Reward-based learning of a redundant task

Irene Tamagnone, Maura Casadio and Vittorio Sanguineti
Dept Informatics, Bioengineering, Robotics and Systems Engineering
University of Genoa
Genoa, Italy
Email: irene.tamagnone@unige.it

***Abstract*—Motor skill learning has different components. When we acquire a new motor skill we have both to learn a reliable action-value map to select a highly rewarded action (task model) and to develop an internal representation of the novel dynamics of the task environment, in order to execute properly the action previously selected (internal model).**
Here we focus on a 'pure' motor skill learning task, in which adaptation to a novel dynamical environment is negligible and the problem is reduced to the acquisition of an action-value map, only based on knowledge of results. Subjects performed point-to-point movement, in which start and target positions were fixed and visible, but the score provided at the end of the movement depended on the distance of the trajectory from a hidden via-point. Subjects did not have clues on the correct movement other than the score value. The task is highly redundant, as infinite trajectories are compatible with the maximum score. Our aim was to capture the strategies subjects use in the exploration of the task space and in the exploitation of the task redundancy during learning.
The main findings were that (i) subjects did not converge to a unique solution; rather, their final trajectories are determined by subject-specific history of exploration. (ii) with learning, subjects reduced the trajectory's overall variability, but the point of minimum variability gradually shifted toward the portion of the trajectory closer to the hidden via-point.

I. Introduction

Neuromotor recovery shares several features in common to motor skill learning [1], [8]. Therefore, understanding the mechanisms underlying the acquisition of a novel skill may help to derive more principled approaches to technology-assisted neuromotor rehabilitation.

The acquisition of a novel skill requires repetitive task performance. Motor skill learning has different components, each with their own peculiar mechanisms of action. Adaptation to a novel environment - an unfamiliar dynamics or a distorted geometry - is believed to require the development of an internal representation (internal model) of such dynamics or distorted geometry [7], [15], [19], which allows the motor system to predict its motor consequences. The development of an internal model is believed to be driven by the prediction error, i.e. the discrepancy between the actual and the predicted disturbance[16]. Tool use is another example of adaptation to a novel environment, in which an internal model of the tool has to be developed; see, for instance [4].

A task is usually described in terms of its degree of successful completion, which can be expressed as a 'score' or 'reward' signal - either explicit or implicit. Developing a 'task model', i.e. a mapping between a movement and its value (reward, or score) in the context of that particular task, is another component of motor skill learning. With respect to adaptation, the computational mechanisms underlying this aspect of motor learning have received less attention. Computational models of motor skill learning are frequently based on reinforcement learning, in which the objective is to find the optimal policy that selects actions so as to maximize reward probability. The reward prediction error, i.e. the discrepancy between the actual and the predicted reward is one likely candidate driving mechanism; see [6]. In many tasks, two components: (i) developing an internal model of dynamics and (ii) developing an internal representation of the action-value map, are tightly coupled and are difficult to dissociate [5]. Moreover, in most tasks the knowledge about the outcome of the movement is not limited to the score value. For instance, in obstacle avoidance tasks subjects not only get a score, but they also see how far they got from the obstacle [14]. This additional information on performance likely plays a role in learning the skill, which makes this component of skill learning more difficult to study experimentally.

Task redundancy - different movements are equivalent in terms of their goal - is another crucial aspect of motor skill learning and neuromotor rehabilitation. After an injury, impaired individual exploit their movement redundancy to find new ways to perform everyday life activities. Therefore, to investigate tasks with redundant solutions is relevant to understand how the brain deals with the redundancy problem. The way redundancy is exploited may provide information about how an internal model of the task is established - the structure of the 'task model'. This can be captured by looking at the pattern of inter-trial variability. Task-relevant features of the movement are expected to exhibit less variability than task-irrelevant ones [12], [9]. A related aspect is that movements equivalent in terms of their associated score may differ in terms of their associated mechanical effort [18].

Here we focus on the 'task model' component of motor learning, by studying a motor learning task in which the role of adaptation is negligible and in which a score, visualized at the end of the movement, is the sole information available on task performance. If subjects don't have clues on task solution other than the score value, their only option will be to explore the space of possible movements in search of high-score regions. Moreover, if the task is redundant, they will need to choose among infinite movement trajectories that are equivalent in terms of task requirements. We will specifically look at how subjects exploit task redundancy to develop an optimal solution for this task.

