PAPER

Motor adaptation via distributional learning

Journal of Neural Engineering

RECEIVED
21 October 2019

REVISED
3 April 2020

ACCEPTED FOR PUBLICATION
16 July 2020

PUBLISHED
24 May 2021

PAPER

Motor adaptation via distributional learning

Brian A Mitchell[1,*], **Michelle Marneweck**[2,5], **Scott T Grafton**[3] **and Linda R. Petzold**[4]

[1] Department of Computer Science, University of California, Santa Barbara, CA, United States of America
[2] Department of Human Physiology, University of Oregon, Eugene, OR, United States of America
[3] Department of Psychological and Brain Sciences, University of California, Santa Barbara, CA, United States of America
[4] Department of Computer Science, Department of Mechanical Engineering, University of California, Santa Barbara, CA, United States of America
[5] Joint co-first author
[*] Author to whom any correspondence should be addressed.

E-mail: brian_a_mitchell@cs.ucsb.edu

Keywords: neuroscience, functional magnetic resonance imaging, learning, control, reinforcement learning

Abstract

Objective. Both artificial and biological controllers experience errors during learning that are probabilistically distributed. We develop a framework for modeling distributions of errors and relating deviations in these distributions to neural activity. *Approach*. The biological system we consider is a task where human subjects are required to learn to minimize the roll of an inverted T-shaped object with an unbalanced weight (i.e. one side of the object is heavier than the other side) during lift. We also collect BOLD activity during this process. For our experimental setup, we define the state of the system to be the maximum magnitude roll of the object after lift onset and give subjects the goal of achieving the zero state. *Main Results*. We derive a model for this problem from a variant of Temporal Difference Learning. We then combine this model with Distributional Reinforcement Learning (DRL), a framework that involves defining a value distribution by treating the reward as stochastic. This model transforms the goal of the controller from achieving a target state, to achieving a distribution over distances from the target state. We call it a Distributional Temporal Difference Model (DTDM). The DTDM allows us to model errors in unsuccessfully minimizing object roll using deviations in the value distribution when the center of mass of the unbalanced object is changed. We compute deviations in global neural activity and show that they vary continuously with deviations in the value distribution. Different aspects might contribute to this global shift or signal difference, including a difference in grasp and lift force at lift onset, as well as sensory feedback of error/roll after lift onset. We predict that there exists a coordinated, global response to errors that incorporates all of this information, which is encoding the DTDM objective and used on subsequent trials enabling success. We validate the utility of the DTDM as a model for biological adaptation by using it to engineer a robotic controller to solve a similar problem. *Significance*. We develop a novel theoretical framework and show that it can be used to model a non-trivial motor learning task. Because this theoretical framework is consistent with state-of-the-art reinforcement learning, we can also use it to program a robot to perform a similar task. These results suggest a way to model the multiple subsystems composing global neural activity in a way that transfers well to engineering artificial intelligence.

1. Introduction

The human brain is capable of controlling movement to achieve adaptation to a changing environment extremely quickly. This adaptation is much faster and more flexible than controllers engineered by humans can achieve, in part because our understanding of how human motor control works is incomplete. We argue that methods that can be applied to both biological and artificial systems are necessary in order to bridge this gap [15]. In particular, one of the gaps is the lack of models relating the behavioral errors optimized during adaptation with neural activity. Specifically, even after a large amount of training,

behavior is stochastic and the variability of this behavior has been correlated with performance [51–55]. Because of the persistent variability of behavior, feedback rewards and errors, which are functions of behavior, are probabilistically distributed. Motor learning would then best be framed in terms of the optimization of a *distribution* of rewards or errors. And yet, to the best of our knowledge, there are no known approaches for modeling distributions of rewards during motor learning and relating their optimization to neural activity.

One common approach is to reduce the probabilistic nature of observed rewards to a deterministic function by looking at the expected reward [63–66, 68]. We argue in this work that such approaches are not a complete representation of the learning process. For example, some neural systems have been shown to optimize the expected future reward, but others may have different, independent objectives: these may include variance reduction or risk-averse learning which involves optimizing the size of the tail(s) of the distribution over future rewards. Moreover, there is a growing body of work that shows that the brain optimizes a Minimum Free Energy (MFE) objective during learning [4–14]. This objective is equivalent to optimizing the KL-Divergence between error and ideal probability distributions. We contribute results to the body of work on MFE theory by modeling rewards as random variables and proposing that the brain is adapting by minimizing deviations between error and ideal distributions of rewards.

We present behavioral and fMRI BOLD data generated from analysis of 16 subjects, each instructed to minimize the rotation of an unbalanced object at and following its lift. At regular intervals, the center of mass of this object is rotated 180 degrees along its vertical plane, forcing the subjects to adapt their strategy to lift the object while minimizing its roll. In this work, we study adaptation to a changing environment over a series of trials (i.e. a series of attempted lifts). We define the state of the system to be the maximum magnitude of the roll of the object for each trial. Our goal is to model the learning objective that is driving the system to the zero state (the target state). A recent extension of Temporal Difference Learning, called the Temporal Difference Model (TDM) framework, suggests a way to incorporate 'closeness' between the current and a target state into a value function [58]. Specifically, if the negative distance from the current to the target state is used as the reward, then the value function quantifies the expected future proximity to the target state. Stochasticity of the reward function can be modeled using Distributional Reinforcement Learning (DRL), where the reward is modeled as a random variable. We combine the DRL and TDM approaches in this work and refer to the complete model as a Distributional Temporal Difference Model (DTDM). DTDM requires the estimation of a value distribution, rather than a value function, which intuitively corresponds to the distribution over future distances from the target. To fit this distribution, temporal differences between an updated value distribution and a past estimate of the value distribution are used; this is in contrast to classic Temporal Difference Learning which relies on comparisons between value functions. Distributional temporal differences can be interpreted as error signals and we show in this work that the optimization of these errors serves as a good model of motor adaptation.

We treat our experimental set up as a short-time horizon problem where the value distribution models distances between the current and target states at the next trial. We show that the value distribution becomes significantly distorted after a change in the center of mass of the object, a distortion which is quickly corrected after a few trials. This correction involves a shift in the mean of the value distribution, in addition to other changes in the structure of the distribution, including a reduction in variance and a shrinking of the size of its tails. To find a neural basis for all of these different characteristics and potential objectives, we look at the *global* neural activity. We show that the magnitude of the distortion of the value distribution varies continuously with the average deviation in global neural activity, suggesting that the brain is optimizing the distortion in the value distribution during motor adaptation. Further, we show that deviations in global neural activity are directly proportional to those of sensorimotor activity, justifying our choice of representation.

We further validate the utility of the distributional temporal difference by using it to train a robot to perform a similar task, that is, to lift an object with minimal roll and do so while adapting to changes in its center of mass. We use the DTDM to update a model of system dynamics for use in Model Predictive Control (MPC), and as seen in human subjects, our optimization scheme results in exponential improvement of the model, both during initial training and during updating. We show that with this prediction error, the robot is able to quickly update its model and minimize the roll of the object.

2. Materials and Methods

2.1. Summary

In our study of motor adaptation, participants ($N = 16$) performed an object lifting task during fMRI scans that required them to minimize the rotation of the object at and during lift. Subjects had to adapt their strategy to the changing of the object's center of mass at regular intervals. Participants performed 7 runs of 40 trials, where each trial required them to use their thumb and index finger to vertically lift an inverted T-shaped object with an unbalanced center of mass while minimizing its roll at lift onset.

Each trial required the subjects to lift the object 5 cm from a flat surface and subjects were notified when the magnitude of the roll of the object exceeded 5°. Every 10 trials, the object was rotated by 180 degrees, requiring the subjects to change their digit positioning, digit load force, or some combination of the two to achieve task success. For each of the 7 unconstrained runs, subjects were free to change the positioning of their thumb and index finger at will. The position of the thumb, index finger and object (and its roll) was tracked during the course of each trial using a 3-camera motion tracking system. Performance was measured by the absolute maximum magnitude roll generated within 250 ms following lift onset (when the object was lifted 1 mm from the table). To allow the subjects to familiarize themselves with the experiment, the first run of 40 trials was allocated for practice and no BOLD activity was measured. For the final 6 runs, BOLD activity was collected for all subjects during all trials. Whole-brain analysis was conducted to identify brain regions activated during 17 time bins, each being 400 ms long, beginning 1.2 s before lift onset. For each block of 20 trials, blocks of contiguous trials were averaged to yield 7 conditions: pre-rotation conditions containing trials 2–4, 5-7, and 8–10; a rotation condition containing trial 11; and post-rotation conditions containing trials 12–14, 15–17, and 18–20. This was done to smooth over short-time variation between trials. Beta values from whole brain analysis were extracted using the Juelich atlas. The vector of all beta values is what we refer to as 'global neural activity' in this work.

2.2. Participants

Twenty healthy subjects participated in this study (median age: 22 years; range: 18–32; 11 women). They were right-handed and had normal or correct to normal vision. We excluded four subjects as a result of equipment failure ($n = 3$) and not finishing the experiment ($n = 1$). Subjects gave written informed consent and all study procedures were approved by the Human Subjects Committee, Office of Research, University of California–Santa Barbara.

2.3. Materials, Design, and Procedure

Subjects were in supine position in the scanner. Excessive head and body motion was minimized with firm cushion padding of the head, neck, and shoulders. Sandbags under the upper right arm minimized upper limb movement. T1 and T2*-weighted scans were collected followed by BOLD measurements while subjects manipulated a symmetrically-shaped object with a hidden asymmetric mass distribution with the aim of preventing object roll.

Specifications of the custom-made inverted T-shaped object with constrained and unconstrained grasp surfaces along its vertical axis can be found in [3]. In short, the object had a horizontal base and a vertical Plexiglass column. On either side of the vertical column were grip surfaces that were either circular (for constrained contact points) or rectangular (for unconstrained contact points) in shape. A brass block, concealed by covers, was positioned on the horizontal base on either side of the vertical column, creating an asymmetric mass distribution (object torque = 180 Newton millimeter (Nmm)). The total mass of the object was 610 g.

The object was placed at arm's length on a table that was placed over the hips of the subject. The object start position was rotated in a counterclockwise direction at a 30° offset from the edge of the table. This position minimized biomechanical constraints that influence object roll (the wrist would be stiffened more when picking up the object when facing forward rather than angled; the former would minimize the object rolling in a clockwise direction). Subjects were asked to press a button that was in a fixed position toward the right of the object between trials. A mirror attached to the head coil gave continuous viewing of the object and the subject's hand.

