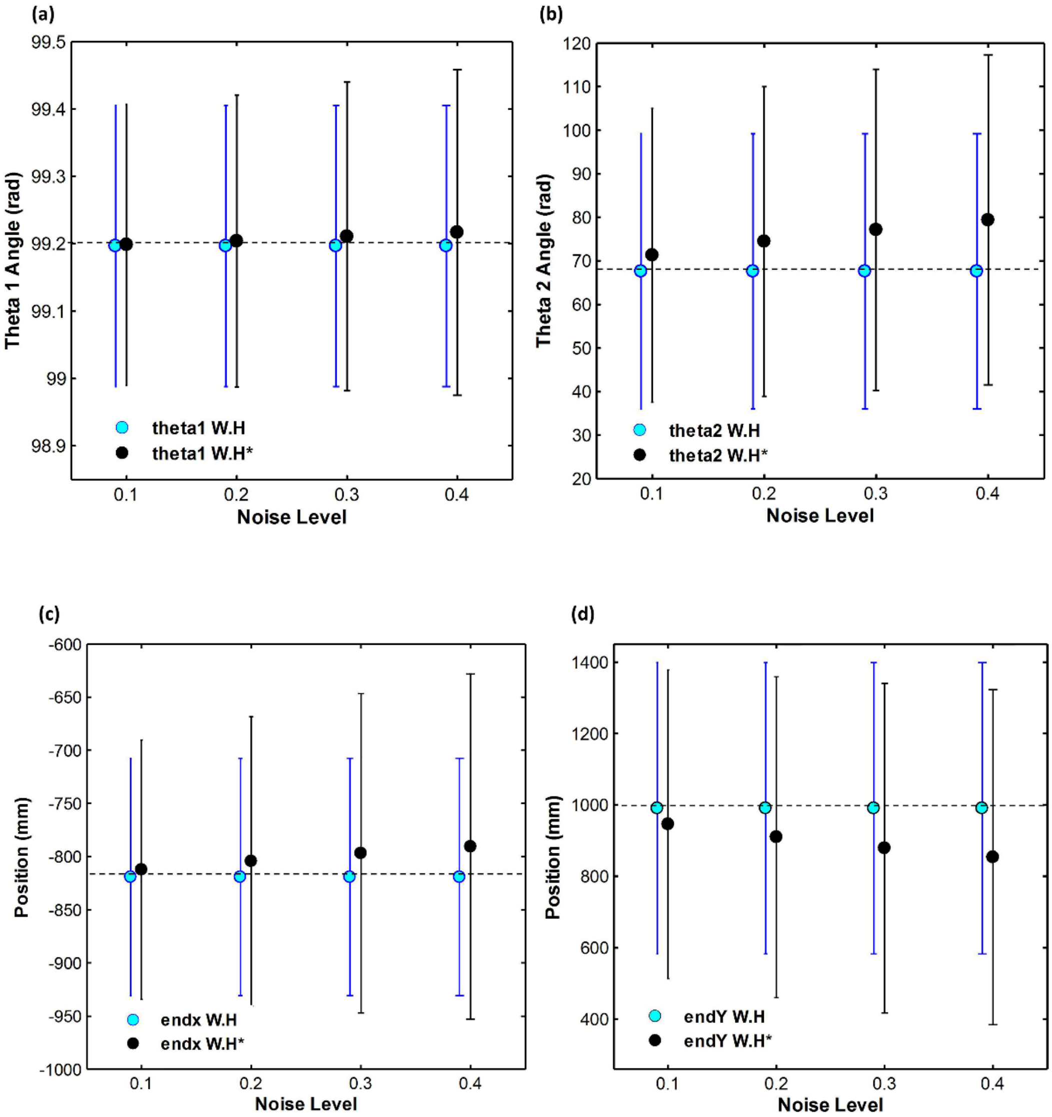

Figure 7. Plot of the error bar. Before applying the weight matrix (W) to the two-link arm model with six muscles, random noise levels were introduced to the coefficients matrix (C), defined by applying different values of x (x = 0.1, 0.2, 0.3, 0.4). Subsequently, both the noise-affected coefficients and the weight matrix were applied to the two-link arm model. The results were compared with those obtained by multiplying the coefficients matrix without noise and the weight matrix after being applied to the six-muscle model. The analysis of the error bar for the W.C matrix (without noise) was compared with the error bar of the W.C* matrix (with noise) (mean and STD) for the 20 subjects at various noise levels (x=0.1, 0.2, 0.3, 0.4). The outputs included Theta 1 (a) and Theta 2 (b) in the EP x (c) and y (d)

