

## Research Article

### THE UNDERSHOOT BIAS: Learning to Act Optimally Under Uncertainty

Sascha E. Engelbrecht, Neil E. Berthier, and Laura P. O'Sullivan

*University of Massachusetts, Amherst*

**Abstract**—*Learning in stochastic environments is increasingly viewed as an important psychological ability. To extend these results from a perceptual to a motor domain, we tested whether participants could learn to solve a stochastic minimal-time task using exploratory learning. The task involved moving a cursor on a computer screen to a target. We systematically varied the degree of random error in movement in three different conditions; each condition had a distinct time-optimal solution. We found that participants approximated the optimal solutions with practice. The results show that adults are sensitive to the stochastic structure of a task and naturally adjust the magnitude of an undershoot bias to the particular movement error of a task.*

Psychology has long recognized that humans are sensitive to the inherent stochasticity of their senses. The basis of signal detection theory is the assumption that particular distal stimuli result in distributions of psychological responses, and that observers adopt decision criteria on the basis of the costs of making errors and correct classifications. In such stochastic problems, one can determine ideal or *optimal* solutions by minimizing or maximizing some cost function, for example, by maximizing the probability of a correct classification (e.g., Dayan & Abbott, 2001). Ernst and Banks (2002) extended these ideas to the sensory fusion problem by showing that adults weigh information from the visual and haptic modalities so as to minimize the variance of estimates of an object's properties.

A second thread in the literature on stochasticity concerns learning about the structure of the world from a stochastic input. For example, Aslin, Saffran, and Newport (1998) studied word segmentation by infants by presenting infants with a stream of nonsense syllables. In this context, "words" were defined as combinations of syllables that had high intersyllable transition probabilities, and boundaries between words were signaled by low intersyllable transition probabilities. During a subsequent test phase, 8-month-old infants responded differently to the words and nonwords. Aslin et al. concluded that the infants kept track of the transition probabilities during familiarization and used them to segment the acoustic stream into candidate words.

Fiser and Aslin (2001) performed a similar experiment involving visual object recognition among adults. In the training set presented to the participants, objects were defined by the joint and conditional probabilities of features. During test, the participants were supposed to choose which of a pair of novel stimuli was more familiar. Training in this experiment was unsupervised, in that the participants were not given any feedback during presentation of the training set. During test, participants chose objects that were consistent with the probabilities of feature combination during training. Further, most subjects lacked awareness of the un-

derlying structure of the data during testing and reported that they were simply guessing.

Although signal detection theory and statistical learning deal with perceptual decision making, complementary issues arise in motor control. When human adults are instructed to repeatedly produce identical movements of prescribed amplitude and duration, they instead produce a distribution of movements that is characterized by a scattering of endpoints, whose standard deviation scales linearly with average movement speed (Meyer, Smith, & Wright, 1982; Schmidt, Zelaznik, Hawkins, Frank, & Quinn, 1979). Apparently, subjects do not, and indeed can not, produce identical movements across a series of trials. Furthermore, there are physiological data showing that stochasticity is a fundamental characteristic of neural information processing (Calvin & Stevens, 1968; Clamann, 1969).

Just as optimization can specify decision criteria in perceptual classification, optimization can be used to specify actions in motor tasks. People often seek to accomplish a goal in a noisy environment in an optimal manner, perhaps trying to reach for an object with the least effort, via the smoothest trajectory, or with the greatest speed. Meyer, Abrams, Kornblum, Wright, and Smith (1988) suggested that the speed-accuracy trade-off captured by Fitts' law is the result of attempting to reach to a target in minimal time in the presence of motor noise. Meyer et al. showed an empirical match between reaching and computed time-optimal movements in a series of experimental studies. Other researchers have extended this idea to the development of reaching movements in human infants (Berthier, 1996) and to saccadic eye movements (Harris & Wolpert, 1998). Engelbrecht (2001) discussed at length the concept of optimality in relation to human movement.