Anatomical and fMRI data were collected using a Siemens 3 T Magnetom Prisma Fit (64-channel phased-array head coil). High-resolution 0.94 mm isotropic T1-weighted (TR = 2500 ms, TE = 2.22 ms, FA = 7°, FOV = 241 mm) and T2*-weighted (TR = 3200 ms, TE = 566 ms, FOV = 241 mm) whole-brain sagital sequence images were taken. During object manipulation, BOLD contrast was measured with a multi-band T2*-weighted echoplanar gradient-echo imaging sequence (TR = 400 ms, TE = 35 ms, FA = 52°, FOV = 192 mm, multi-band factor 8). A functional image contained 48 slices acquired parallel to the AC–PC plane (3 mm thick; 3 × 3 mm in-plane resolution).

The position and roll of the object were measured using three motion tracking cameras that were radiofrequency-shielded (Precision Point Tracking System, Worldviz; see [3] for the in-scanner setup). With this system, we recorded positions with six degrees of freedom using near-infrared LEDs (frame rate: 150 Hz; camera resolution: 640 × 480 VGA; at the focal distance, the spatial accuracy is sub-millimeter). An individual LED marker was positioned on either side of the T-shaped object on the outer tip of the aluminum rods (to measure object roll).

2.3.1. Experimental Design and Procedure

The experimental task consisted of four conditions: manipulating the left- and right-weighted object at constrained and unconstrained contact points. Before scanning, subjects completed 40 practice trials to familiarize them with the audio cues instructing when and how to lift the object on a given trial. The 40 trials consisted of 10 blocked trials for each of the 4 conditions (20 trials at unconstrained and 20 trials at constrained grasp contact points). We focus on the data generated from the unconstrained trials in this work.

Each trial began with the subject's hand relaxed on the button. An audio cue instructed subjects to release the button and to reach, grasp, and lift the object to a height marker (5 cm) until the next audio cue (4 s after button-release time) that instructed them to return the object and hand to their respective start positions. The start cue of the first trial was aligned with a functional image. An error cue was given after trial completion if the object roll exceeded 5° at any time during the trial. Stimulus timings for each block of trials were controlled by a custom script (Vizard Virtual Reality Software Toolkit, version 4.0, Worldviz), and the inter-trial interval was randomly chosen to be between 2- 6 s, with a rest period between each of the four blocks of trials. Trial order within a given block was counterbalanced across runs and subjects.

Following practice, BOLD contrast was measured as subjects completed 40 trials in each of 6 functional runs (for a total of 240 trials). For each run's fMRI analyses, we parsed these trials in the following way, giving 7 conditions of interest for unconstrained and constrained conditions, respectively:

1. early pre-rotation trials 2–4
2. mid pre-rotation trials 5–7
3. late pre-rotation trials 8–10
4. rotation trial 11
5. early post-rotation trials 12–14
6. mid post-rotation trials 15–17
7. late post-rotation trials 18–20.

2.4. Kinematic data processing

Kinematic data were filtered using a fourth-order Butterworth filter (cutoff frequency = 5 Hz). We defined object roll as the angle of the object in the frontal plane, with peak object roll extracted shortly after lift onset (250 ms) before somatosensory feedback resulted in corrective responses to counter object roll. Trials with object roll > 5 were classified as errors. Lift onset was defined as the timepoint when the object was lifted 1 mm and remained above this value for at least 20 samples.

2.5. MRI data preprocessing

MRI data were pre-processed and analyzed in SPM12 (Wellcome Trust Center for Neuroimaging, London, UK). Specifically, functional images across all runs were spatially realigned to a mean functional image using 2nd degree B-spline interpolation, which were then co-registered to each subject's structural T1 image. Between-subject spatial normalization steps were conducted with SPM's normalize function aligning each subject's T1 and its co-registered functional images into standard ICBM/MNI-152 atlas space (interpolation: 4th degree B-spline; voxel size: 3x3x3 mm).

We used a deconvolution-based general linear model (GLM) approach to model BOLD activity, with a finite impulse response (FIR) function selected as a basis function (window length: 6.8 s; order: 400 ms), yielding 17 400 ms time bins. Bins 0 and 1 relate to neural activity present before lift onset; lift onset occurs at the start of bin 3. As described above, for each run, we modeled 7 conditions for unconstrained and constrained trials, respectively, with three pre-rotation conditions containing trials 2–4, 5–7, and 8–10; a rotation condition containing trial 11; and three post-rotation conditions containing trials 12–14, 15–17, and 18-20.

Finally, we used the RobustWLS Toolbox in SPM [48] to account for movement artifact by an unbiased estimation of noise variance of each imaging and down-weighting of images with high variance. Nevertheless, head motion mean rotations and translations (with minimum and maximum values in parentheses) were minimal: *x*: −.02 mm (−.38, .34); *y*: -.29 mm (-.86, .29); z: .76 mm (-.42, 1.72); pitch: -.008° (-.02, .008); roll: -.001° (-.009, .006); yaw: .002 ° (-.005, .01).

Before use in estimating the neural deviation, the BOLD values across different ROI's were aggregated into vectors. Given that the task under consideration was a sensorimotor task, it would be natural to restrict the regions under consideration to sensorimotor regions. We show in figure 1 that this is unnecessary, as the deviations generated by sensorimotor regions (vertical axis) are directly proportional to those generated by global activity (horizontal axis). The red lines demonstrate approximate equivalence: the sample deviations cluster about this line for all conditions. The sensorimotor ROI's selected here were the bilateral anterior intraparietal sulcus (AIPS), the Cerebellum, Insula, motor 4a, motor 4p, parietal operculum, primary somatosensory cortex, and superior parietal lobule (SPL). Before deviations were computed, the BOLD vectors were mapped to a lower dimensional space (the space used was ten dimensional). A basis for this space was computed using the Treelet Transform [42] because of its ability to capture sparse, hierarchical structure in covariance matrices.

One explanation for the strong effect of errors on global neural activity is that activity is largely driven by motor brain regions. We address this possibility by considering a seed-based functional connectivity analysis, where the seed regions selected were bilateral motor 4a and motor 4p. This sort of analysis could involve attempting to explain the variance in global activity with variation in the activity of the seed regions. We find that not only can these four regions not explain a significant portion of the global activity, there are no four brain regions that can. In fact, we find that 30-40 regions are required to explain over 50% of the variance in global neural activity. To show this, for each FIR bin, we collected vectors of BOLD activity over all subjects and trials. We then computed a Principal Component Analysis (PCA) and looked at the explained variance ratios for different numbers of principal components. The results of this experiment

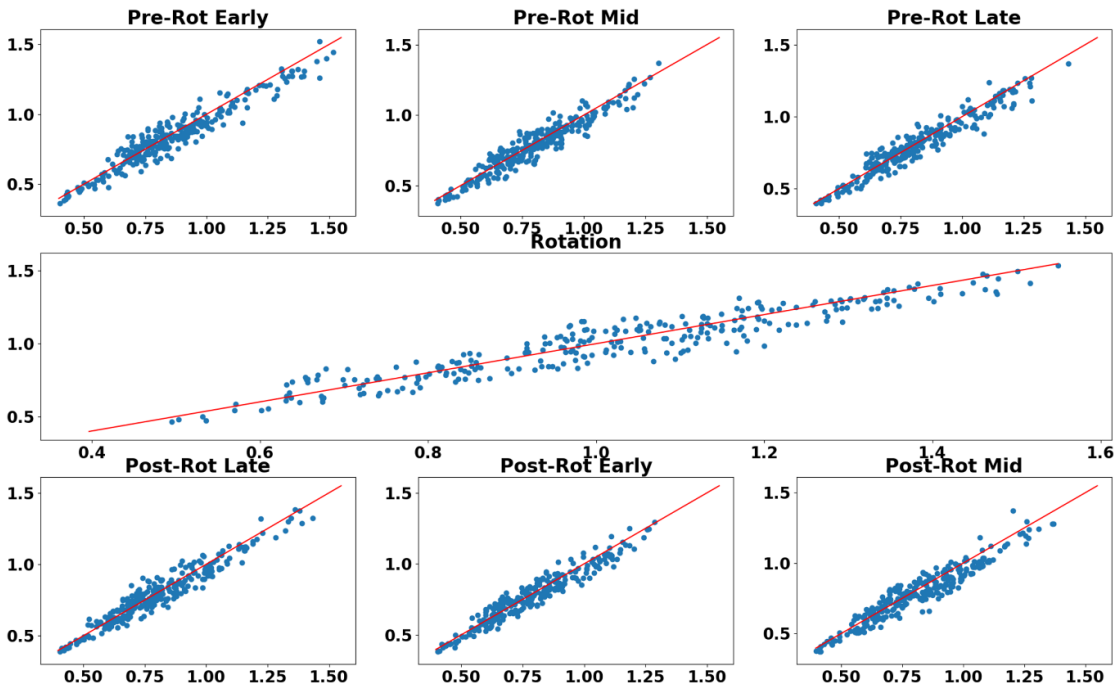


Figure 1. All plots are of sensorimotor deviations (vertical axis on each plot) against global deviations (horizontal axis on each plot). The red line gives perfect equality and the samples cluster about this line. Figure 2 shows that a small number of brain regions cannot be used to explain most of the variance in global neural activity, but this does not preclude the possibility that a small number of brain regions may better encode the DTDM objective. A reasonable candidate for this small set of regions is the set of sensorimotor brain regions, but the plot above shows that sensorimotor and global neural deviations are strongly correlated. This suggests that sensorimotor regions are roughly encoding the same information as the set of all brain regions.