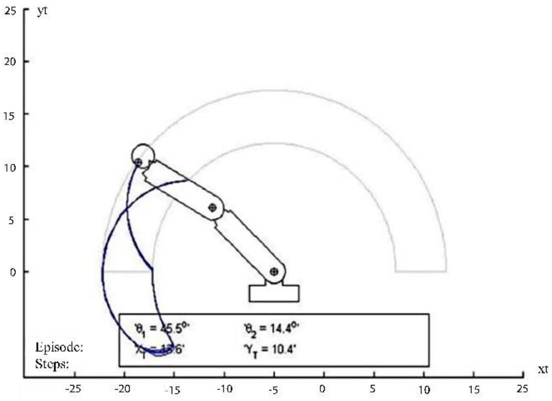

Figure 8. Plot of the arm trajectory. The two-link arm model trajectory controlled by the MA-SARSA algorithm

Table 4. Comparison of MA-SARSA and NMF- MA-SARSA algorithms in control of six-muscle two-link arm

Trajectory	The average number of steps	
Longer route and wider space	32±12.27	MA-SARSA algorithm
Shorter path and Limited space	25±10.63	NMF-MA-SARSA algorithm

RL algorithm is based on survival and growth performed by the reinforcement learning's agent, which in turn leads to finding the best solution for the desired action. The RL controller can be considered a powerful approach to developing capable and robust robot controllers [31]. In a previous study [15], the RL controller and the two-link arm model were used to successfully reach the target. However, a challenge encountered is the path to reach the target, and in addition, the MA-SARASA algorithm serves as the long-gain target. However, a challenge encountered is the path to reach the target, and in addition, the MA-SARSA algorithm requires time to achieve the target and slow learning speed, as well as, the algorithm is not optimizing the trajectory [15]. Since the algorithm is based on trial and error, it must explore a large number of states to achieve the desired outcome. The analysis of the high number of states contributes to the algorithm's sluggishness. Such as the NMF method, can be beneficial in the research for optimal paths within a smaller space.

In the study, the combination of the NMF method (to extract synergy patterns) and the two-link arm model with six muscles, controlled by the RL algorithm, allowed to achievement of the desired end-point position and path. The NMF method has been widely utilized in numerous studies [14, 20, 21] to extract synergy patterns. Previous research [13, 14] has shown that highly modular and similar muscular synergies are found among subjects who perform the same movement in many cases.

This similarity was also observed in the results of the extracted patterns. In this study, we used these similarities as a positive aspect of the NMF-MASARSA algorithm.

Determining the optimal number of synergies extracted by the NMF method not only reduces computational complexity [32] but also provides a

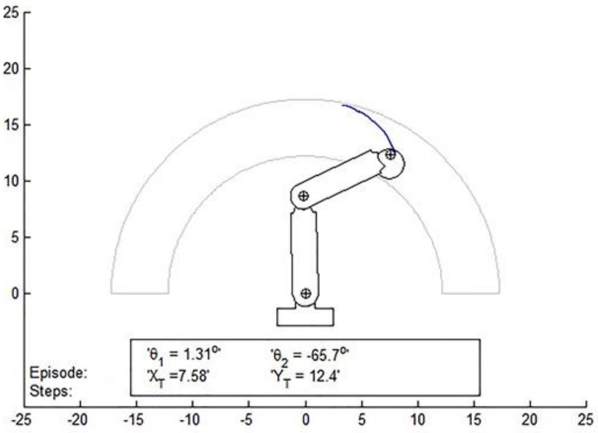

Figure 9. Plot of the arm trajectory. The NMF method, the two-link arm model trajectory, controlled by the NMF-MA-SARSA algorithm. In these 100 episodes the model has managed to find the optimal route

clear description of the arm-reaching space, as demonstrated in the present study. This optimum number of synergies can be achieved by selecting the best VAF threshold. In our study, the NMF method was used to extract synergetic patterns, which effectively reduced the movement space, thus reducing the computational burden of the MA-SARSA algorithm.