978-1-4673-6024-1/13/$31.00 ©2013 IEEE

II. Materials and Methods

A. Subjects

Six healthy right-handed subjects (4 male - 2 female, age 29 ± 3 years) participated in the experiment. The research conforms to the ethical standards laid down in the 1964 Declaration of Helsinki that protects research subjects. Each subject signed a consent form that conforms to these guidelines.

B. Experimental apparatus and Task

Subjects sat in front of a computer screen and grasped the handle of the planar robot manipulandum Braccio di Ferro (BdF) (Celin srl, La Spezia, Italy); see [2] for details. The robot did not generate forces. We only used its optical encoders to record the end effector position during the movements. Subjects had to perform point-to-point movements between two fixed locations, from a starting point (white circle, $\oslash 3$ cm) to a target point (yellow circle, $\oslash 3$ cm), both displayed on the computer screen at a horizontal distance of 20 cm; see Figure 1. The current position of the end effector was continuously displayed as a red cursor ($\oslash 1$ cm); the ongoing trajectory of the end effector was also displayed as a red trace. At the end of the movement, a numeric score (0-100) was displayed on the screen. In addition, a text message warned the subjects if the movement was either too slow (speed < 1 m/s) or too fast (speed > 1.2 m/s). However, subjects were not penalized if the movement speed was not in the suggested range. Subjects were told that their goal was to vary the shape of the movement trajectory in order to achieve the maximum score. They were also told that the score was related to trajectory shape, not to its duration. The score was calculated in terms of the minimum distance d of the movement trajectory from a (hidden) point, placed at a fixed location (see Figure 1). If $d \leq 1$ cm, the score was set to 100. For $d > 1$ cm, the score decreased according to a Gaussian profile, whose standard deviation was calculated so that the score was 0 for $d \geq 4$ cm. The task is highly redundant (any trajectory that passes through the via-point gets the same score) and the only information provided to subjects is the score value at the end of each movement.

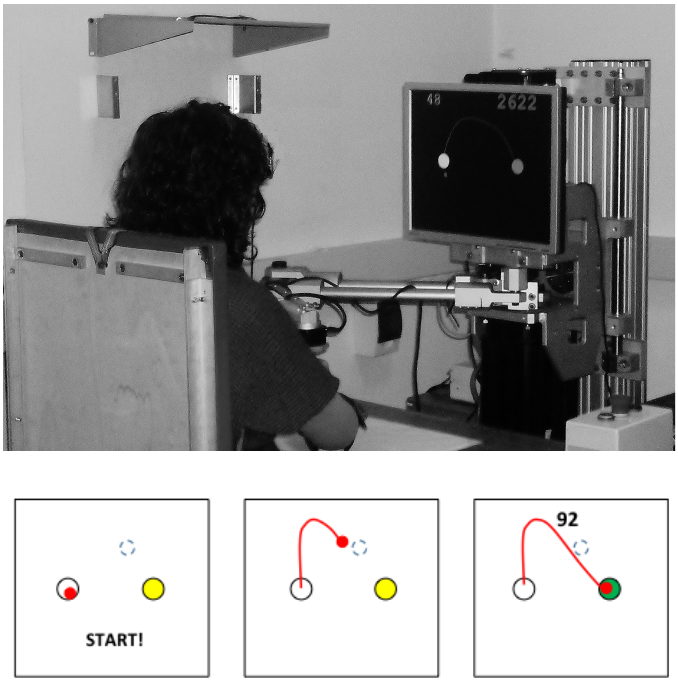

Fig. 1. Experimental apparatus and task. Participants start moving from the starting position toward the target. At the end of the movement a score is displayed, which reflects the distance of the trajectory from a (hidden) via-point