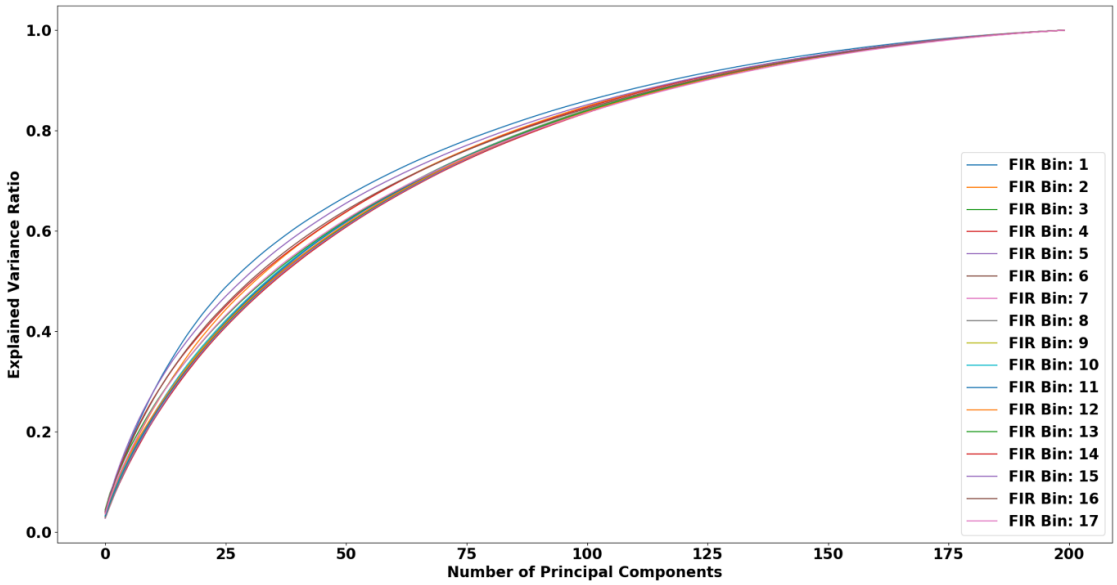


Figure 2. In this plot we show how the explained variance ratio changes in response to increasing the number of principal components in the basis set. A possible explanation for the responsiveness of global activity to errors is that global activity is strongly driven by a small number of brain regions (e.g. primary motor cortex). This plot illustrates the point that 30-40 principal components (depending on the FIR bin) are needed to explain at least 50% of the variance in global neural activity. This shows that there does not exist a small set of brain regions that can explain most of the variance in global activity.

are shown in figure 2: this plot highlights the fact that there does not exist a small group of brain regions that can explain most of the variance in global neural activity.

2.6. Robotic Simulation Details

The OpenAI Gym Pick and Place environment was modified to replicate the experimental task described in this paper. Specifically, the block to be moved was

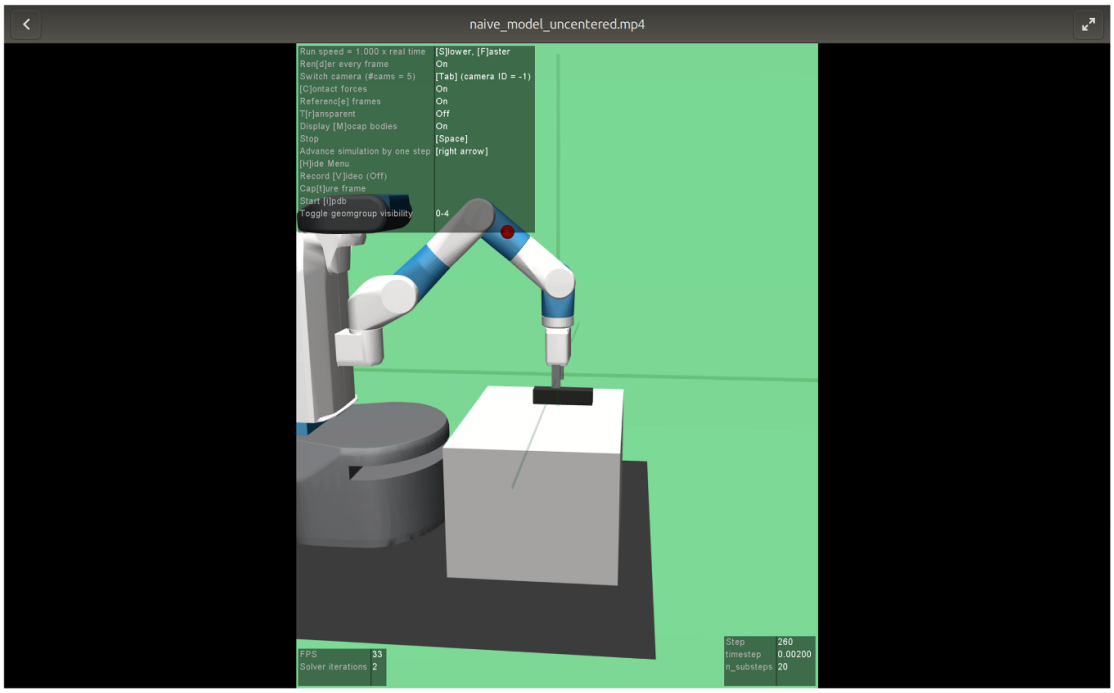


Figure 3. A still frame taken from the robotic simulation. Reproduced from [71]. Copyright (c) 2016 OpenAI (https://openai.com).

extended along a single axis to allow for shifting of the center of mass of the block along this extended axis. Adapting to lift this unbalanced weight with minimal roll along the extended axis would then test the ability of the robot to perform a similar task to that accomplished by the human subjects. Two prior policies were trained using Deep Deterministic Policy Gradients (DDPG) and Hindsight Experience Replay (HER) to lift the block with minimal roll when its center of mass is centered and uncentered, respectively [69]. The parameters used for training were the defaults given in [69]. A single frame taken of the simulation is shown in figure 3.

The dynamics model used for Model Predictive Control (MPC) was a deep neural network with 3 layers and 256 neurons per layer. This network was trained using ADAM with learning rate 0.001 and batch size 256 [70]. Mini-batches were sampled from a uniform distribution over elements of the replay buffer, which had a maximum size of 1×10^{-6} elements. A zero'th order policy optimization scheme was used within the MPC framework. For this optimization scheme, 500 rollouts were used, each of length 15 timesteps.

3. Results

3.1. Errors During Motor Learning are Probabilistically Distributed

First, we examine behavioral performance during the adaptation task. Across all participants, trials and conditions, the maximum magnitude roll over the course of a trial, averaged over all subjects and runs, was observed to be low for pre-rotation conditions, high for rotation conditions, and low again for post-rotation conditions. This point is illustrated in figure 4, where we show the distributions over states for all conditions. These plots not only make clear the presence of errors and the fact that they are quickly corrected, but also that the distributions over states contain meaningful information that would be lost by considering only the mean. For example, the distribution generated by the rotation condition has a different shape from any of those generated from the pre-/post-rotation conditions (pre/post vs rot, $A^2 = 89.21\,431$, $p<0.01$; throughout this work, pre/post refers to the combination of pre-rotation and post-rotation samples and rot refers to rotation samples). This result holds after bootstrap resampling of samples to correct for sample-size differences between pre-rotation/post-rotation conditions and the rotation condition. Plots of the resampled histograms are shown in the second row of figure 4. The resampled histograms were generated by resampling pre-rotation and post-rotation samples uniformly at random to generate sample sizes equal to that of the rotation condition.

Analysis of these results requires a representation of the error that takes into account the observed distributional information. We observed that these distributions over distances follow a Weibull distribution

$$p(d_p;\gamma,\beta) = \frac{\gamma}{\beta}\left(\frac{d_p}{\beta}\right)^{\gamma-1} e^{-\left(\frac{d_p}{\beta}\right)^{\gamma}}, \qquad (1)$$

where d_p is the L_p distance and γ and β are parameters. This provides a convenient, closed-form

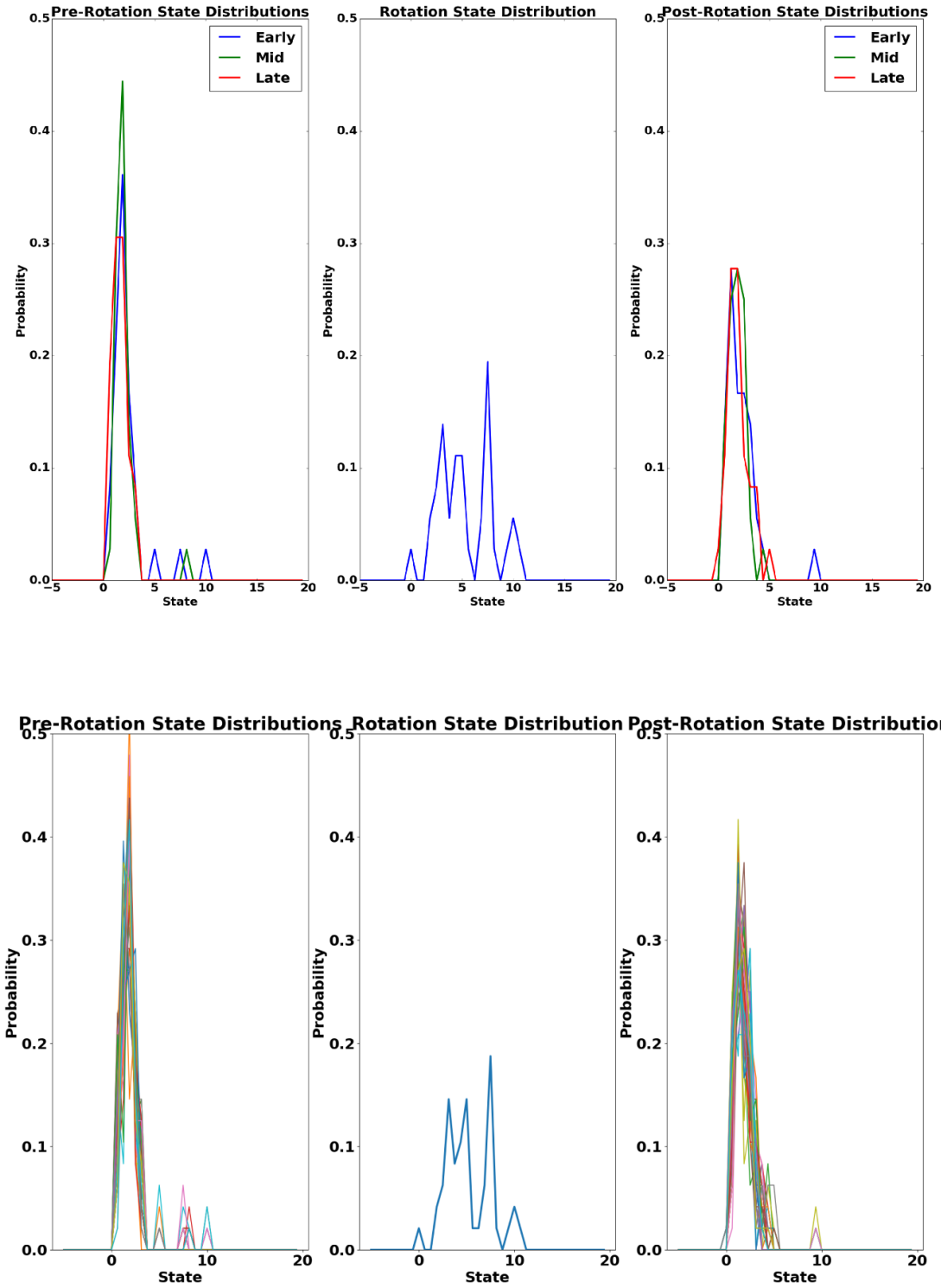


Figure 4. Top: state distributions. The rotation condition results in a state distribution that has both higher mean and an entirely different shape. **Bottom**: bootstrap resampling of histograms to yield balanced sample sizes across pre-rotation, rotation, and post-rotation conditions results. The difference in the rotation histogram compared with the pre-rotation and post-rotation histograms is preserved even in the case of bootstrap resampling.

mathematical representation for errors that we revisit throughout this work. To validate that our distances are indeed Weibull distributed, consider first the necessary and sufficient conditions for distances between feature vectors to be Weibull distributed. Given feature vectors $X = [X_1, \ldots, X_n] \in \mathbb{R}^n$ and $Y = [Y_1, \ldots, Y_n] \in \mathbb{R}^n$, the L_p distance between X and Y is Weibull distributed if $|X_i - Y_i|^p$ are non-identical, correlated, and upper bounded, for all $1 \leq i \leq n$. Rather than construct a mathematical proof that these assumptions hold for human movement, we instead demonstrate that Weibull distributions can be successfully fit to our data.