Figure 5 (a and b) provides information on the average number of synergies extracted from the 20 subjects. It is evident that the biceps, triceps, PMJ, and DEL muscles were the most involved in arm-reaching movements. These results highlight the suitability of the two-link arm model with six muscles as the best model in this context. In the present study, we utilized the similarity among the extracted synergetic patterns observed in many cases. Furthermore, as demonstrated in the study details, determining the optimal number of synergies that can be achieved by NMF and the best VAF threshold method would reduce the calculations. Furthermore, the MA-SARSA algorithm, which serves as a long-gain target [17], can benefit from an optimization algorithm (such as the NMF method) to find the optimal path. Therefore, these synergy patterns (represented by the W-weight matrix) were used as input to the two-link arm model with six muscles, and the MA-SARSA algorithm was used for control. As can be seen in Figure 8, without using the NMF method, the trajectory has traveled a long way in reaching the target. On the other hand, by utilizing NMF-MA-SARSA the trajectory has gone through a shorter route to the target (see Figure 9). It is seen in Table 4, the average steps produced by NMF-MA-SARSA algorithm was shorter than MA-SARSA algorithm, 25 and 32 respectively. Additionally, the NMF-MA-SARSA algorithm had a lower variance in the number of steps compare with the MA-SARSA algorithm. Through this approach, the MA-SARSA algorithm could learn to generate suitable actions represented by the desired C matrix and achieve the target on a desired path (see Figure 10).

According to Figure 10b, in each episode of control of the two-link arm model with six muscles, the MA-SARSA algorithm generated a C matrix of size [4*1]. When this value was multiplied by the W matrix (which was the weight matrix of size [6*4]), it resulted in the movement of the two-link arm model with six muscles. As mentioned above, the first number of synergies (W1) had the highest value compared to other numbers of synergies, while the first row of the C matrix had the minimum value compared to other rows of the C matrix.

4.1. Limitations and Future Directions

The present study has some limitations regarding the sample and methodology that should be considered. All participants were male, so future studies may include female participants. In addition, factors such as age and individual characteristics (e.g., activity level or athletic experience) may also influence the results.

To survey the robustness of the two-link arm model with six muscles when the MA-SARSA algorithm controls it by producing a random C matrix, future studies may utilize non-Gaussian noise on the model to explore how the system might operate under more realistic conditions, as non-Gaussian noise can pose challenges for data analysis. Furthermore, it is suggested that researchers focus on models of other parts of the body in future research.

Additionally, future studies should explore more than two links and/or more than six muscles involved in arm-reaching movements to investigate whether the results are consistent. However, increasing the number of muscles and links may raise the computational load, which could impact the final results.

5. Conclusion

The approach proposed in this study involves employing techniques like NMF and VAF methods to compute the W.C matrix. Subsequently, this matrix is applied to a two-link arm model with six muscles. As well as controlling the model with the MA-SARSA algorithm. The results of the NMF-MA-SARSA algorithm demonstrate that the controller was more similar to human motor control, reduced the computational requirements needed to reach the target, optimized the trajectory, and improved arm movement towards a specific target. The results indicate that the methods mentioned successfully achieve the desired destination and end-point position. Additional parameters should be identified and optimized to further improve the result.