Although optimal action is something humans seem to do naturally, computing optimal solutions to real problems is extremely difficult in practice because of the curse of dimensionality (Bellman, 1961; Sutton & Barto, 1998). One must compute the quality of all possible solutions to determine which is the best. For most real tasks, such as reaching movements, this would require computing the quality of all combinations of possible reaching movements, varying in direction, amplitude, and speed, with all possible types of grasp and arm posture.

However, recent advances in computational theory provide a way to compute *approximately* optimal solutions to stochastic control problems. These algorithms conceive of the problem as one in which an actor explores the results of his or her actions from various states. These methods are usually called "reinforcement learning" (Sutton & Barto, 1998) or "neurodynamic programming" (Bertsekas & Tsitsiklis, 1996), and they have been shown to be of practical utility in controlling artificial systems. Interestingly, these methods were inspired by results from animal learning.

The new computational algorithms view the learning problem as intermediate to supervised and unsupervised learning. In a supervised learning situation, a "teacher" provides a desired output for each input. The popular connectionist technique, backpropagation of error, is an algorithm that uses supervised learning. In contrast, unsupervised learning involves the

Address correspondence to Neil E. Berthier, Department of Psychology, Tobin Hall, University of Massachusetts, Amherst, MA 01003; e-mail: berthier@psych.umass.edu.

The Undershoot Bias

adaptive, statistical structuring of some set of inputs without feedback (e.g., Fiser & Aslin, 2001). In an exploratory or reinforcement learning task, feedback is provided during learning, but that feedback is an overall assessment of the system's performance, not explicit guidance, and may be corrupted by noise (Sutton & Barto, 1998).

What is learned or computed by these algorithms is similar to a motor program or motor plan, but is closed loop. When the algorithms converge, the program generates a mapping between states of the dynamical system and the elementary actions that are optimal for those states. Optimal movements can then be made by executing the optimal action from the start state, observing the resulting state, executing the optimal action from that state, and so on, until the goal state is reached. This framework can result in a blended sequence of movements and does not commit the actor to a discrete sequence of separable movements. This view of movement is similar to the iterative correction model of Crossman and Goodeve (1983) and Keele (1968).

The use of stochastic optimal control theory as a model of infant reaching (Berthier, 1996) and eye movements (Harris & Wolpert, 1998) has proved valuable in that movement kinematics predicted by the mathematical theories generally match observed behavior. One prediction of the theories is that the initial movement of a sequence should generally undershoot the target, and empirical studies of both reaching and eye movements have supported this prediction (Aslin & Salapatek, 1975; Berthier, 1996; Chua & Elliott, 1993; Harris & Wolpert, 1998; Henson, 1978; Hofsten, 1991; Toni, Gentilucci, Jeannerod, & Decety, 1996; Vercher, Mageses, Prablanc, & Gauthier, 1994). The most influential theory of the speed-accuracy trade-off (Meyer et al., 1988) uses a more restricted mathematical formulation that assumes the initial movement does not undershoot the target, an assumption that is not consistent with either reaching or eye movement data (cf. Elliott, Helsen, & Chua, 2001).

The observations of Berthier (1996) and Harris and Wolpert (1998) that motor noise and policies that control human reaching and eye movements go hand in hand might be dismissed as mere correlations. Further, the fact that plans for action in noisy situations are difficult to discover also suggests that humans might be unable to solve stochastic optimal control tasks by exploration. In the current study, we directly manipulated the amount of stochasticity in a task to determine if adults discover and adopt time-optimal movement kinematics appropriate for the different levels of noise. Adults were presented with a novel task that involved using two keys on a computer keyboard to move a cursor to a target. We directly manipulated the amount of motor error by adding noise to the command for movement. Participants were asked to move the cursor to the target in the shortest time, but were otherwise unpracticed and uninformed about the nature of the task. An important feature of this experiment was the fact that we were able to numerically compute the optimal policy for each level of stochasticity using dynamic programming. If stochastic optimal control theory and reinforcement learning are good models of human learning, then participants should be able to develop near-optimal plans for solving our task, and those plans should track the task's level of stochasticity.