In figure 5 we show that the empirical distributions over distances resulting from comparing pre/-post samples with rotation samples (called rotation

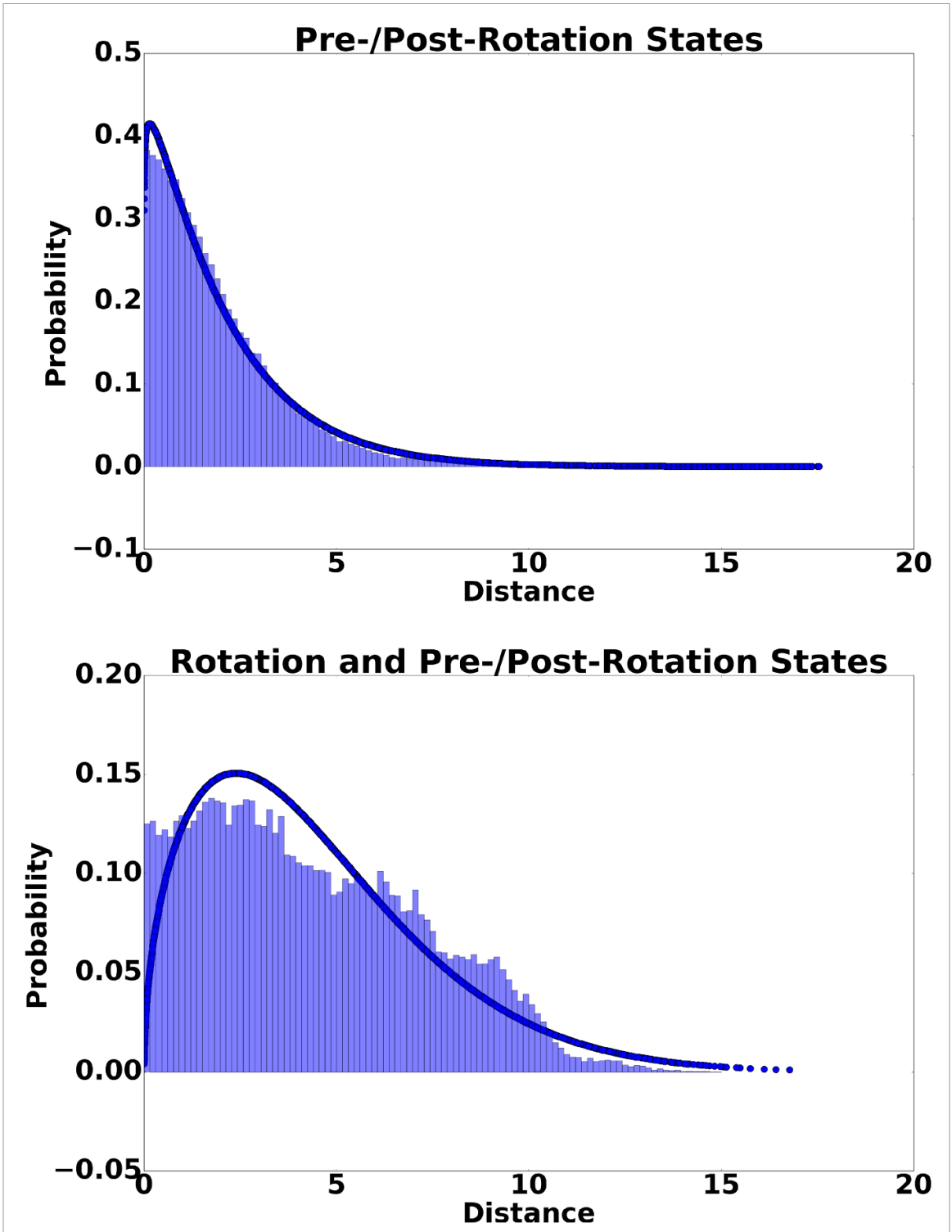


Figure 5. The empirical density estimates are presented as histograms and the Weibull fits are superimposed and given by the blue dots. **Top**: distances are between pre-/post-rotation conditions. **Bottom**: distances are between rotation and pre-/post-rotation conditions.

or rot) differ significantly from the empirical distributions generated by comparing pre/post samples with other pre/post samples (pre/post-rot vs pre/post-pre/post, $A^2 = 60.66\,115$, $p<0.01$). Moreover, fitting Weibull distributions to these empirical distributions using Maximum Likelihood Estimation (MLE), we are able to generate accurate fits, suggesting that the Weibull is indeed a good model for these data (pre/post empirical vs pre/post Weibull, $\beta = 2.047$, $\gamma = 1.062$, $A^2 = 1.184\,024$, $p>0.2$; rot empirical vs rot Weibull, $\beta = 4.947$, $\gamma = 1.504$, $A^2 = 1.72\,791$, $p>0.2$). We call the pre/post-rot Weibull the error Weibull (W_e) and we call the pre/post-pre/post Weibull the ideal Weibull (W_i). As subjects adapt and W_e is transformed back to W_i, a number of characteristics of W_e change: its mean shifts towards 0, its long tail becomes reduced in size, its variance shrinks, and its skew decreases. From

this, it seems as if some notion of the deviation between W_e and W_i would have to be used as feedback to a controller in order to incorporate all of this information.

3.2. A Distributional model for prediction errors

A model of learning that relies on the deviation between W_e and W_i can be derived from a Temporal Difference Model (TDM) [58], which can in turn be derived from a Temporal Difference Learning (TDL) update. TDL is a recursive scheme to maximize expected future rewards and requires the definition of a value function, $V(s_t|\pi)$, where s_t is the state at time t and π is a policy. The value function can be defined as

$$V(s_t|\pi) = \mathrm{E}_{p(s_{t+1}|s_t,a_t),\pi}[r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \ldots], \tag{2}$$

$$V(s_t|\pi) = \mathrm{E}_{p(s_{t+1}|s_t,a_t),\pi}[r_t + \gamma V(s_{t+1}|\pi)], \tag{3}$$

where $\gamma \in [0, 1)$ is a discount factor, $p(s_{t+1}|s_t, a_t)$ is a model of the system dynamics, $\mathrm{E}[\cdot]$ is the expectation operator, and r_t is the reward at time t. Perhaps the simplest approach to fitting $V(s_t|\pi)$ using TDL, called TD(0), relies on computing an estimator $\hat{V}(s_t|\pi)$ using the update equation

$$\hat{V}(s_t|\pi) = \hat{V}(s_t|\pi) + \alpha[r_t + \gamma \hat{V}(s_{t+1}|\pi) - \hat{V}(s_t|\pi)], \tag{4}$$

where $\alpha \in \mathbb{R}$ is the learning rate. This update involves a comparison between $r_t + \gamma \hat{V}(s_{t+1}|\pi)$ and $\hat{V}(s_t|\pi)$. The intuition for this update is, since the former has slightly more information from the environment than $\hat{V}(s_t|\pi)$, $\hat{V}(s_t|\pi)$ should be updated to be closer to it. TDM's define a reward function using the notion of a goal state, s_g, where $r_t = r(s_t, a_t, s_{t+1}, s_g) = -d_p(s_{t+1}, s_g)$ and d_p is the L_p distance. This reward results in a value function that quantifies the expected future proximity of the system to the goal state. As applied to our experimental system, if we let t and $t+1$ be trial numbers, since the goal state is a roll of zero, $V(s_t|\pi)$ would then indicate the expected magnitude of the roll over future trials.

TDM's rely on a kind of reward prediction error to update the value function, and can also act as a bridge between state prediction and reward prediction errors. To be clear, state prediction error is the error in predicting the next state given the current state, and reward prediction error is the ability to predict the future reward given the current state. These errors, when applied to our system, quantify the ability to predict future rolls in expectation. This is shown in the Supplement, where we give conditions for the equivalence of state and reward prediction in the TDM framework. In our experimental system, we are not simply interested in defining V using an expectation over $p(s_{t+1}|s_t, a_t)$ and π. We would like to be able to use all of the information contained in the distribution over rewards. To this end, we incorporate TDM's into the Distributional RL framework

$$Z(s_t, a_t, s_g) \overset{D}{=} R(s_t, a_t, s_{t+1}, s_g) + \gamma Z(s_{t+1}, a_{t+1}, s_g), \tag{5}$$

where Z and R are the value and reward distributions, respectively, and $\overset{D}{=}$ indicates equality in distribution [50]. Similar to TD(0), Distributional RL updates an estimator of the value distribution, $\hat{Z}(s_t, a_t, s_g)$, by comparing $R(s_t, a_t, s_{t+1}, s_g) + \gamma \hat{Z}(s_{t+1}, a_{t+1}, s_g)$ with $\hat{Z}(s_t, a_t, s_g)$. Because these are probability distributions, $\hat{Z}(s_t, a_t, s_g)$ is updated to minimize

$$D_{KL}(R(s_t, a_t, s_{t+1}, s_g) + \gamma \hat{Z}(s_{t+1}, a_{t+1}, s_g)||\hat{Z}(s_t, a_t, s_g)), \tag{6}$$

where $D_{KL}(\cdot||\cdot)$ is the KL-divergence. This update is analogous to the temporal difference learning update, generalized to the setting where rewards are probabilistically distributed. This distributional objective, with $R(s_t, a_t, s_{t+1}, s_g)$ defined as the distribution over $-d_p(s_{t+1}, s_g)$, is relevant in the context of the results presented thus far. Specifically, in the case of short-time horizon problems, those where $\gamma = 0$, then $Z(s_t, a_t, s_g) \overset{D}{=} R(s_t, a_t, s_{t+1}, s_g)$ follows a Weibull distribution.