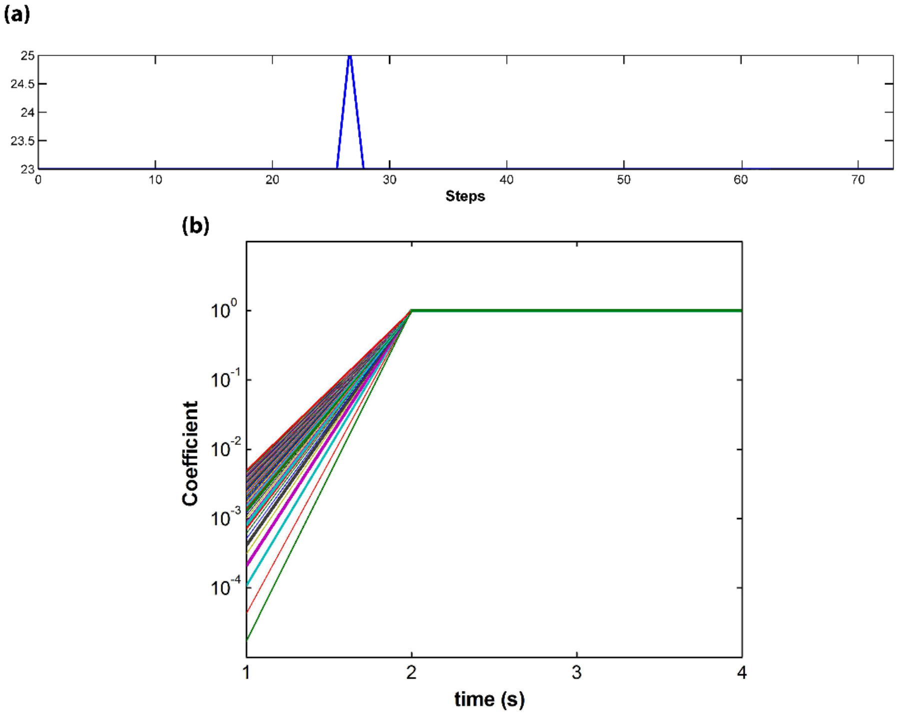

Figure 10. a) The learning chart for the two-link six-muscular arm using the NMF-MA-SARSA algorithm in the best pathway episode. The horizontal and vertical axes represent the steps and the total reward which are achieved in a best pathway episode, respectively. The total reward was achieved after approximately 27 steps in the best pathway episode and the total reward amount was 25. b) Plot of the action coefficient matrix for each episode (C best). The action coefficient matrix output in each episode of the model was controlled by MA-SARSA. As can be seen, the values of the C matrix range between 0 and 1

Acknowledgment

The authors express their gratitude to all volunteers and participants who participated in the study trial, along with those who participated in recording the data and carrying out the experiments.

References

1- Padmaja Kulkarni, Jens Kober, Robert Babuška, and Cosimo Della Santina, "Learning assembly tasks in a few minutes by combining impedance control and residual recurrent reinforcement learning." *Advanced Intelligent Systems,* Vol. 4 (No. 1), p. 2100095, (2022).

2- Claire Glanois *et al.*, "A survey on interpretable reinforcement learning." *Machine Learning,* pp. 1-44, (2024).

3- Frensi Zejnullahu, Maurice Moser, and Joerg Osterrieder, "Applications of Reinforcement Learning in Finance--Trading with a Double Deep Q-Network." *arXiv preprint arXiv:2206.14267,* (2022).

4- Dong Han, Beni Mulyana, Vladimir Stankovic, and Samuel Cheng, "A survey on deep reinforcement learning algorithms for robotic manipulation." *Sensors,* Vol. 23 (No. 7), p. 3762, (2023).

5- Nat Wannawas and A Aldo Faisal, "Towards AI-controlled FES-restoration of arm movements: neuromechanics-based reinforcement learning for 3-d reaching." In *2023 11th International IEEE/EMBS Conference on Neural Engineering (NER)*, (2023): *IEEE*, pp. 1-4.

6- Jonas Tebbe, Lukas Krauch, Yapeng Gao, and Andreas Zell, "Sample-efficient reinforcement learning in robotic table tennis." In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, (2021): *IEEE*, pp. 4171-78.

7- Kathleen M Jagodnik, Philip S Thomas, Antonie J van den Bogert, Michael S Branicky, and Robert F Kirsch, "Training an actor-critic reinforcement learning controller for arm movement using human-generated rewards." *IEEE Transactions on Neural Systems and Rehabilitation Engineering,* Vol. 25 (No. 10), pp. 1892-905, (2017).

8- Tobias Johannink *et al.*, "Residual reinforcement learning for robot control." In *2019 International*

Conference on Robotics and Automation (ICRA), (2019): *IEEE*, pp. 6023-29.

9- Meng Fan-Cheng and Dai Ya-Ping, "Reinforcement learning adaptive control for upper limb rehabilitation robot based on fuzzy neural network." In *Proceedings of the 31st Chinese Control Conference*, (2012): *IEEE*, pp. 5157-61.

10- Don Liang *et al.*, "Synergistic activation patterns of hand muscles in left-and right-hand dominant individuals." *Journal of Human Kinetics,* Vol. 76 (No. 1), pp. 89-100, (2021).

11- N Bernstein, "The co-ordination and regulation of movements, Oxford Pergamon." *Search in* (1967).

12- Cristiano Alessandro, Ioannis Delis, Francesco Nori, Stefano Panzeri, and Bastien Berret, "Muscle synergies in neuroscience and robotics: from input-space to task-space perspectives." *Frontiers in computational neuroscience,* Vol. 7p. 43, (2013).

13- Emilio Bizzi and Vincent CK Cheung, "The neural origin of muscle synergies." *Frontiers in computational neuroscience,* Vol. 7p. 51, (2013).