METHOD

Participants

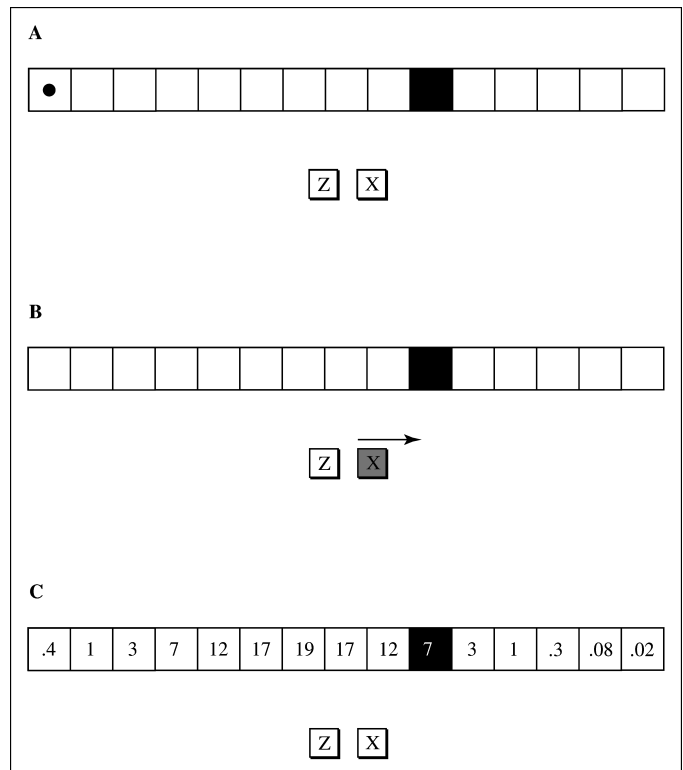
One male and five females participated in the experiment. The participants were psychology graduate students who were paid an hourly

fee for their participation. None of the participants were knowledgeable about the aims of the study, nor were they informed about any details of the mathematical model used in the experiment.

Task and Procedure

The experimental task is shown in Figure 1a. An array of 15 squares (36 pixels to a side) was shown on a 14-in. computer screen. Each square represented a state in a one-dimensional state space. A dot (radius of 3 pixels) displayed in one of the squares informed the participant of the current state of the dynamical system. Participants controlled the dot's movement by pressing one of two keys ("Z" or "X") on a standard QWERTY keyboard. The task was to move the dot from its initial position to the goal state indicated by an outlined box. The dot was to be moved from its position at the start of the trial to the goal as quickly as possible (i.e., in minimum time).

The state was given by the current box ( $x \in \{0, 1, \dots, 14\}$ ). At the beginning of each trial, the dot was located at the center of the left-most square ( $x_0 = 0$ ), and the target was nine squares to the right ( $x_T = 9$ ). In general, the "X" key moved the dot to the right, and the "Z" key moved the dot to the left. The dot was moved in a sequence of discrete movements. As soon as either the "X" or the "Z" key was pressed, the



**Fig. 1.** The task. The black dot, initially located at the center of the left-most square (a), is to be moved to the dark area (in the experiment, the target was indicated instead by an outlined box), located nine squares to the right. If the subject presses the right key for a certain duration (b), the dot disappears immediately and remains invisible for the full duration of the key press. Upon release of the button, the dot reappears. The degree of movement error in the high-noise ( $\sigma = 0.35$ ) condition is illustrated in (c). The number in each square shows the expected number of dot appearances in that square in response to 100 identical 0.5-s key presses (i.e.,  $\Delta x = 6$ ) from the start state.