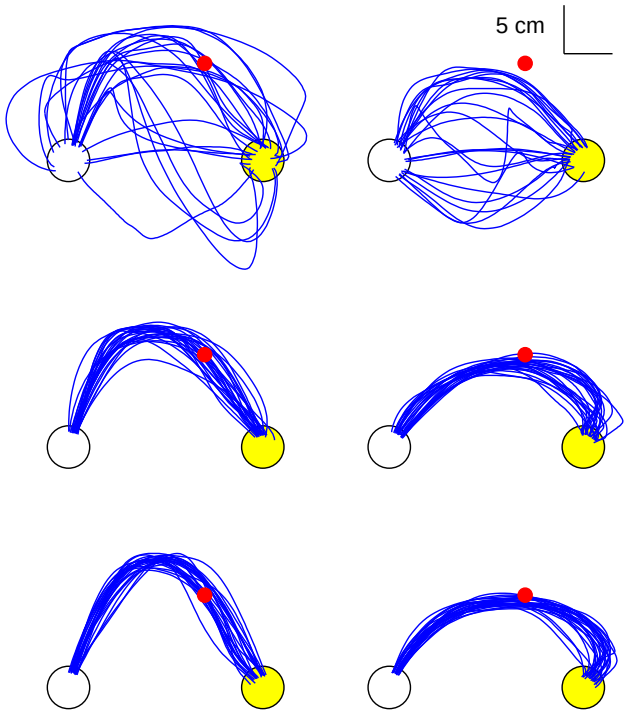

Fig. 2. Movement trajectories in the early, middle and late phases of the experiment for two typical subjects, S4 and S6. The red dot denotes the (hidden) via-point

C. Experimental Protocol

The experiment consisted of 150 movements, splitted into 3 epochs (50 movements each). A pause of 60 sec was scheduled among the epochs.

D. Data Analysis

Hand trajectories were sampled at 60 Hz and stored for subsequent data analysis. The recorded trajectories and all the relevant derivatives were smoothed by means of a 6th order Savitzky-Golay filter with a 170 ms time window (cut-off frequency: 7.5 Hz).

1) Learning and exploration: As performance indicator we took the reward signal (score). We also looked at what movement subjects converged to, and how did they get there, i.e. how they explored the space of possible movements. To do so, we looked at inter-trial movement variability (similarity from two consecutive trajectories). We specifically focused on two similarity measures: figural distance and the correlation of the speed profiles. While the first indicator provides information related to the spatial component of the trajectories, speed profile correlation also accounts for the time component.

a) Figural Distance: Given two trajectories A and B (consisting respectively of n_A and n_B points each), assuming that $d_{AB}(i) = \min_j \|\overrightarrow{x}_A(i) - \overrightarrow{x}_B(j)\|$ is a vector containing the distances between the trajectory B and each point in A, whereas the vector $d_{BA}(i)$ contains the distances between the trajectory A and each point in B, the figural distance (FD) between A and B is defined [3] as:

$$FD_{AB} = \frac{1}{n_A + n_B}\left[\sum_{i=1}^{n_A} d_{AB}(i) + \sum_{i=1}^{n_B} d_{BA}(i)\right] \quad (1)$$

FD measures differences in shape, irrespective of the differences in speed. We calculated the FD between each subsequent pair of trajectories.

b) Speed Profile Correlation: Given the speed profiles of trajectories A and B, $v_A(i)$ and $v_B(i)$, the correlation between A and B is defined as:

$$C_{AB} = \frac{\max_\tau [c_{AB}(\tau)]}{\sqrt{c_{AA}(0) \cdot c_{BB}(0)}} \quad (2)$$

where $c_{AB}(\tau) = \frac{1}{n}\sum_{i=1}^{n}[(v_A(i) - \overline{v}_A) \cdot (v_B(i+\tau) - \overline{v}_B)]$ with $n = \min(n_A, n_B)$ is the cross-covariance of the two speed profiles.