Keeping with the notation of the previous section, we can think of $\hat{Z}(s_t, a_t, s_g)$ as being equivalent to W_i during the pre-rotation conditions. When the center of mass changes, $Z(s_t, a_t, s_g)$ is actually W_e, though $\hat{Z}(s_t, a_t, s_g)$ is still W_h. The deviation between $\hat{Z}(s_t, a_t, s_g)$ and $Z(s_t, a_t, s_g)$, that is, W_e and W_i, is optimized during adaptation. For the experimental system studied in this work, there are a number of potential explanations for this deviation, from errors in the model of system dynamics to errors in the behavioral policy. The identification of the precise source of the deviation between $\hat{Z}(s_t, a_t, s_g)$ and $Z(s_t, a_t, s_g)$ is beyond the scope of this work. Our goal is to present a framework for modeling learning with stochastic rewards in a manner amenable to both biological modeling and robotic control. With this in mind, we note that $\hat{Z}(s_t, a_t, s_g)$ may be parameterized by θ, which includes parameters for every component of the controller used to solve the unbalanced lifting task. We can now propose a model for motor learning, specifically, a model for learning to dynamically update a controller to lift an object in response to its changing physical properties. Our model is that the brain attempts to solve the following optimization problem

$$\underset{\theta}{\text{minimize}}\ D_{KL}(W_i||W_e), \tag{7}$$

To the best of our knowledge, the fusion of temporal difference learning and distributional reinforcement learning is a novel theoretical framework for reinforcement learning. The optimization problem above is a special case of the full DTDM optimization, but throughout the rest of the paper, when we refer to the DTDM problem, we are referring to equation (7).

3.3. Global neural activity optimizes the distributional temporal difference objective

We have already shown that behavior is updated to optimize the deviation between W_e and W_i, that is, behavior is updated according to equation (7). To see the effect of the object rotation condition on global neural activity, we first processed brain activity in consecutive time intervals using finite impulse response (FIR) modeling. We then selected FIR time bins that are likely encoding information about the lift of the apparatus. Details of the method used to select the 'lift' FIR bins are given in the Supplement. Briefly, we first identify 'pre-lift' FIR bins as those before lift onset: this occurs at FIR bin 3. We then interpret the hemodynamic response as a stochastic process and note that there are two distinct stimuli within each trial: the pre-lift and lift stimuli. Given that these stimuli are separated in time, their respective hemodynamic responses will peak at different times. This allows for the segmentation of the FIR bins as most likely generated from either the pre-lift or the lift process. Those most likely generated from the lift process (bins 15-16) are called 'lift' bins and are used to estimate the deviation of global neural activity resulting from lift. These bins are identified using a hard threshold based upon a model of the hemodynamic response (i.e. the Canonical Hemodynamic Response Function, CHRF) [2]. We interpret the CHRF as a mixture of Gamma Distributions. Using two CHRF's (one corresponding to pre-lift and one corresponding to post-lift), we are able to segment the FIR bins as most likely exhibiting BOLD activity from pre-lift or post-lift. Further details on this method are given in the Supplement.

These results are shown in figure 6. For each condition (pre-rotation early/mid/late, rotation, post-rotation early/mid/late), Weibull distributions were generated by comparing the betas generated during that condition with the betas generated during all others. Example Weibull distributions generated during FIR bin 0 and FIR bin 15 are shown in the left and right columns of the top two rows of figure 6. The distribution generated using the rotation condition exhibits a significant deviation from the others at FIR bin 15 but not FIR bin 0. Because pre/post-pre/post and pre/post-rot Weibull distributions are statistically different for lift bins but not pre-lift bins (pre/post-rot vs pre/post-pre/post for pre-lift bins, t(df) = -1.572 965, p>0.2; pre/post-rot vs pre/post-pre/post for lift bins, t(df) = -8.73 572, p<0.01), this suggests that global neural activity is perturbed by the rotation condition, and then moves back to become indistinguishable from the pre-rotation state. Thus we call the pre/post-pre/post Weibulls 'ideal beta Weibull distributions' or W_i^b and the pre/post-rot Weibull the 'error beta Weibull' or W_e^b.

Our results suggest that the brain may be sensitive to $D_{KL}(W_i||W_e)$. In figure 7 (bottom row) we show that the difference in the means of W_i^b and W_e^b (using lift FIR bins) is directly proportional to the deviation between W_i and W_e (i.e. $D_{KL}(W_i||W_e)$; $R^2 = 0.55$). We show in figure 7 (left, middle row) that global neural activity is also directly proportional to the TDM error, that is, errors in expected future reward ($R^2 = 0.44$). To understand this result, we present histograms estimating W_i and W_e from two representative subjects. The transport of W_e to W_i involves more than just a shift in the mean for both subjects, but for both (and for all other subjects as well), the mean is indeed shifted during adaptation.

It is important to note that TDM does not contain a complete description of the errors. To see this quantitatively, we use a Conditional Value at Risk (CVAR) model [16]. CVAR models offer a means of taking advantage of the information contained in the value distribution, beyond its mean. These models involve optimizing the expected value in the tails of the value distribution. For example, minimizing lower tail values results in controllers that are risk averse. Risk aversion in our experimental system would involve minimizing the use of actions leading to outcomes in the tail of W_i. For example, suppose that subjects initially used lifting strategies that sometimes led to states near zero (extremely successful outcomes), but also often led to the apparatus being dropped, resulting in high roll. A risk averse learning process would avoid this strategy, leading to fewer observations in the tail of W_i. Interestingly, because this may also reduce the observation of as many low roll states, the mean of W_e may be unaffected by risk averse learning. We show in figure 7 that the CVAR error (i.e. the expected lower-tail value) is also proportional to mean neural deviation ($R^2 = 0.46$). Because CVAR error is a characteristic of the value distribution and is independent of the means of W_i and W_e, this suggests that the global neural deviation is, in fact, also encoding more than just the expected future reward.

The error $D_{KL}(W_i||W_e)$ can be interpreted in a number of ways since different aspects of neural activity could contribute to this shift. Sensory activity as well as error signaling could contribute to such a shift. In addition, compensatory behaviors were also observed during the course of a lift. When a subject perceived a tilt, they would often attempt to change the forces and torques used during the course of the lift, often resulting in reduced roll. We hypothesize that there exists a coordinated, global response to

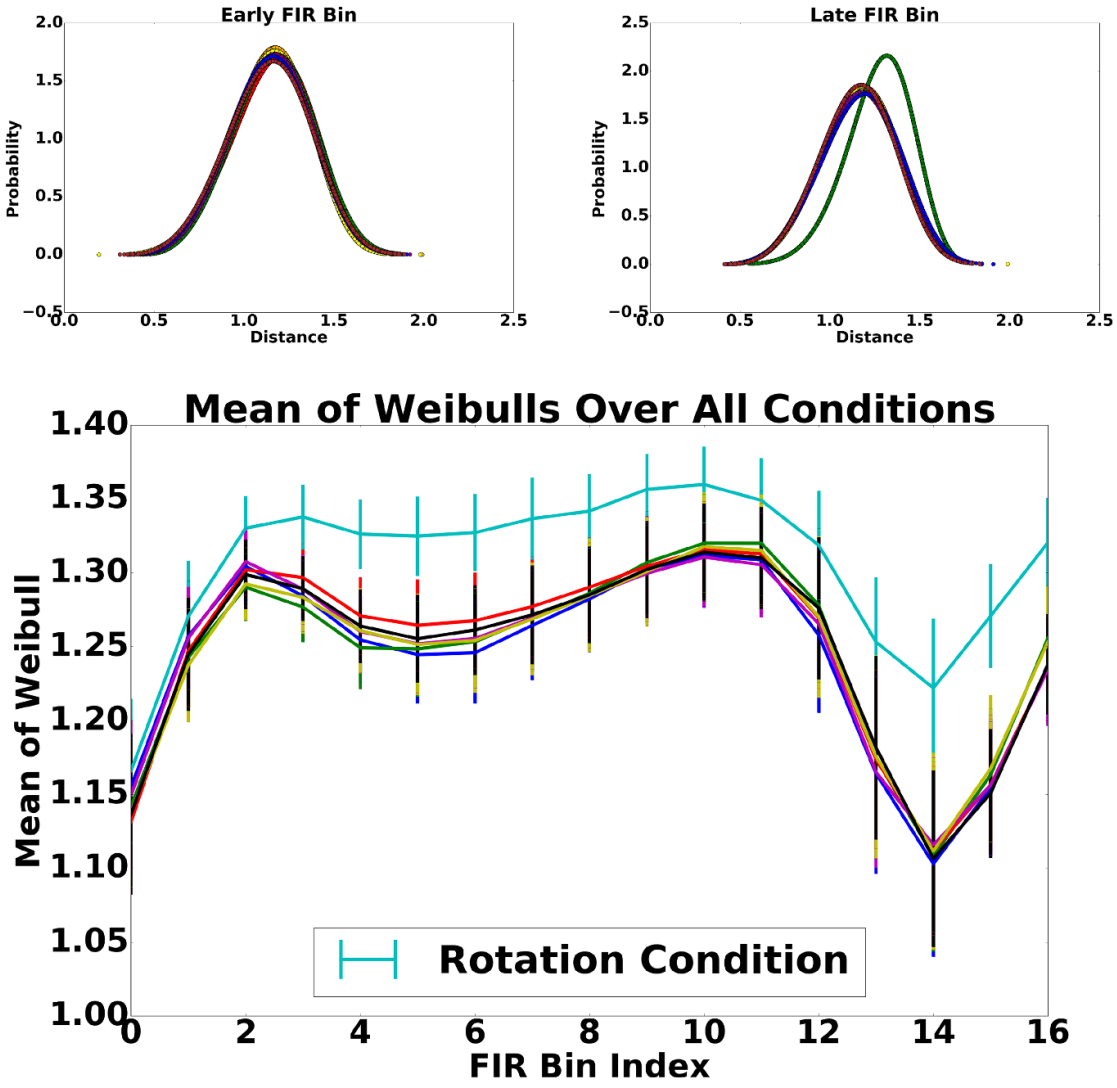


Figure 6. The **top** row contains the fits generated using FIR bins 0 and 15. As in the rest of this work, the distance is the L2 norm of the difference between global neural activities. This is a unit-less measure of deviation between feature vectors. W_i^b becomes distinguishable from W_e^b around FIR bin 15. This point is illustrated in the **middle** row. This row contains the estimates of the means of the Weibull distribution for each condition against the FIR bin index. The significant deviation of W_i^b from W_e^b for late FIR bins is captured by these plots.

errors that incorporates all of this information and that it is proportional to $D_{KL}(W_i||W_e)$. To show that the global shift in neural activity can be directly used as a feedback error signal, we use $D_{KL}(W_i||W_e)$ to fit a robotic controller.