dot disappeared from the screen (Fig. 1b). When the key was released, the dot reappeared at a new location. The distance traveled was a function of the duration of the key press and random noise. Specifically, the choice of key and the duration of the key press were first converted into a command,  $\Delta x_k$ , according to the following equation:

$$\Delta x_k(t_k) = \begin{cases} -12t_k & \text{if } Z \\ +12t_k & \text{if } X, \end{cases}$$

where  $t_k$  is the duration (in seconds) of the  $k$ th key press within a trial (i.e., a key press of 1 s resulted in an expected movement of 12 squares). The next state of the system was determined by adding Gaussian noise to the movement command,  $\Delta x_k$ , and adding the result to the current position. Specifically,

$$x_{k+1} = x_k + \Delta x_k + \eta(0, \sigma|\Delta x_k|), \quad (1)$$

where  $\eta(0, \sigma|\Delta x_k|)$  was a normally distributed random variable with a mean of zero and standard deviation equal to  $\sigma|\Delta x_k|$ . The value of  $x_{k+1}$  was limited to be between 0 and 14. This stochastic model of the movement can then be thought of as linear scaling of the endpoint error with the speed of movement (Meyer et al., 1982; Schmidt et al., 1979). Participants performed movements in separate conditions at three levels of motor noise,  $\sigma = 0.0, 0.2, \text{ or } 0.35$ . Figure 1c illustrates the degree of movement error in the high-noise ( $\sigma = 0.35$ ) condition.

If the dot’s new location was inside the target zone ( $x = x_T$ ), the trial was terminated; otherwise, the subject performed another key press, whereupon the dot’s location was again updated according to the procedure just outlined. This cycle was repeated until the dot was at the target location.

For each noise level (i.e., for each value of  $\sigma$ ), a minimum-time policy was numerically computed by formulating the task as a Markov decision process. The time of a particular movement was computed as the total sum of times required to execute a discrete sequence of movements from the start state to the goal. The state transition from one state to the next was given by Equation 1. The state ( $X \in \{0, 1, \dots, 14\}$ ) and control ( $U \in \{-14, \dots, 14\}$ ) spaces were finite. The time required to make the state transition was

$$g(x, \Delta x) = \begin{cases} 0 & \text{if } x = x_T, \Delta x = 0 \\ |\Delta x|/12 + t_R & \text{otherwise,} \end{cases}$$

where  $|\Delta x|/12$  was equivalent to key-press duration (in seconds), and  $t_R$  was the average reaction time (in seconds), as determined from the experimental data. The optimal policy for the decision process (i.e., the minimum-time policy) was determined using value iteration, a standard dynamic programming technique (see, e.g., Bertsekas, 1987).

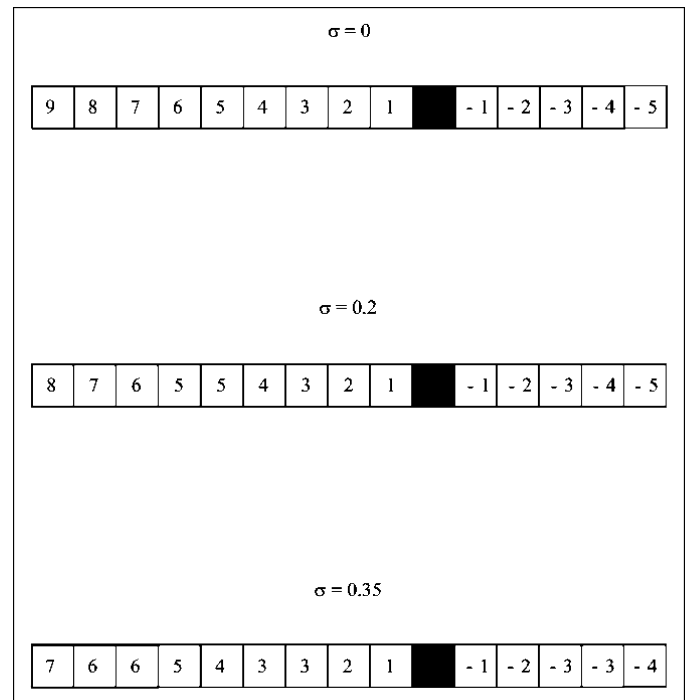
Participants were simply told to “try and move the dot to the goal as quickly as possible.” There was a separate session for each of the three noise levels ( $\sigma = 0.0, 0.2, \text{ or } 0.35$ ), and each participant took part in all three sessions. For 3 participants, the order of the noise levels was 0.0, 0.2, 0.35, and for the other 3, the order was 0.0, 0.35, 0.2. (The deterministic condition was always first because it permitted the participants to learn the basic task characteristics, such as the relation between key-press duration and object movement, in a reasonable amount of time.) Participants worked in blocks of 50 trials, with short rest periods between blocks. At the end of each block, participants were told the average movement time over the 50 trials of that block. Because of the increasing difficulty of the task with increasing stochasticity, partici-