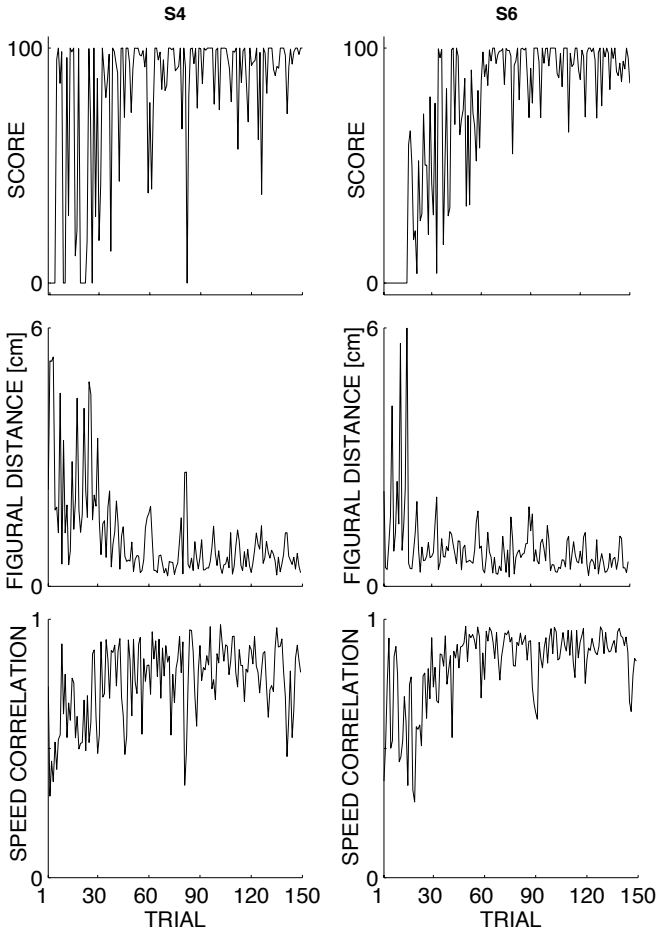

Fig. 3. Trial by trial evolution of score (top), and trial-by-trial similarity, measured in terms of figural distance (middle) and speed correlation (bottom), for subjects S4 and S6

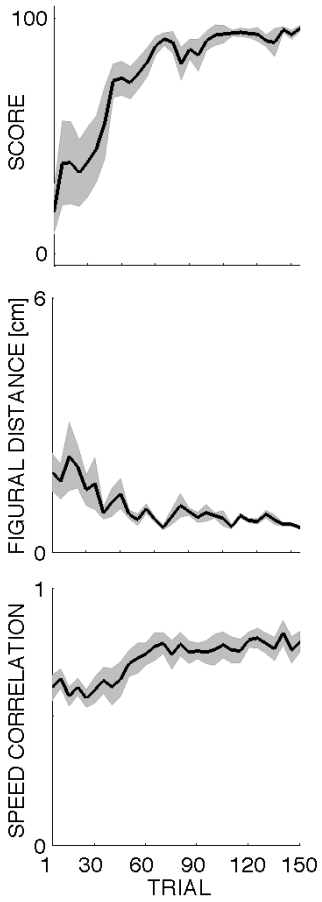

Fig. 4. Trial by trial evolution of score (top), and trial-by-trial similarity, measured in terms of figural distance (middle) and speed correlation (bottom). The line represents the average of all subjects, the dashed area the SE

2) Spatial variability: A look at movement variability may also shed some light on how redundancy is exploited. If subjects learn optimal trajectories, they may keep exploring the high-score portion of the solution space to find less effortful solutions. Alternatively, they may keep the first solution that they get into, and no further optimization occurs. To understand what parts of the trajectory are, respectively, more and less variable, for each group of 25 consecutive movements we calculated a spatial variability index; see also [14]. First, we calculated the mean spatial path, by resampling each path in a fixed number of equidistant samples (100) and averaging each trial to compute the mean spatial path. Then, for each location along the path (in 10% increments), the spatial variability was computed as follows: (i) We calculated the normal direction with respect to the mean tangential velocity; (ii) For each individual trajectory, we selected the data points that intersected the normal direction, and (ii) we took the standard deviation of these points with respect to the average trajectory as a measure of spatial variability for that portion of the trajectory.