3.4. Robots can also optimize the distributional temporal difference objective

Conveniently, the optimization problem in equation (7) leads to a form that can be optimized by an artificial agent. To see this, we consider an optimization problem similar to those used to update models of system dynamics for use in Model-Based RL. A popular objective for fitting a model of system dynamics is

$$\underset{\theta}{\text{minimize}} ||s^*_{t+1} - f_\theta(s_t, a_t)||_2^2, \tag{8}$$

where f_θ is a model parameterized by θ, s^*_{t+1} is the true state at time $t+1$, and $f_\theta(s_t, a_t)$ is the predicted state at $t+1$. On its face, it may not be obvious how equation (8) is related to equation (7). The latter involves fitting W_e, which is a distribution over distances between the current and target states, while the former involves comparing predicted and actual states at time $t+1$. To see the connection, consider the fact that if f_θ is a probabilistic model, even if its performance is optimized via equation (8) (with some steps taken to preserve non-zero variance), the distances $||s^*_{t+1} - f_\theta(s_t, a_t)||_2$ will be Weibull distributed. We can think of this Weibull as W_i. In the case where the environment changes and the state at time $t+1$ is no longer s^*_{t+1} but instead s'_{t+1}, the performance of f_θ is no longer measured by $||s^*_{t+1} - f_\theta(s_t, a_t)||_2$. Instead, $||s'_{t+1} - f_\theta(s_t, a_t)||_2$ is used. The distribution over these new distances is no longer

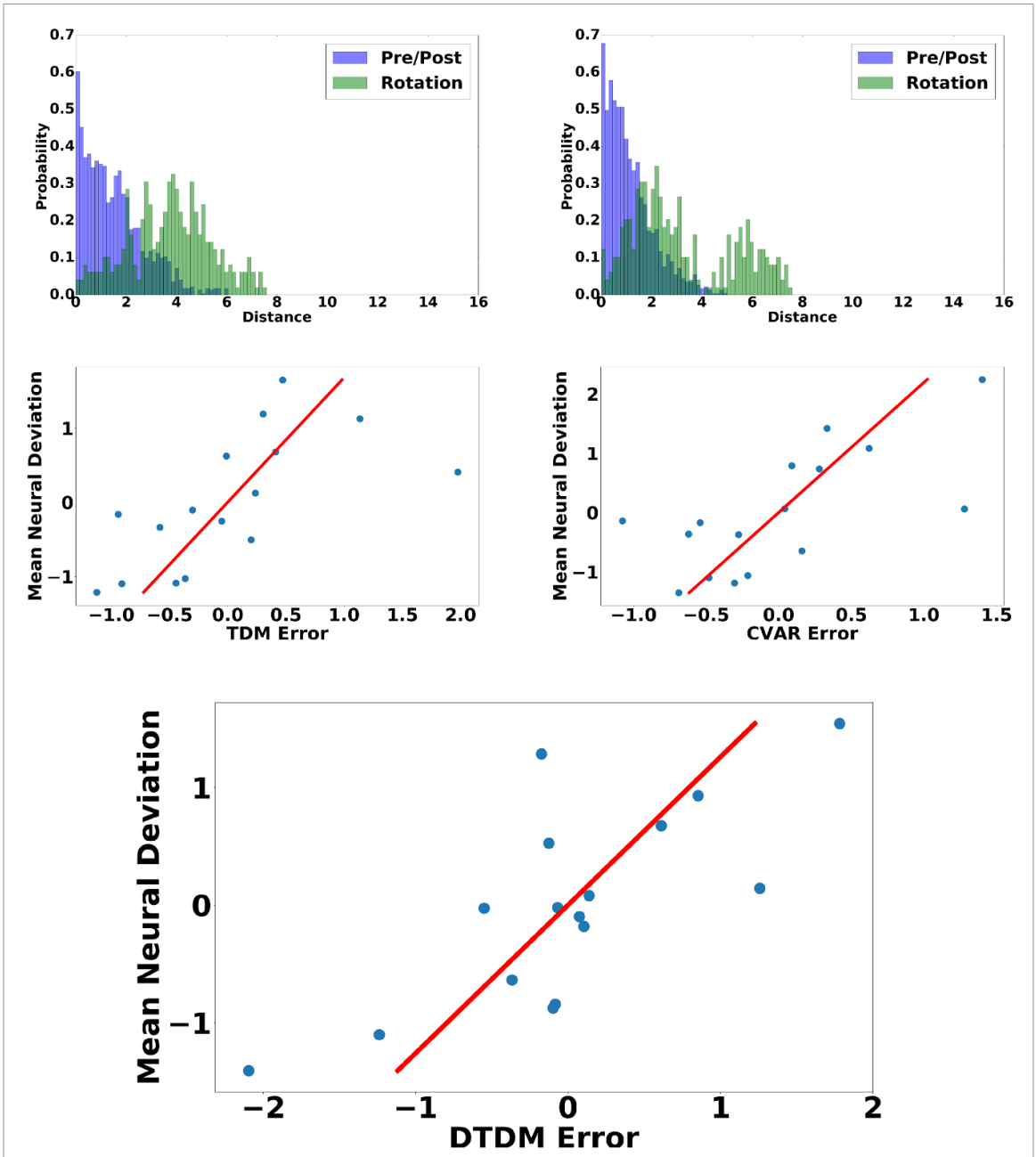
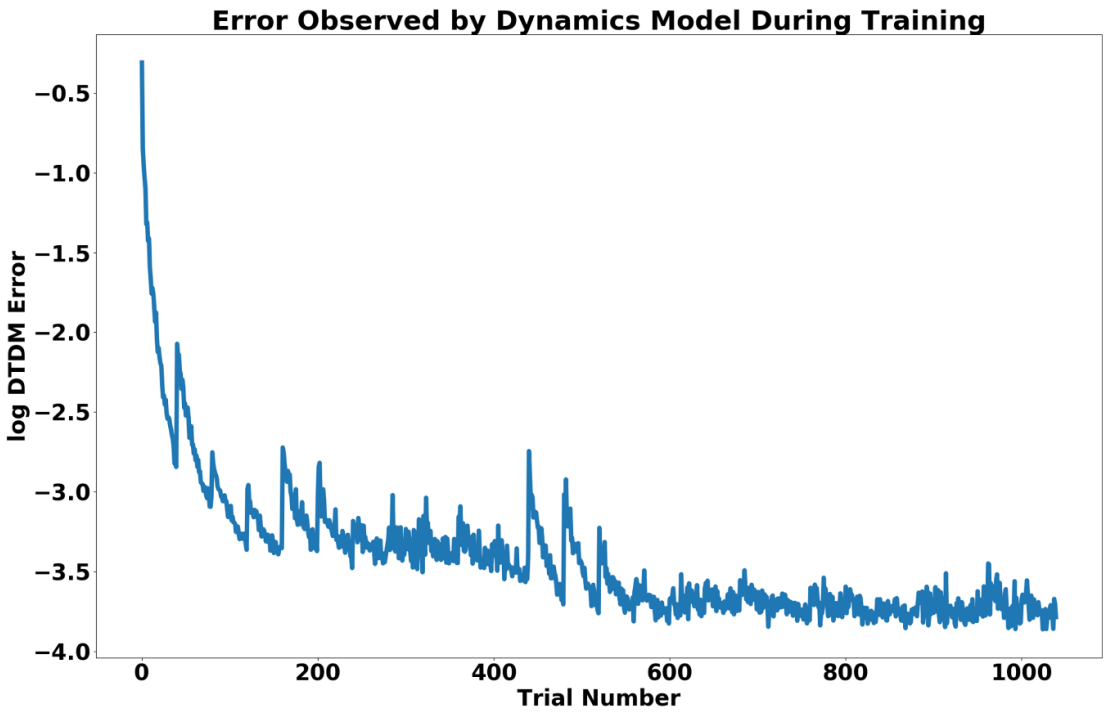


Figure 7. The **top** row contains histograms of pre-/post-rotation and pre-/post-rotation with rotation state distances generated from two subjects. Modeling the movement of the mean of the rotation distribution is not sufficient to completely characterize the learning objective. The **middle** row relates TDM and CVAR errors with the mean beta deviation, where each point is a subject. The **bottom** plot illustrates the relationship between the DTDM error and the mean beta deviation. Lines of best fit are shown in red and are generated using the RANSAC algorithm because of its robustness to outliers. Arguably, the mean neural deviation is encoding both TDM and CVAR errors (as well as other relationships between W_i and W_e).

W_i, and we call this new Weibull W_e. Updating the dynamics model using equation (7) would then amount to bringing the predictions of $f_\theta(s_t, a_t)$ as close to s'_{t+1} as they had been to s^*_{t+1} before the environment changed. We incorporate equation (7) into a model-based RL approach. We use this model-based framework to allow a simulated robotic arm to learn to lift a block when the location of its center of mass is periodically shifted.