14- Vincent CK Cheung *et al.*, "Muscle synergy patterns as physiological markers of motor cortical damage." *Proceedings of the National Academy of sciences,* Vol. 109 (No. 36), pp. 14652-56, (2012).

15- José Antonio Martin and H De Lope, "A distributed reinforcement learning architecture for multi-link robots." in *4th International Conference on Informatics in Control, Automation and Robotics*, (2007), Vol. 192, p. 197.

16- Fereidoun Nowshiravan Rahatabad and Parisa Rangraz, "Combination of reinforcement learning and bee algorithm for controlling two-link arm with six muscle: simplified human arm model in the horizontal plane." *Physical and Engineering Sciences in Medicine,* Vol. 43 (No. 1), pp. 135-42, (2020).

17- Jun Izawa, Toshiyuki Kondo, and Koji Ito, "Biological arm motion through reinforcement learning." *Biological Cybernetics,* Vol. 91 (No. 1), pp. 10-22, (2004).

18- Albert Albers, Wenjie Yan, and Markus Frietsch, Application of reinforcement learning to a two DOF robot arm control. *Na*, (2009).

19- Xiaoling Chen *et al.*, "Muscle activation patterns and muscle synergies reflect different modes of coordination during upper extremity movement." *Frontiers in Human Neuroscience,* Vol. 16, p. 912440, (2023).

20- Vincent CK Cheung *et al.*, "Plasticity of muscle synergies through fractionation and merging during development and training of human runners." *Nature Communications,* Vol. 11 (No. 1), p. 4356, (2020).

21- Yushin Kim, Thomas C Bulea, and Diane L Damiano, "Novel methods to enhance precision and reliability in muscle synergy identification during walking." *Frontiers in Human Neuroscience,* Vol. 10, p. 455, (2016).

22- Akira Saito, Aya Tomita, Ryosuke Ando, Kohei Watanabe, and Hiroshi Akima, "Similarity of muscle synergies extracted from the lower limb including the deep muscles between level and uphill treadmill walking." *Gait & posture,* Vol. 59, pp. 134-39, (2018).

23- Alessandro Scano, Robert Mihai Mira, and Andrea d'Avella, "Mixed matrix factorization: A novel algorithm for the extraction of kinematic-muscular synergies." *Journal of Neurophysiology,* Vol. 127 (No. 2), pp. 529-47, (2022).

24- Daniel Soukup and Ivan Bajla, "Robust object recognition under partial occlusions using NMF." *Computational intelligence and neuroscience,* Vol. 2008 (No. 1), p. 857453, (2008).

25- Jingcheng Chen, Yining Sun, and Shaoming Sun, "Muscle synergy of lower limb motion in subjects with and without knee pathology." *Diagnostics,* Vol. 11 (No. 8), p. 1318, (2021).

26- Vincent CK Cheung, Andrea d'Avella, Matthew C Tresch, and Emilio Bizzi, "Central and sensory contributions to the activation and organization of muscle synergies during natural motor behaviors." *Journal of Neuroscience,* Vol. 25 (No. 27), pp. 6419-34, (2005).

27- Sensor Locations. [Online]. Available: http://seniam.org/sensor_location.html.

28- Hyun K Kim, Jose M Carmena, S James Biggs, Timothy L Hanson, Miguel AL Nicolelis, and Mandayam A Srinivasan, "The muscle activation method: an approach to impedance control of brain-machine interfaces through a musculoskeletal model of the arm." *IEEE Transactions on Biomedical Engineering,* Vol. 54 (No. 8), pp. 1520-29, (2007).

29- Fereidoun Nowshiravan Rahatabad and Elham Farzaneh Bahalgerdy, "The Most Effective VAF Threshold for Extracting the Optimum Number of Synergies for Reaching Movement in a Two-Link Arm Model with Two DoF." *Frontiers in Biomedical Technologies,* (2024).

30- Kenji Tahara, Zhi-Wei Luo, Suguru Arimoto, and Hitoshi Kino, "Task-space feedback control for a two-link arm driven by six muscles with variable damping and elastic properties." in *Proceedings of the 2005 IEEE International Conference on Robotics and Automation*, (2005): *IEEE*, pp. 223-28.

31- Tools for Reinforcement Learning, Neural Networks, and Robotics (Matlab and Python). [Online]. Available: http://jamh-web.appspot.com/download.htm#Reinforcement_Learning.

32- Ilge Akkaya *et al.*, "Solving Rubik'sRubik's cube with a robot hand." *arXiv preprint arXiv:1910.07113,* (2019).