III. Results

Figure 2 depicts the temporal evolution of the movement trajectories during the early ($1-25$ trials), middle ($51-75$

trials) and late ($125 - 150$ trials) phases of the experiment, for two different subjects. Subjects initially attempted different movements, until they hit a non-zero score region. As soon as subjects got a non-zero score, they narrow down the exploration. At the beginning of the middle phase, the trajectory has almost stabilized.
A closer look at the temporal evolution of score and trial-by-trial variability - see Figure 3 - suggests that subjects do not completely give up exploration after the early phase of learning, as shown by the occasional increases of the figural distance and/or the decreases in the speed correlation. The limited degree of co-variation between figural distance and speed correlation suggests that exploration may involve the spatial, the temporal aspects of the movement, or both. On average over all subjects, these same quantities suggest that exploration continues to the very end of the trial (non-zero figural distance and speed correlation less than one). Other factors - for instance, fatigue and loss of attention/motivation in the later phases of the experiment - may also play a role; see Figure 4.

At the end of the experiment, all subjects converged to a relatively stable, consistent trajectory. However, the learned trajectory was not the same for all subjects. Figure 5 depicts the last ten movement trajectories made by each subject - which can be interpreted as the movements 'learned' by that subject - together with their respective speed profiles. After completion of the experiment, all subjects were questioned about the strategy they had developed. In general, the explanations were rather vague. The majority reported that they targeted regions of the space that they had identified as associated with an high score. One subject suggested he was aiming at a specific, high-score spatial path; no one guessed that the score was determined by the distance from a single via-point. Figure 5 suggests that the final trajectories are not the same for all subjects, and are determined by the (subject-specific) history of exploration.

Another important information is provided by the spatial variability observed along the average path. In point-to-point movements with a via-point, the optimal feedback control hypothesis of motor control [18] predicts that when crossing the via-point, the optimal trajectory exhibits a minimum of spatial variability.
A look at the evolution of spatial variability in the trajectories learned by different subjects, see Figure 6, suggests that subjects do not generally converge to the optimal control solution. The variability decreases across trials in all points of the trajectory, but the lowest variability, and hence the greatest repeatability, is always observed in the first half of the trajectory. In some subjects the point of minimum variability gradually gets closer to the point at minimum distance from the hidden via-point. This would have been consistent with the optimal control hypothesis, but is not true for all subjects.
Finally, the point of maximum curvature does not exhibit a consistent relation with the via-point. In the Figure 2, in the two subjects the maximum curvature is found, respectively, before (subject S4) or after (subject S6) the via-point.

IV. Discussion

We have investigated a 'pure' motor skill learning task in which a complex trajectory has to be learned solely on the basis of knowledge of results (a numeric score). The task is inherently redundant as infinite trajectories are task-equivalent i.e. they give the same score. The issue is relevant to neuromotor rehabilitation because, when trying to recover a functional behavior, stroke survivors need to explore the space of their residual movements and to exploit redundancy for identifying movements that are more efficient in terms of the associated effort. In this process, knowledge of results is often the only available information.

A. *Is there a unique 'optimal' solution?*

Our results clearly demonstrate that each subject converged to a different solution. An optimal control formulation of this problem - based on a complete knowledge of the task (position of the via-point) - predicts that the 'optimal' trajectory in terms of minimization of variability and effort would correspond to matching the hidden via-point with the point of minimum path variability. In contrast, our subjects do not converge to a solution in which these two points are coincident. Possible interpretations are that either (i) subjects don't use optimality principles to select their movement, or (ii) the observed trajectory is still optimal when lack of knowledge of via-point position is accounted for. This is a matter of future investigation.

