Interpreting reach adaptation in the actor-critic framework

We aim to identify how a graded reward signal affects motor adaption; specifically, we explore the differences in motor adaptation when visuomotor error & reward signals are included or excluded from the subject's feedback. The subfield of reward-based or reinforcement learning has been coarsely influential in human motor adaptation studies, but the direct connection between the theory and the behavior has been elusive. Recent literature has equated the acquisition of a target with "reward" (Izawa and Shadmehr, 2011), but the binary signal of success or failure on a trial provides only a coarse signal to compare to theory or, per our current study, to reverse-engineer how people differentially adapt in response to error versus reward.

Subjects perform sets of reaching movements while holding a robotic manipulandum. On each day of training, subjects perform 160 movements in a null (zero external force) environment, followed by 160 movements in a perturbing environment. Each movement starts in front of the subject's chest, ends 10 cm further from the subject's chest, and is intended to be 750±100 ms long. The reward is displayed at the end of each movement and its values range from 0 to 100 points; it is decreasing function of the absolute area between the direct trajectory and the subject's path. The subjects have no prior knowledge of how the score is related to their movement, but they are aware that their monetary compensation for participating is proportional to the total number of points they earn on each day. When reward is present in the task, subjects are asked to predict the score they are going to receive (i.e. a guess between 0 & 100) before each upcoming trial.

We use the actor-critic model as an interpretive framework to analyze human motor adaptation data. A key prediction from this theory is that predictions of reward will change in step sizes proportional to the previous prediction error. If the subject has no access to relevant sensory feedback, they should be relying mostly upon the reward signal for adaptation & they should also make incremental changes in movement strategy that are also proportional to the reward prediction error.

With the first 12 subjects, we compare adaptation in the constant viscous field with reward & visuomotor feedback against adaptation without reward & only visuomotor feedback in the oppositely signed field. First, we discovered that all 12 subjects update their reward predictions in proportion to the reward prediction error (Figure #1A); α_C represents this proportionality and is generated by fitting a generalized linear model over the reward prediction updates and the reward prediction error. For all 12 subjects, this value is significantly greater than zero ($p < 0.001$). Similarly, for 10 out of 12 subjects trial-by-trial changes in movement error (i.e. perpendicular displacement at peak speed) are significantly correlated with the reward prediction error (Figure #1B; $p < 0.001$). Overall, this experiment suggests that the verbal reward prediction can function as a measure of the subject's action-value function; the error signal generated from this prediction seems to affect the trial-by-trial changes in movement error, as predicted by the actor-critic model. In the presence of visuomotor feedback, the addition of reward did not appear to affect the completeness or time constant of adaptation (paired t-test: $p > 0.05$).

With the second 12 subjects, the 160 perturbed movements took place in a movement channel; the robot generates two virtual viscoelastic walls that constrain the subject to the path directly between the start and target points. Using this technique, we can make the subjects generate lateral forces (i.e. perpendicular to the target direction) without allowing them to deviate from the straight path. The reward signal is closest to 100 when the subject generates lateral forces that are proportional to their velocity. The channel and the reward signal together create a 'virtual' viscous field. We view adaptation in the virtual viscous field akin to adaptation in the real viscous field without the visuomotor feedback that reveals deviation from the direct trajectory.

We averaged the force-profiles of five consecutive trials & fit each average to a state dependent model: $F(t) = -K*y(t)-B*(dy/dt)$. We use these parameters to describe each subject's action strategy; the maximally reward strategy is $B = \pm 15$ Ns/m (sign is constant for each subject) and $K = 0$ N/m. Ten out of 12 subjects updated reward prediction values in proportion to the prediction error (Figure #2A, $p < 0.001$) suggesting that they are at least attempting to learn how the reward works. The mapping between force output and reward is highly nonlinear & the state-dependent model does not explain all of the variance in force output. Even though subjects were considering a complex action space, many of them were able to discover a highly rewarded behavior (Figure #3D). Nine out of twelve subjects were able to discover that pushing in one direction is rewarded more than pushing in the other. Seven out of those nine were able to generate the appropriate magnitude of viscous force generation as well.

Subjects are capable of adapting to the virtual viscous field with only reward feedback, but it is difficult for them to predict how incremental changes in actions will affect their reward. This experiment suggests that in the absence of visuomotor feedback subjects can develop a state- or action-value function to relate movements and rewards. This particular training regimen within this particular reward space was sufficient to induce lasting, rather than briefly transient, changes in perpendicular force output.

