Optimization strategies in human reinforcement learning

Heiko Hoffmann, Evangelos Theodorou, and Stefan Schaal
University of Southern California, USA

Some human movement skills require optimizing a movement such that a future event has a desired outcome. Such skills are, e.g., hitting a ball with a bat or swinging a golf club to achieve that the ball has a desired trajectory. Learning a movement given only reward feedback is a typical reinforcement learning problem (Sutton and Barto, 1998). While several researchers studied reinforcement learning in robotics and machine learning, little is known about human reinforcement learning for movement skills. For example, we do not even know which learning strategy humans choose in one of the simplest reinforcement learning settings, i.e., with immediate reward feedback at the end of a movement.

Here, we investigate this question using a behavioral paradigm mimicking a ball-hitting task (Fig. 1 A). Subjects (n=10) sat in front of a computer screen and moved a stylus on a tablet towards an unknown target. This invisible target was located on a line that the subjects had to cross. Every subject did 100 movement trials. During each movement, visual feedback of the stylus position on the screen was suppressed. After the movement, a reward was displayed graphically as a colored bar. As reward, we used a Gaussian function of the distance between the target location and the point of line crossing. The choice of this function was inspired by the work of Koerding and Wolpert (2004), which suggested an inverted Gaussian loss function in sensorimotor tasks.

Subjects learned to adapt their movements towards the hidden target (Fig. 1 B and C). To investigate *how* they updated their movement choice, we hypothesize three optimization strategies:

1) Reward-weighted average (RW): $\tilde{x}_{i+1} = \frac{R_i x_i + R_{i-1} x_{i-1}}{R_i + R_{i-1}}$

2) Random search (RS): $\tilde{x}_{i+1} = \text{argmax}_{\{x_i, x_{i-1}\}} R(x)$

3) Gradient ascent (GA): $\tilde{x}_{i+1} = x_i + \eta \frac{R_i - R_{i-1}}{x_i - x_{i-1}}$

For simplicity, we assume subjects encode a movement with a single parameter, the point of line crossing, x_i. Thus, we assume the following scenario: at trial i, subjects choose a movement target $\tilde{x}_i$, experience a movement error ν_i, and observe the resulting movement $x_i = \tilde{x}_i + \nu_i$ and the corresponding reward $R_i(x_i)$. Based on these observations, subjects choose a new movement target $\tilde{x}_{i+1}$ according to one of the above strategies. Without the noise ν_i, only GA would converge to the goal if it is outside the interval $[x_{i-1}, x_i]$, i.e., the noise assists exploration of new solutions. The parameter η is the learning rate.

The above strategies make specific predictions on the dependence of the expectation value of $-(x_{i+1} - x_i)/(x_i - x_{i-1})$ on $R_{i-1}/(R_{i-1} + R_i)$, see Fig 2 A. Interestingly, only the prediction of RW was consistent with the data of all 10 subjects (Fig 2 B). We can further quantify this result. In the case $R_i > R_{i-1}$, the three different strategies predict distinct frequencies p of data points fulfilling $-(x_{i+1} - x_i)/(x_i - x_{i-1}) > 0$: for RW, $p > 0.5$, for RS, $p = 0.5$, and for GA, $p < 0.5$, under the assumption that the noise ν_i has mean 0. This distinction becomes intuitively clear by inspecting Fig 2 A for $R_{i-1}/(R_{i-1} + R_i) < 0.5$, and it can be proven. We computed p for each subject (Fig 2 C). The mean of p is significantly above 0.5 (t-test: p=0.005, Wilcoxon signed-rank test: p=0.01). For simplicity, we limited the update rule to a time window of two data points, x_{i-1} and x_i, but we can prove and show experimentally a similar distinction as above for larger window sizes.

The result that humans may prefer reward-weighted averaging over gradient ascent seems surprising. The literature on reinforcement learning is dominated by gradient-ascent methods. These methods are indeed preferable if the movement variance (noise) is low. However, for the same noise variance as observed in subjects, we found in simulation that reward-weighted averaging converges faster than gradient ascent. Thus, one could hypothesize that humans choose an optimization strategy that is the most suitable for their own movement variance.