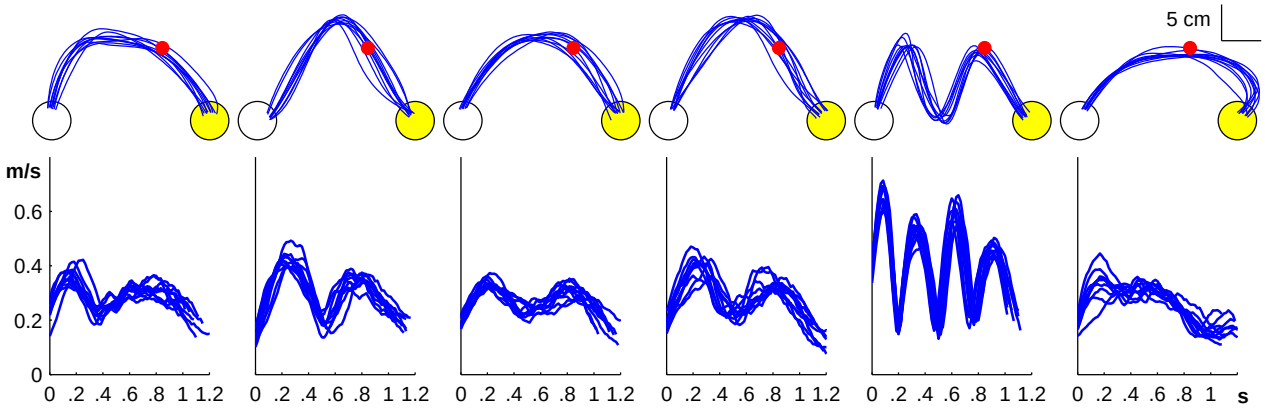

Fig. 5. Movements (top) and speed profiles (bottom) learned (last 10 movement trajectories) by each individual subject. From left to right: subjects S1-S6

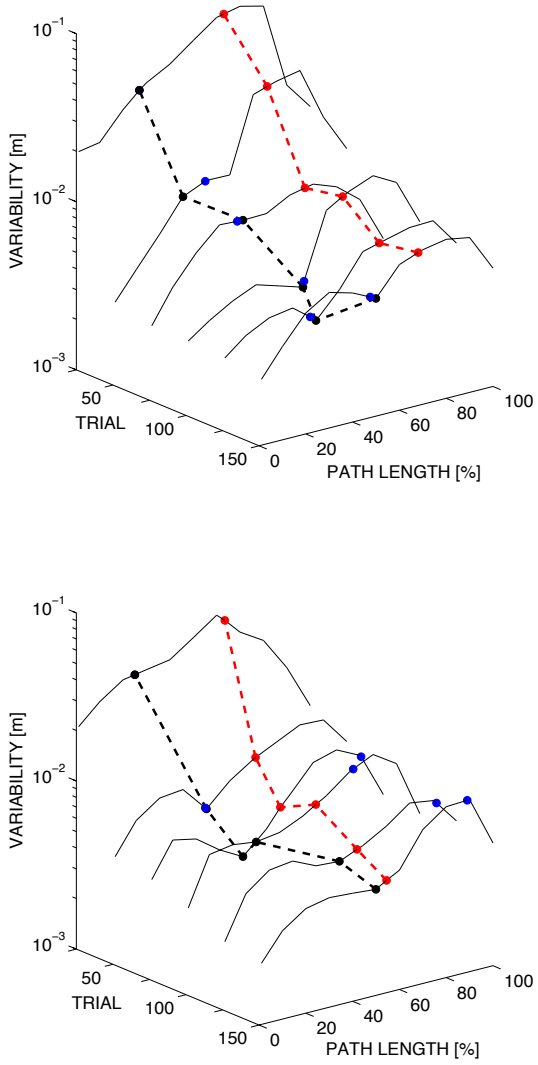

Fig. 6. Evolution of movement variability throughout the course of the learning epochs, for subjects S4 and S6, calculated at different fractions of the total path length. The red dashed line denotes the the point of minimum distance from the hidden via-point. The black dashed line represents the point of minimum path variability. Blue dots indicate the point of maximum curvature

B. Task-relevant vs task-irrelevant variability

Many studies - e.g. [10], [11], [13] - have shown that motor skill learning is characterized by a decreased variability in the task-relevant components of the movement. Similar to [14], [17], here we found that spatial variability decreases with training over the whole movement path; see Figure 6. However, a greater decrease is observed in the portions of the path that are closer to the hidden via-point. A look at the pattern of spatial variability along the movement path suggests that there is a point in which variability reaches a minimum - the black dots in Figure 6). Moreover, this point gradually converges to the portion of the path that has minimum distance to the hidden via-point - the red line in Figure 6. This suggests that subject gradually incorporate the (hidden) structure of the task in their 'task model'.

ACKNOWLEDGMENTS

This work was partly supported by the EU Grant FP7-ICT- 271724 (HUMOUR), by a grant from the Italian Ministry of Research (PRIN 2009), and by the COST Action TD1006 (European Network on Robotics for NeuroRehabilitation).