pants were given 10 blocks at  $\sigma = 0.0$ , 15 blocks at  $\sigma = 0.2$ , and 20 blocks at  $\sigma = 0.35$ . Performance had stabilized at the end of training, as indicated by a stabilization of the initial key-press duration ( $\Delta x_0$ ). Participant 4 received 30 blocks of trials at  $\sigma = 0.35$  because of movement command variability, but when this participants’ data for the 20th and 30th blocks were compared, no differences in the average command were detected. Also, because of a procedural error, 1 participant received only 9 blocks of trials at  $\sigma = 0.0$ .

## RESULTS

Figure 2 shows numerically computed minimum-time policies for the three experimental conditions. The number in each box is the optimal command for that state. For example, from state  $x = 0$ , the optimal command was 9 when  $\sigma = 0.0$ , 8 when  $\sigma = 0.2$ , and 7 when  $\sigma = 0.35$ . This means that the time-optimal command from the start state undershoots the target one square when  $\sigma = 0.2$ , and undershoots the target two squares when  $\sigma = 0.35$ .

In general, undershooting is optimal in stochastic reaching or aiming problems for two reasons (e.g., Elliott et al., 2001). First, when movements involve inertia (masses), overshooting incurs heavy penalties because one must decelerate the mass to move back to the target. Second, movements past the target require the actor to move longer distances because the hand backtracks to cover the length of the overshoot, thus covering the distance twice. Given that our task involved no inertia, it is the second reason that caused undershoots to be optimal in the current experiment. In a task involving inertia, undershooting would require only adjusting the gain of the movement, whereas over-



**Fig. 2.** Minimum-time policies for the three noise-level conditions. The number in each box refers to the time-optimal action, for that state. The positive commands would result from pressing the “X” key; the negative commands, from pressing the “Z” key.

The Undershoot Bias

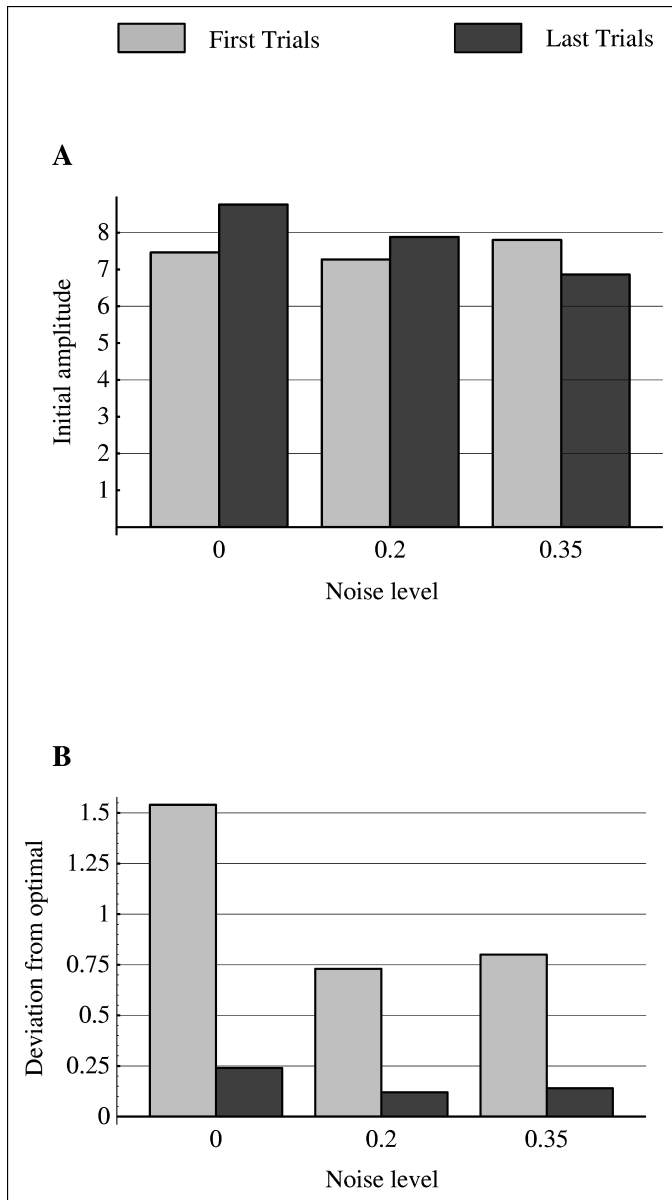
shooting would require reprogramming to reverse the direction of movement.