The controller we use assumes the existence of two stochastic policies: one that is capable of lifting an object with a centered center of mass and another that is capable of lifting an object with an unbalanced center of mass. We make this assumption because in learning to adapt to a shifting center of mass, the human subjects in our experiment already know how to lift the object in both orientations. The task is assessing their ability to adapt, thus this is the focus of our robotic experiment as well. At time t of the simulation, R possible actions are sampled from the policies. Rollouts from these actions are simulated forward in time to $t + T$ using a dynamics model and the policies. This results in R state-action trajectories of length T. These trajectories



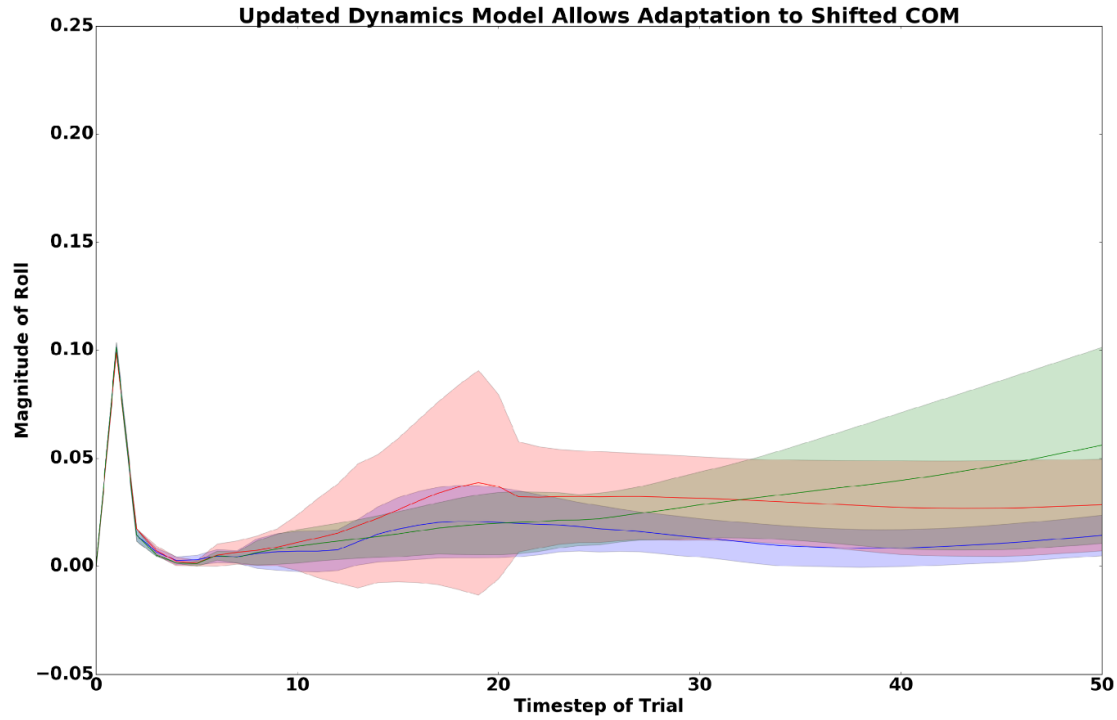


Figure 8. Top: Error of the dynamics model against the number of iterations of training algorithm. Training was performed using mini-batch sampling, which explains the saw-tooth improvement in the error. The center of mass is switched half-way through training, resulting in a jump in the error. This jump is corrected during subsequent updating. **Bottom**: The **red** curves are generated by the controller fit using a centered weight with dynamics trained on a centered weight, the **green** curves are generated using an uncentered weight with a dynamics model trained on a centered weight and not updated, and the **blue** curves are generated using an uncentered weight with an updated model. The updated model is able to outperform the controller without a model update.

are compared using the cumulative reward over all T timesteps, $\sum_{h=t}^{t+T} c(s_h^i, a_h^i)$, where $i \in \{1, \ldots, R\}$ and $c(s_h^i, a_h^i)$ is the absolute value of the roll of the object at time h. The action at time t yielding the lowest cost trajectory is the one selected and this process is repeated for each timestep.

The results of this experiment are shown in figure 8. The top plot shows the error generated by

the dynamics model with respect to the trial number. Shortly after trial 400, the center of mass is switched, causing a spike in the error. Within about 50 trials, the model has adapted and its performance has improved to be better than it was before the switch. The bottom plot shows the performance of the controller as measured by the absolute value of the roll over the course of the trial. The results show that the robot is able to adapt quickly to the changing center of mass, albeit not as quickly as a human. The robot is able to adapt in a little over 100 trials, while the human is able to adapt within 1-2 trials.

There are many possible sources of inefficiency for the robotic controller that could explain this performance gap. First, the dynamics model is updated using random batch sampling from past experience. Arguably, humans do not randomly sample from all past experiences with the object when faced with sub-optimal performance. They are able to draw from past experience based upon hypotheses as to the cause of the errors. Next, the dynamics model is relatively simple and contains no prior knowledge before training about how such objects behave. The human mind contains an enormous amount of past experience to draw on to generate hypotheses explaining errors. The representation of the object in the human mind is also much higher dimensional than the representation used by the robot, containing tactile, visual, and auditory information. Yet, while the performance of the robotic controller is not at the level of the human, these experiments demonstrate that the DTDM objective can actually be used to solve a control problem that is similar to the one solved by humans. Videos demonstrating the robot's performance can be found at (Naive model: https://youtu.be/Amm5ziMSv7U; Updated model: https://youtu.be/2zqlX-2TvCU).

4. Discussion

We have proposed a distributional learning objective for use during motor control and used this representation to construct a model of motor learning. To so do, we extended Temporal Difference Models to Distributional Temporal Difference Models. We have shown that behavior appears to optimize this distributional objective and that deviations in global neural activity are proportional to the magnitude of the distortion of the value distribution. DTDM is not simply useful as a model of motor learning. We have shown that it can be incorporated into a robotic controller and used for engineering applications. The strong connection implied between neural and robotic systems suggests that improved understanding of the brain can be directly used to improve robotic engineering. Our work also suggests that work exploring the converse claim may be successful as well. This claim is often made indirectly, for example, by citing the neuroscientific origins of machine learning, though there is currently no formal framework for extracting neuroscientific principles for the purpose of engineering AI [15]. We hope that this work will be a step in this direction.

Our results also contribute to the accumulating body of evidence in support of the Minimum Free Energy (MFE) theory of neural learning [4–14]. Our results concern motor learning, while the MFE theory is posited to apply to neural learning in general. This theory posits that learning proceeds through the optimization of a free energy of the form $\mathrm{E}_{\rho}[V] + \mathrm{H}[\rho]$, where V is a potential function, $\mathrm{E}[\cdot]$ is the expectation operator, and ρ is a probability measure. Interestingly, optimization of this free energy is equivalent to optimization of $\mathrm{D}_{KL}(\rho||e^{-V})$ [4]. In this work, we have not explored the extent to which W_i can be approximated by a measure of the form e^{-V}: this information would allow for equation (7) to be directly related to a free energy functional. This may be an interesting direction for future work. The connection between this work and MFE theory raises questions related to the neural origins of the DTDM objective. Specifically, is this mathematical objective explicitly encoded by some population of neurons? Previous work has shown that populations of neurons are capable of performing Bayesian inference and both MFE learning and optimization of the DTDM objective are examples of variational Bayesian inference [1]. Further work is necessary to explicitly demonstrate the mechanism of how neurons are capable of encoding the DTDM objective.

We have proposed a distributional framework for motor learning, but have not explored in depth how different aspects of the value distribution could be used during motor adaptation. By extending TDM's to the setting where policies are able to optimize the structure of value distributions, we allow for explicit modeling of properties of learning that cannot be captured by TDM's alone. One such property of human learning is risk aversion, where humans not only learn to maximize reward, they also learn to avoid poor performance (in the case of this work, low reward). In figure 5, we show that W_e has a much longer tail than W_i in addition to having a higher mean, and that the shape of the tail of W_e is also optimized during learning. This is caused by a reduction in the relative number of large rolls (shrinking the size of the tails of W_e), which is a form of risk aversion. Using the difference in the means of W_e and W_i obscures this information, despite the fact that it is useful in a number of different settings. Certainly, there are situations in which risk-averse behavior is best and situations where it results in overly cautious behavior. By maintaining a representation of the value distribution, the brain is able to generate policies by optimizing different aspects of the distribution. These policies can be selected from, to produce behavior that is appropriately cautious for a given situation. The issue of selecting from amongst a population of possible actions is

interesting in the context of DTDM for other reasons as well. Often, the representation of errors used in control problems and in modeling neural controllers is subordinate to the type of controller used, for example, either model-based or model-free. This work suggests that from both a neurological and an engineering standpoint, this manner of thinking may be reversed. Specifically, it may be better to develop a representation of errors that can be used for either model-free or model-based control, and then develop a controller that can best optimize this error in the system of interest. In the context of neurological systems, this suggests the existence of a generic error encoding that are independent of the class of controller. The utility of such a generic error representation would facilitate, for example, action selection in the setting where a number of candidate actions must be selected from and the candidate actions are generated from both model-free and model-based systems [17–21]. In this setting, a generic representation of error would allow for a universal way of comparing the performance of controllers and selecting actions.

Acknowledgments

This project was funded from the Institute for Collaborative Biotechnologies through grant W911NF-19-2-0026 from the U.S. Army Research Office the U.S. Army Research Laboratory, Grant W911NF-16-1-0474 from the Army Research Office, mission funding to the U.S. Army Research Laboratory, and the National Health and Medical Research Council (CJ Martin Biomedical Fellowship, GNT1110090 to MM). The views, opinions, and/or findings expressed are those of the author(s) and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government.