REFERENCES

[1] E. Burdet, V. Sanguineti, H. Heuer, and D. Popovic. Motor skill learning and neuro-rehabilitation (guest editorial). *IEEE Trans Neural Syst Rehabil Eng*, 20(3):237–238, May 2012.

[2] M. Casadio, V. Sanguineti, P. G. Morasso, and V. Arrichiello. Braccio di ferro: a new haptic workstation for neuromotor rehabilitation. *Technol Health Care*, 14(3):123–42, 2006.

[3] M. A. Conditt, F. Gandolfo, and F. A. Mussa-Ivaldi. The motor system does not learn the dynamics of the arm by rote memorization of past experience. *J Neurophysiol*, 78(1):554–60, Jul 1997.

[4] J. B. Dingwell, C. D. Mah, and F. A. Mussa-Ivaldi. Manipulating objects with internal degrees of freedom: evidence for model-based control. *J Neurophysiol*, 88(1):222–35, Jul 2002.

[5] J. Izawa and R. Shadmehr. Learning from sensory and reward prediction errors during motor adaptation. *Plos Computational Biology*, 7(3), March 2011.

[6] G. Dam, K. Kording, and K. Wei. Credit Assignment during Movement Reinforcement Learning. *PLoS ONE*, 8(2):e55352, February 2013.

[7] M. Kawato and H. Gomi. The cerebellum and VOR/OKR learning models. *Trends in Neuroscience*, 15(11):445?453, November 1992.

[8] J. Krakauer. Motor learning: its relevance to stroke recovery and neurorehabilitation. *Current Opinion in Neurology*, 19:84–90, 2006.

[9] M. Latash, J. P. Scholz, and G. Schöner. Motor control strategies revealed in the structure of motor variability. *Exercise and Sport Sciences Reviews*, 30(1):26–31, October 2001.

[10] X. Liu, K. M. Mosier, F. A. Mussa-Ivaldi, M. Casadio, and R. A. Scheidt. Reorganization of finger coordination patterns during adaptation to rotation and scaling of a newly learned sensorimotor transformation. *J Neurophysiol*, 105(1):454–73, Jan 2011.

[11] K. M. Mosier, R. A. Scheidt, S. Acosta, and F. A. Mussa-Ivaldi. Remapping hand movements in a novel geometrical environment. *J Neurophysiol*, 94(6):4362–72, Dec 2005.

[12] H. Müller and D. Sternad. Decomposition of variability in the execution of goal-oriented tasks: three components of skill improvement. *J Exp Psychol Hum Percept Perform*, 30(1):212–33, Feb 2004.

[13] F. A. Mussa-Ivaldi, M. Casadio, Z. C. Danziger, K. M. Mosier, and R. A. Scheidt. Sensory motor remapping of space in human-machine interfaces. *Prog Brain Res*, 191:45–64, 2011.

[14] R. Ranganathan and K. M. Newell. Influence of motor learning on utilizing path redundancy. *Neurosci lett*, 3(469):416–20, Jan 2010.

[15] R. Shadmehr and F. Mussa-Ivaldi. Sensorimotor adaptation is believed to be driven by the prediction error. *Journal of Neuroscience*, 14:3208–3224, May 1994.

[16] R. Shadmehr, M. A. Smith, and J. W. Krakauer. Error correction, sensory prediction, and adaptation in motor control. *Annu Rev Neurosci*, 33:89–108, 2010.

[17] I. Tamagnone, A. Basteris, and V. Sanguineti. Robot-assisted acquisition of a motor skill: Evolution of performance and effort. In *Biomedical Robotics and Biomechatronics (BioRob), 2012 4th IEEE RAS EMBS International Conference on*, pages 1016 –1021, june 2012.

[18] E. Todorov and M. Jordan. Optimal feedback control as a theory of motor corrdination. *Nature Neuroscience*, 5(11):1226–1235, November 2002.

[19] D. Wolpert, Z. Ghahramani, and J. M.I. An internal model for sensorimotor integration. *Science*, 269(5232):1880–1882, September 1995.