Figure 3 summarizes the subjects' performance. The data in both panels are averages over 150 trials. They are grouped by noise level ( $\sigma = 0.0, 0.2, \text{ and } 0.35$ ) and practice (first 150 trials vs. last 150 trials). Figure 3a shows the control command for the initial movement of each trial ( $\Delta x_0$ ). An analysis of variance (ANOVA) did not show any differences across the three noise levels in the initial command for the initial 150 trials,  $F(2, 10) = 0.68, p > .05$ . At the start of training and in all three conditions, subjects aimed to undershoot the target by between 1 and 1 1/2 squares. Clearly, this was a suboptimal pattern. A larger control

command would have been preferable in the no-noise condition, whereas a smaller one should have been used in the high-noise condition.

Practice led to a differentiation of the strategies in the three noise-level conditions. In the no-noise and medium-noise conditions, the initial amplitude increased, whereas in the high-noise condition, it decreased. ANOVA indicated that the average initial amplitude at the end of training differed by level of stochasticity,  $F(2, 10) = 10.09, p < .005$ , and that initial commands were better adapted to the particular noise levels at the end of training than at the beginning. This is more easily seen in Figure 3b, which shows the magnitude of deviation of the subjects' average control command from the one prescribed by the time-optimal policies. In each noise-level condition, practice resulted in a large reduction of the error.

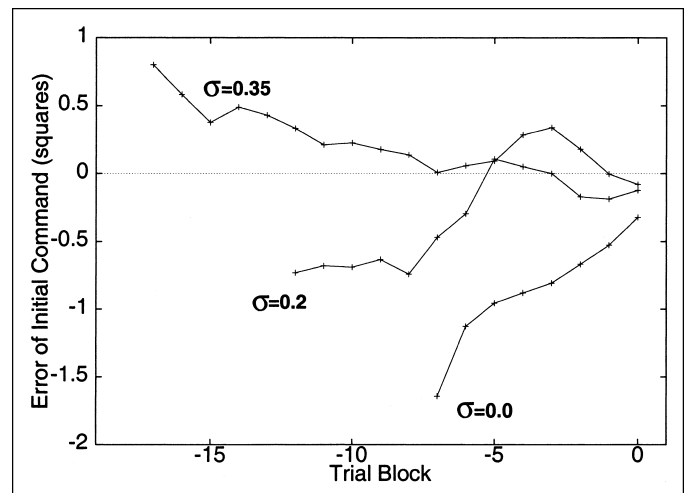
Figure 4 shows the learning curve for participants in each of the three conditions. The data are running averages over three blocks of trials, with Block 0 being the last block of training. Plotted performance is the error of the first movement from the computed optimal movement. The figure shows that the deviation from optimal was substantial on the first blocks of training, but that by the end of training, participants closely approximated the optimal initial movement.



**Fig. 3.** Initial control command (a) and absolute difference between the initial control command and the one prescribed by the minimum-time policy (b) as a function of noise level ( $\sigma = 0.0, 0.2, \text{ and } 0.35$ ) and practice (first 150 trials vs. last 150 trials). All data are averages over 150 trials and 6 subjects.