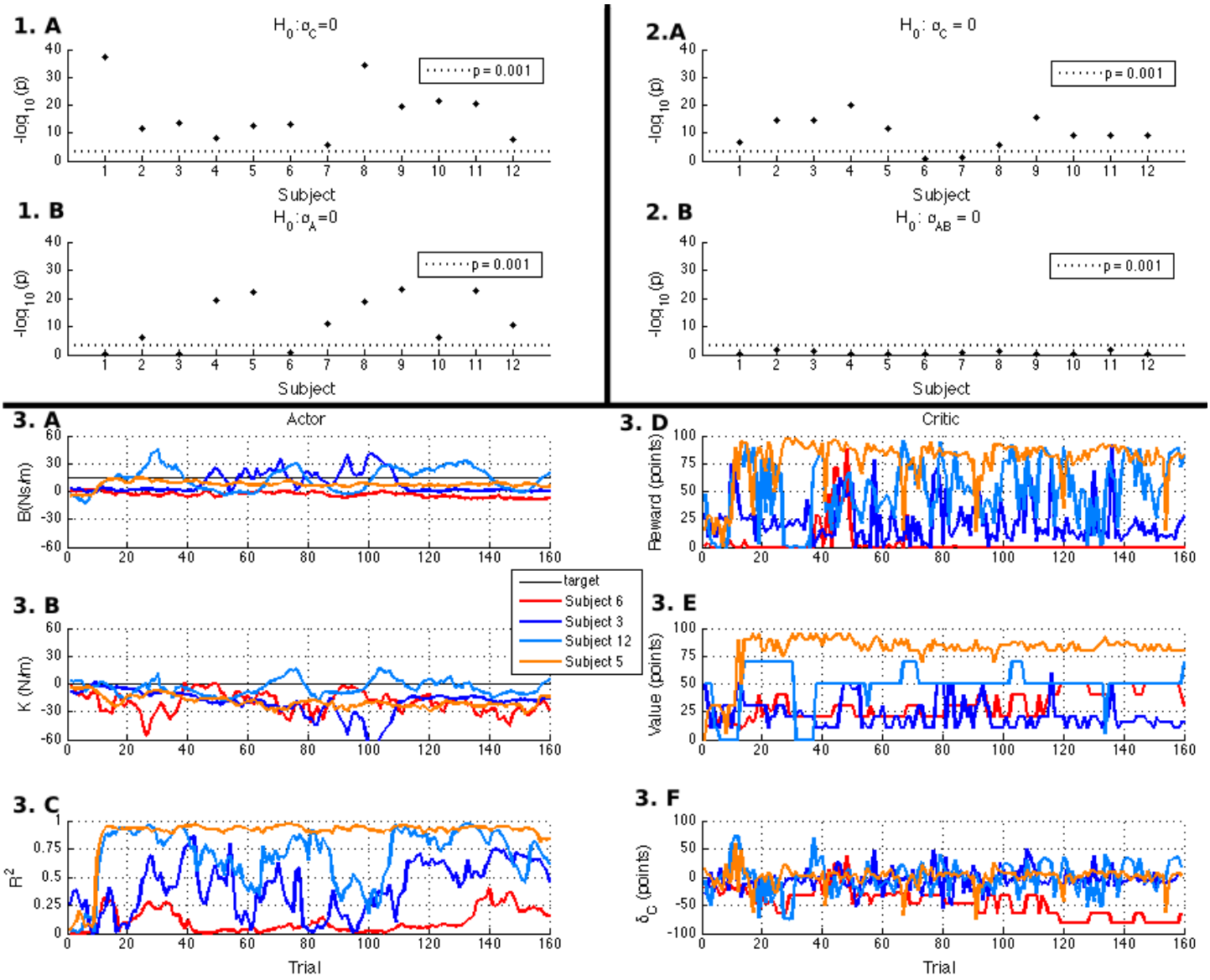

Figure 1: Critic & actor learning rates with sensory feedback – Each figure shows the p-values on the fits of generalized linear model. The null hypothesis in each case is that the parameter is equal to zero; we reject the null hypothesis at significance $p < 0.001$. (A) All 12 subjects showed a significant correlation between reward prediction update and reward prediction error. (B) Nine of twelve subjects showed a significant correlation between change in movement error between trials and reward prediction error.

Figure 2: Critic & actor learning rates without sensory feedback – Each figure shows the p-values on the fits of generalized linear model. The null hypothesis in each case is that the parameter is equal to zero; we reject the null hypothesis at significance $p < 0.001$. (A) Ten of twelve subjects showed a significant correlation between reward prediction update and reward prediction error. (B) No subject updated movement viscosity in proportion to the reward prediction error. Though some subjects were still able to find some rewarding behavior, this suggests that subjects are exploring a much higher-dimensional space than the two-dimensional state-dependent model.

Figure 3: Actions, rewards & predictions without sensory feedback – The relevant actor-critic model variables are plotted for each trial. Four subjects are shown to simplify presentation. Subject 6 represents a subject who did not learn to push into the correct wall. Subjects 3 & 12 demonstrate that subjects can find a well-rewarded movement strategy, but that they do not always stay with that strategy. Subject 5 represents a subject who found a highly rewarded behavior, and varied only slowly from this strategy. (A) The timecourse of movement viscosities. (B) The timecourse of movement elasticities. Almost every subject generated non-zero position-related forces; mostly leftward forces (negative) torward the end of movement. (C) State-dependent model fits for each trial. The state-dependent model does not accounut for much of the variance in force output for many subjects. (D) The timecourse of rewards for each subject. (E) The time course of value predictions. (F) The timecourse of temporal difference errors based off of reward predictions. Subject 6 is unable to reduce the magnitude of the TD error; he/she over-estimates rewards for much of the training set.

[1] Doya, K. (2000). Reinforcement learning in continuous time & space. *Neural Computation*, *12*¸ 219-245.

[2] Doya, K. (2002). Metalearning and neuromodulation. *Neural Networks*, *15*, 495-506.

[3] Izawa, J., & Shadmehr, R. (2011). Learning from sensory and reward prediction errors during motor adaptation. *PLoS Computational Biology*, *7*(3), e1002012.