ORCID iD

Brian A Mitchell https://orcid.org/0000-0003-0870-0434

References

[1] Sohn H, Narain D, Meirhaeghe N and Jazayeri M 2019 Bayesian Computation through cortical latent dynamics *Neuron* **103** 934–47
[2] Lindquist M A, Loh J M, Atlas L Y and Wager T D 2008 Modeling the hemodynamic response function in fMRI: efficiency, bias and mis-modeling *Neuroimage* **45** S187–98
[3] Marneweck M, Barany D A, Santello M and Grafton S T 2019 Neural representations of sensorimotor memory- and digit position-based load force adjustments before the onset of dexterous object manipulation *J Neurosci* **38** 4724–37
[4] Mitchell B A, Lauharatanahirun N, Garcia J O, Wymbs N, Grafton S T, Vettel J M and Petzold L R 2019 A minimum free energy model of motor learning *Neural Comput.* **31** 1945–63
[5] Adams R A, Shipp S and Friston K J 2013 Predictions not commands: active inference in the motor system *Brain Struct Funct* **218** 611–43
[6] Braun D A, Ortega P A, Theodorou E and Schaal S 2011 Path integral control and bounded rationality *Adaptive Dynamic Programming And Reinforcement Learning* 202–9
[7] Friston K 2010 The free-energy principle: a unified brain theory? *Nat Rev Neurosci* **11** 127–38
[8] Friston K, Mattout J and Kilner J 2011 Action understanding and active inference *Biol Cybern* **104** 137–60
[9] Friston K, Samothrakis S and Montague R 2012 Active inference and agency: optimal control without cost functions *Biol. Cybernetics* **106** 523–41
[10] Kappen H J, Gomez Y and Opper M 2012 Optimal control as a graphical model inference problem *Mach. Learn.* **87** 159–82
[11] Ortega P A and Braun D A 2010 A minimum relative entropy principle for learning and acting *J. Artif. Int. Res.* **38** 475–511
[12] Ortega P A and Braun D A 2013 Thermodynamics as a theory of decision-making with information-processing costs *Proc. R. Soc.* A **469** 2153
[13] van den Broek J L, Wiegerinck W A J J and Kappen H J 2010 Risk-sensitive path integral control *UAI* **6** 1–8
[14] Haarnoja T, Tang H, Abbeel P and Levine S 2017 Reinforcement learning with deep energy-based policies (arXiv:1702.08165v2)
[15] Neftci E O and Averbeck B B 2019 Reinforcement learning in artificial and biological systems *Nat. Mach. Intell.* **1** 133–43
[16] Morimura T, Sugiyama M, Kashima H, Hachiya H and Tanaka T 2010 Nonparametric return distribution approximation for reinforcement learning *ICML*
[17] Schernaum S, Dshemushadse M, Rischer R and Goschke T 2010 How decisions evolve: the temporal dynamics of action selection *Cognition* **117** 407–16
[18] Gallivan J P, Logan L, Wolpert D M and Flanagan J R 2016 Parallel specification of competing sensorimotor control policies for alternative action options *Nat. Neurosci.* **19** 320–9
[19] Kim B and Basso M A 2010 A probabilistic strategy for understanding action selection *J Neurosci* **30** 2340–55
[20] Zhang J, Hughes L E and Rowe J B 2012 Selection and inhibition mechanisms for human voluntary action decisions *NeuroImage* **63** 392–402
[21] Kurth-Nelson Z and Redish A D 2009 Temporal-Difference Reinforcement Learning with distributed representations *Plos One 1371/journal.pone.0007362* **4** e7362
[22] Benjamini Y and Yekutieli D 2001 The control of the false discovery rate in multiple testing under dependency *Ann. Stat.* **29** 1165–88
[23] Walther A, Nili H, Ejaz N, Alink A, Kriegeskorte N and Diedrichsen J 2016 Reliability of dissimilarity measures for multi-voxel pattern analysis *Neuroimage* **137** 188–200
[24] Barany D A, Della-Maggiore V, Viswanathan S, Cieslak M and Grafton S T 2014 Feature interactions enable decoding of sensorimotor transformations for goal-directed movement *J Neurosci* **34** 6860–73
[25] Saad Z S, Reynolds R C, Argall B, Japee S and Cox R W 2004 SUMA: an interface for surface-based intra- and inter-subject analysis with AFNI *Biomedical Imaging: Nano to Macro IEEE Int. Symp.* **2** 1510–13
[26] Tomassini V, Jbabdi S, Klein J C, Behrens T E, Pozzilli C, Matthews P M, Rushworth M F and Johansen-Berg H 2007 Diffusion-weighted imaging tractography-based parcellation of the human lateral premotor cortex identifies dorsal and ventral subregions with anatomical and functional specializations *J Neurosci* **27** 10259–69
[27] Picard N and Strick P L 2001 Imaging the premotor areas *Curr Opin NeuroBiol* **11** 663–72
[28] Destrieux C, Fischl B, Dale A and Halgren E 2010 Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature *Neuroimage* **53** 1–15
[29] Geyer S, Ledberg A, Schleicher A and Kinomura S 1996 Two different areas within the primary motor cortex of man *Nature* **382** 805
[30] Schneider T and Hermsdörfer J 2016 Anticipation in object manipulation: behavioral and neural correlates *Progress in*

Motor Control ed Laczko & Latash Advances in Experimental Medicine and Biology **957** pp 173–94
[31] Eickhoff S B, Stephan K E, Mohlberg H, Grefkes C, Fink G R, Amunts K and Zilles K 2005 A new SPM toolbox for combining probabilistic cytoarchitectonic maps and functional imaging data *Neuroimage* **25** 1325–35
[32] Eickhoff S B, Heim S, Zilles K and Amunts K 2006 Testing anatomically specified hypotheses in functional imaging using cytoarchitectonic maps *Neuroimage* **32** 570–82
[33] Eickhoff S B, Paus T, Caspers S, Grosbras M H, Evans A C, Zilles K and Amunts K 2007 Assignment of functional activations to probabilistic cytoarchitectonic areas revisited *Neuroimage* **36** 511–21
[34] Eisenberg M, Shmuelof L, Vaadia E and Zohary E 2010 Functional organization of human motor cortex: directional selectivity for movement *J Neurosci* **30** 8897–905
[35] Eisenberg M, Shmuelof L, Vaadia E and Zohary E 2011 The representation of visual and motor aspects of reaching movements in the human motor cortex *J Neurosci* **31** 12377–84
[36] Haxby J V, Connolly A C and Guntupalli J S 2014 Decoding neural representational spaces using multivariate pattern analysis *Annu. Rev. Neurosci.* **37** 435–56
[37] Ejaz N, Hamada M and Diedrichsen J 2015 Hand use predicts the structure of representations in sensorimotor cortex *Nat. Neurosci.* **18** 1034–40
[38] Fabbri S, Stubbs K M, Cusack R and Culham J C 2016 Disentangling representations of object and grasp properties in the human brain *J Neurosci* **36** 7648–62
[39] Kriegeskorte N, Mur M and Bandettini P 2008 Representational similarity analysis–connecting the branches of systems neuroscience *Front Syst Neurosci* 2
[40] Diedrichsen J, Provost S and Zareamoghaddam H 2016 On the distribution of cross-validated Mahalanobis distances (arXiv:160701371)
[41] Walther A, Nili H, Ejaz N, Alink A, Kriegeskorte N and Diedrichsen J 2016 Reliability of dissimilarity measures for multi-voxel pattern analysis *Neuroimage* **137** 188–200
[42] Lee A B and Nadler B 2007 Treelets — A Tool for Dimensionality Reduction and Multi-Scale Analysis of Unstructured Data *arXiv* 0707.0481v3
[43] Ashburner J and Friston K J 2005 Unified segmentation *Neuroimage* **26** 839–51
[44] Diedrichsen J 2006 A spatially unbiased atlas template of the human cerebellum *Neuroimage* **33** 127–38
[45] Diedrichsen J, Balsters J H, Flavell J, Cussans E and Ramnani N 2009 A probabilistic MR atlas of the human cerebellum *Neuroimage* **46** 39–46
[46] Diedrichsen J, Maderwald S, Küper M, Thürling M, Rabe K, Gizewski E, Ladd M E and Timmann D 2011 Imaging the deep cerebellar nuclei: a probabilistic atlas and normalization procedure *Neuroimage* **54** 1786–94
[47] Diedrichsen J and Zotow E 2015 Surface-based display of volume-averaged cerebellar imaging data *PloS one* **10** e0133402
[48] Diedrichsen J and Shadmehr R 2005 Detecting and adjusting for artifacts in fMRI time series data *Neuroimage* **27** 624–34
[49] Fu Q, Hasan Z and Santello M 2011 Transfer of learned manipulation following changes in degrees of freedom *J Neurosci* **31** 13576–84
[50] Bellemare M G, Dabney W and Munos R 2017 A Distributional Perspective on Reinforcement Learning (arXiv:1707.06887v1)
[51] Haar S, Donchin O and Dinstein I 2017 Individual movement variability magnitudees are explained by cortical neural variability *J Neurosci* **37** 9076–85
[52] Wu H G, Miyamoto Y R, Castro L N G, Lvecsky B P and Smith M A 2014 Temporal structure of motor variability is dynamically regulated and predicts motor learning ability *Nat. Neurosci.* **17** 312–21
[53] Olveczky B P, Andalman A S and Fee M S 2005 Vocal experimentation in the juvenile songbird requires a basal ganglia circuit *PLOS Biology* **3** e153
[54] Kao M H, Doupe A J and Brainard M S 2005 Contributions of an avian basal ganglia-forebrain circuit to real-time modulation of song *Nature* **433** 638–43
[55] Tumer E C and Brainard M S 2007 Performance variability enables adaptive plasticity of 'crystallized' adult birdsong *Nature* **450** 1240–4
[56] Nagabandi A, Kahn G, Fearing R S and Levine S 2017 Neural Network Dynamics for Model-Based Deep Reinforcement Learning with Model-Free Fine-Tuning *NIPS*
[57] Hassabis D, Kumaran D, Summerfield C and Botvinik M 2017 Neuroscience-Inspired Artificial Intelligence *Neuron* **95** 245–58
[58] Pong V, Gu S, Dalal M and Levine S 2018 Temporal Difference Models: Model-Free Deep RL for Model-Based Control *ICLR*
[59] Bellemare M G, Dabney W and Munos R 2017 A Distributional Perspective on Reinforcement Learning *ICML*
[60] Clavera I, Nagabandi A, Fearing R S, Abbeel P, Levine S and Finn C 2018 Learning to Adapt: Meta-Learning for Model-Based Control *NIPS*
[61] Rao A V 2010 A Survey of Numerical Methods for Optimal Control *Advances in the Astronautical Sciences* **135** 1
[62] Burghouts G J, Smeulders A W M and Geusebroek J M 2008 The Distribution Family of Similarity Distances *NIPS*
[63] Glascher J, Daw N, Dayan P and O'Doherty J P 2010 States versus Rewards: Dissociable Neural Prediction Error Signals Underlying Model-Based and Model-Free Reinforcement Learning *Neuron* **66** 585–95
[64] Glascher J, Adolphs R, Damasio H, Bechara A, Rudrauf D, Calamia M, Paul L K and Tranel D 2012 Lesion mapping of cognitive control and value-based decision making in the prefrontal cortex *Proc. Natl Acad. Sci.* **109** 14681–6
[65] Guo R, Bohmer W, Hebart M, Chien S, Sommer T, Obermayer K and Glascher J 2016 Interaction of Instrumental and Goal-Directed Learning Modulates Prediction Error Representations in the Ventral Striatum *J Neurosci* **36** 12650–60
[66] Schultz W 2017 Reward Prediction Error *Current Biology* **27** R369-R371
[67] Burda Y, Edwards H, Storkey A and Klimov O 2018 Exploration by Random Network Distillation (arXiv:1810.12894v1)
[68] Starkweather C K, Babayan B M, Uchida N and Gershman S J 2017 Dopamine reward prediction errors reflect hidden state inference across time *Nat. Neurosci.* **20** 581–9
[69] Andrychowicz M, Wolski F, Ray A, Schneider J, Fong R, McGrew B, Tobin J, Abbeel P and Zaremba W 2017 Hindsight Experience Replay *NIPS*
[70] Kingma D P and Lei Ba J 2015 ADAM: a method for stochastic optimization *ICLR* (https://arxiv.org/pdf/1412.6980.pdf.)
[71] Brockman G, Cheung V, Pettersson L, Schneider J, Schulman J, Tang J and Zaremba W 2016 OpenAI Gym. (arXiv:1606.01540v1)