DISCUSSION

The results clearly indicate that subjects are sensitive to variations in a task's stochastic component, and that they are able to adapt their behavior accordingly. This is an impressive result given the difficulty of the task. First, the participants were not familiar with the dynamics of the task and received no information about the stochastic component involved in it. Second, participants never received any information about how to improve their performance. Third, performance and estimates of movement time from a single trial were often misleading. Given the level of noise in the task, it was likely that on some trials participants chose the optimal action, but because of the particular random number chosen on that trial, the movement time was very long. Likewise, the choice of a very nonoptimal action could result in a short movement time if the random number chosen on that trial was small. Consequently, if participants based their movement strategies



**Fig. 4.** Learning curves for the three levels of stochasticity. The plotted data are running averages over three trial blocks (150 trials), with Block 0 being the last block of training.

on the results of the movements from single trials, those strategies would likely be in error (see, e.g., Sutton & Barto, 1998). Somehow, our participants integrated information over a series of trials to improve performance.

Although the current task is not meant to be a direct model of reaching, pointing, or eye movements, it contains central aspects of those tasks. Actions were made in the presence of noise that required on-line correction. The fact that participants adjusted action plans to the level of noise through practice suggests that adults could also learn near-optimal solutions to real-world tasks with a reasonable amount of practice. This finding, coupled with the fact that reaching and eye movements appear to be time-optimal, supports optimal control models such as those of Meyer et al. (1988), Berthier (1966), and Harris (1995), and models of exploratory learning such as those of Berthier (1996) and Sutton and Barto (1998).

Optimal solution of our task required an undershoot bias in the presence of noise. Undershoot biases have also been observed in arm (Hofsten, 1991; Toni et al., 1996) and eye (Aslin & Salapatek, 1975; Henson, 1978) movements. Other researchers have computationally demonstrated that undershoot biases are often required for optimal behavior. Harris (1995) showed that saccadic undershoot is a stochastic optimal control strategy that minimizes total flight time. Berthier (1996) extended this analysis to infant reaching movements, which are typically composed of multiple movements. As infants develop, the number of these units decreases, the amplitude of the initial unit increases, and the speed of the individual units follows an N-shaped curve (increase–decrease–increase). Berthier showed that this complex developmental pattern can be reproduced if one assumes that infants behave according to a stochastic time-optimal strategy that is continuously adapted to the level of motor noise, which decreases with the infant's age. The observed and predicted undershoot biases are not, however, consistent with the stochastic optimized-submovement model of Meyer et al. (1988), who assumed that the average initial movement takes the hand to the target.

It appears that one must assume the existence of a mechanism for learning stochastic optimal control policies from (relatively unspecific) evaluative feedback. There has been a great deal of theoretical work concerning potential mechanisms for enabling such a powerful learning process (Bertsekas & Tsitsiklis, 1996; Sutton & Barto, 1998), and there is now growing evidence that mechanisms of this kind are, in fact, realized in the nervous system (Schultz, Dayan, & Montague, 1997). Our work suggests that adults could generate near-optimal plans in a reasonable period of time by exploration. Such plans would be sensitive to the current level of motor noise, a result that would be of significant benefit to the actor because action plans would be tailored to an individual's degree of motor error and also track any changes in motor error that might occur as the individual changes in size and strength. We hope that the present research gives additional impetus to the study of these learning mechanisms and that it further encourages the convergence of computational and experimental approaches that is instrumental to this enterprise.

**Acknowledgments**—The reported research was supported by National Science Foundation Grant 97-20345 and Grant JSMF 96-25 from the McDonnell-Pew Foundation for Cognitive Neuroscience. We would like to thank Andy Barto, Neil MacMillan, Sandy Pollatsek, and two anonymous reviewers for their very helpful comments on the manuscript.