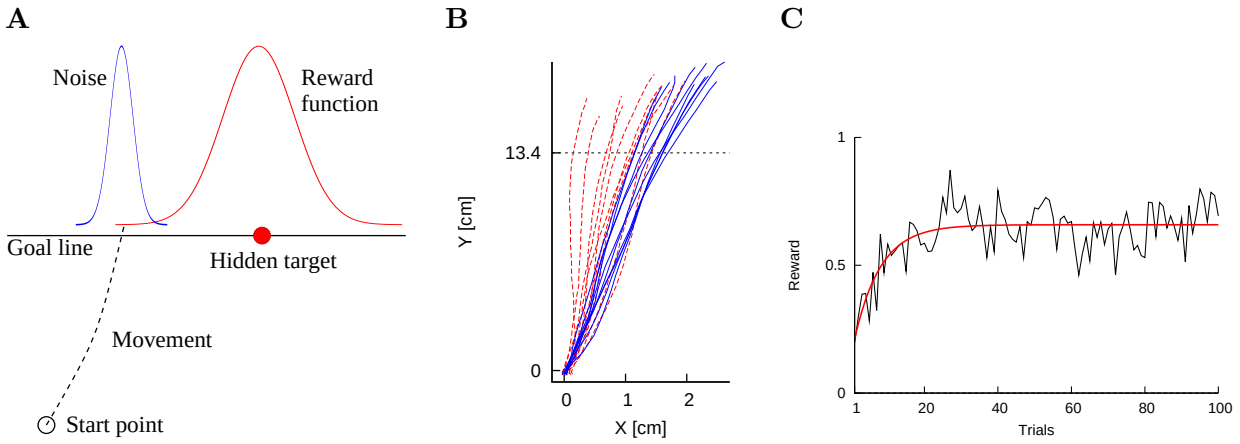

Figure 1: Experiment and raw data. **A**: Subjects move from a start point and need to cross a goal line. The only feedback is a reward signal at the end of a movement. This reward is a Gaussian function of the point of goal-line crossing. **B**: Movement adaptation through learning for a typical subject. The first 10 (dashed red) and the last 10 movements (solid blue) are shown. **C**: Trial-by-trial evolution of the reward, averaged across all 10 subjects. An exponential function is fitted to the data.

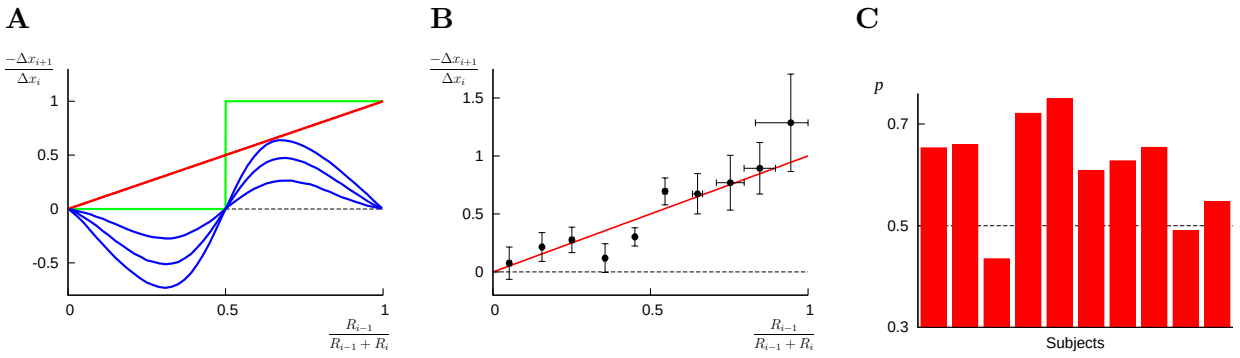

Figure 2: Comparison of optimization methods and results. **A**: Three optimization methods are compared: reward-weighted averaging (red), random search (green), and gradient ascent (blue). For gradient ascent, the results for three different learning rates are shown; these results were obtained from simulation. **B**: On the same graph, experimental results (error bars show standard errors) are compared to the prediction of reward-weighted averaging (red line). Before averaging, experimental results were binned into 10 intervals equally spaced along the x-axis. **C**: Probabilities $p = p(\frac{-\Delta x_{i+1}}{\Delta x_i} > 0 \,|\, R_i > R_{i-1})$ are shown for all subjects. Only for reward-weighted averaging, the average of p is expected to be above 0.5. The predicted p itself has variance, which depends on the number of movement trials for each subject.

References

Sutton, R S and Barto, A G (1998), Reinforcement learning: An introduction. MIT Press.

Körding, K P and Wolpert D M (2004), The loss function of sensorimotor learning. PNAS, 101, pp. 9839-9842.