## REFERENCES

- Aslin, R.N., Saffran, J.R., & Newport, E.L. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological Science*, *9*, 321–324.
- Aslin, R.N., & Salapatek, P. (1975). Saccadic localization of visual targets by the very young human infant. *Perception & Psychophysics*, *17*, 293–302.
- Bellman, R.E. (1961). *Adaptive control processes*. Princeton, NJ: Princeton University Press.
- Berthier, N.E. (1996). Learning to reach: A mathematical model. *Developmental Psychology*, *32*, 811–823.
- Bertsekas, D.P. (1987). *Dynamic programming: Deterministic and stochastic models*. Englewood Cliffs, NJ: Prentice Hall.
- Bertsekas, D.P., & Tsitsiklis, J.N. (1996). *Neuro-dynamic programming*. Belmont, MA: Athena Scientific.
- Calvin, W.H., & Stevens, C.F. (1968). Synaptic noise and other sources of randomness in motoneuron interspike intervals. *Journal of Neurophysiology*, *31*, 574–587.
- Chua, R., & Elliott, D. (1993). Visual regulation of manual aiming. *Human Movement Science*, *12*, 365–401.
- Clamann, H.P. (1969). Statistical analysis of motor unit firing patterns in a human skeletal muscle. *Biophysics Journal*, *9*, 1233–1251.
- Crossman, E.R.F.W., & Goodeve, P.J. (1983). Feedback control of hand movement and Fitts' Law. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology*, *35(A)*, 251–278.
- Dayan, P., & Abbott, L.F. (2001). *Theoretical neuroscience*. Cambridge, MA: MIT Press.
- Elliott, D., Helsen, W.F., & Chua, R. (2001). A century later: Woodworth's (1899) two-component model of goal directed aiming. *Psychological Bulletin*, *127*, 342–357.
- Engelbrecht, S.E. (2001). Minimum principles in motor control. *Journal of Mathematical Psychology*, *45*, 497–542.
- Ernst, M.O., & Saksida, L.M. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, *415*, 429–433.
- Fiser, J., & Aslin, R.N. (2001). Unsupervised statistical learning of higher-order spatial structures from visual scenes. *Psychological Science*, *12*, 499–504.
- Harris, C.M. (1995). Does saccadic undershoot minimize saccadic flight-time? A Monte-Carlo study. *Vision Research*, *35*, 691–701.
- Harris, C.M., & Wolpert, D.M. (1998). Signal-dependent noise determines motor planning. *Nature*, *394*, 780–784.
- Henson, D.B. (1978). Corrective saccades: Effects of altering visual feedback. *Vision Research*, *18*, 63–67.
- Hofsten, C., von. (1991). Structuring of early reaching movements: A longitudinal study. *Journal of Motor Behavior*, *23*, 253–270.
- Keele, S.W. (1968). Movement control in skilled motor performance. *Psychological Bulletin*, *70*, 387–403.
- Meyer, D.E., Abrams, R.A., Kornblum, S., Wright, C.E., & Smith, J.E.K. (1988). Optimality in human motor performance: Ideal control of rapid aimed movements. *Psychological Review*, *95*, 340–370.
- Meyer, D.E., Smith, J.E.K., & Wright, C.E. (1982). Models for the speed and accuracy of aimed movements. *Psychological Review*, *89*, 449–482.
- Schmidt, R.A., Zelaznik, H.N., Hawkins, B., Frank, J.S., & Quinn, J.T. (1979). Motor-output variability: A theory for the accuracy of rapid motor acts. *Psychological Review*, *86*, 415–451.
- Schultz, W., Dayan, P., & Montague, P.R. (1997). A neural substrate of prediction and reward. *Science*, *275*, 1593–1599.
- Sutton, R.S., & Barto, A.G. (1998). *Reinforcement learning*. Cambridge, MA: MIT Press.
- Toni, I., Gentilucci, M., Jeannerod, M.O., & Decety, J. (1996). Differential influence of the visual framework on end point accuracy and trajectory specification of arm movements. *Experimental Brain Research*, *111*, 447–454.
- Vercher, J.L., Magenes, G., Prablanc, C., & Gauthier, G.M. (1994). Eye-head-hand coordination in pointing at visual targets: Spatial and temporal analysis. *Experimental Brain Research*, *99*, 507–523.

(RECEIVED 4/10/02; REVISION ACCEPTED 8/31/02)