

Human Optimization Strategies under Reward Feedback

Heiko Hoffmann, Evangelos A. Theodorou, and Stefan Schaal

Departments of Computer Science and Neuroscience, University of Southern California, Los Angeles, CA, USA

Motivation

How do humans learn if given at the end of a movement a continuous reward feedback?

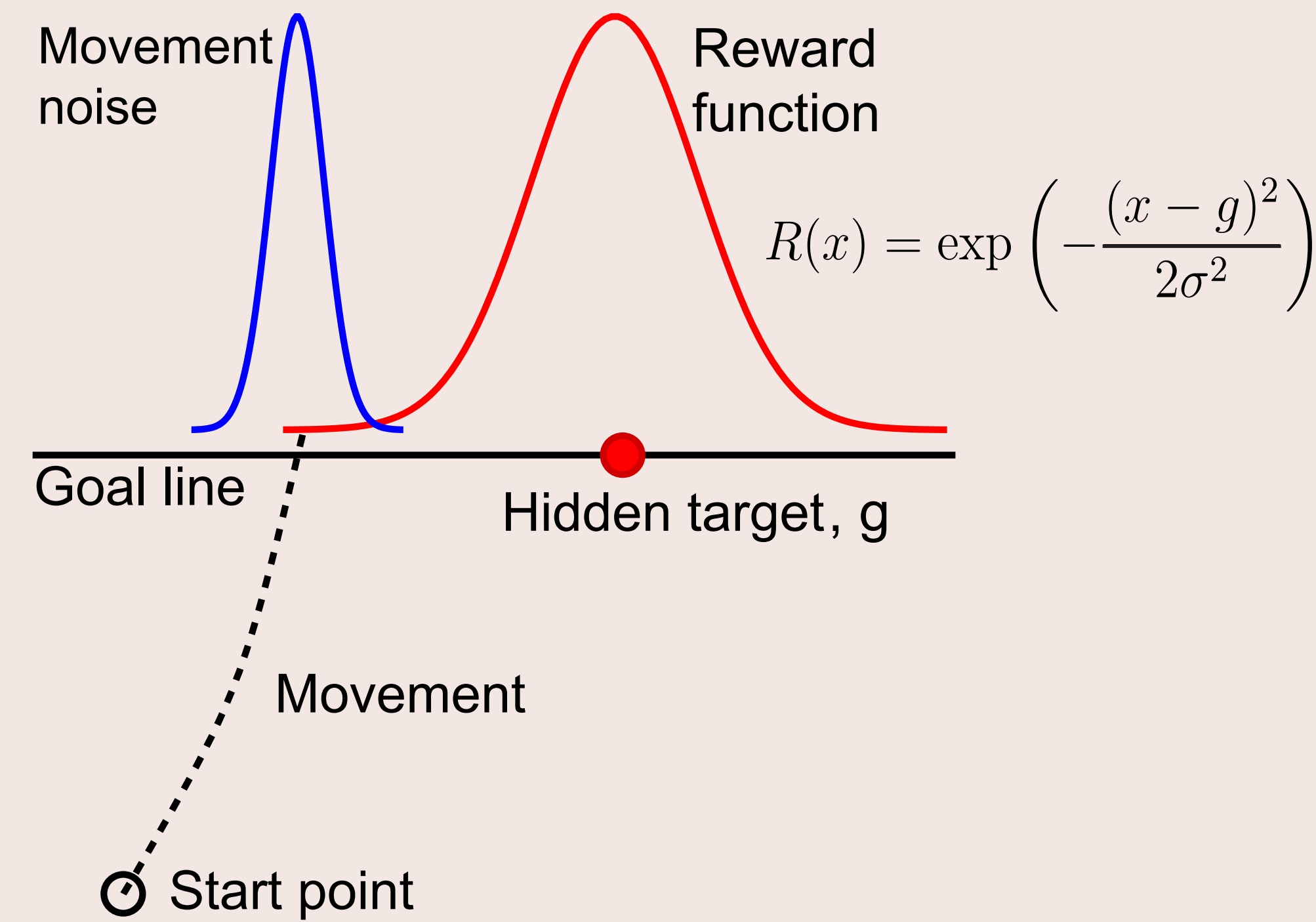
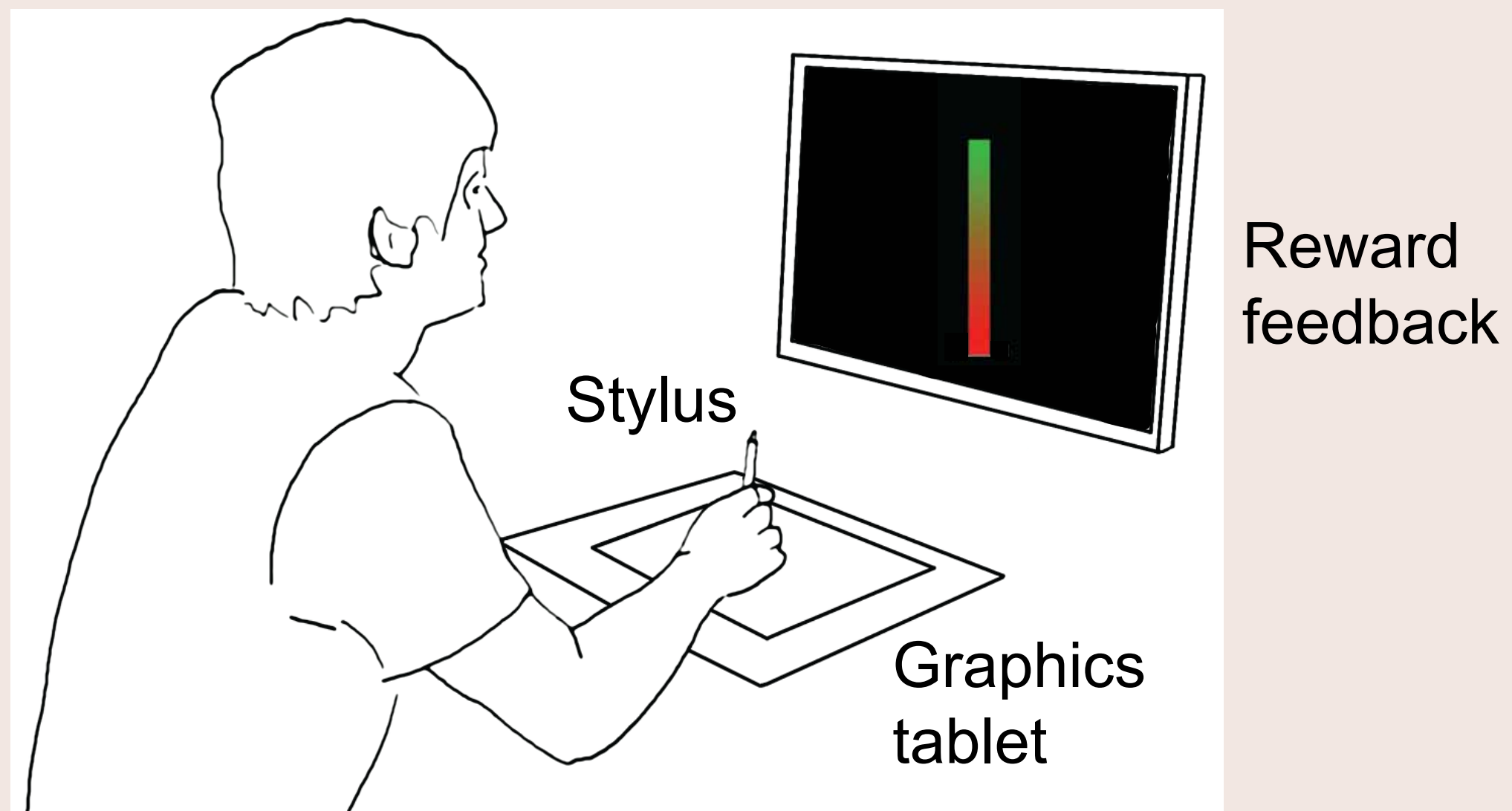
Lack of experiments studying movement-skill acquisition in a reward-learning setup.



Experiment 1

We designed an experiment that mimics a ball-hitting task. 12 naive, healthy, right-handed subjects participated.

The subjects move a stylus towards a goal line (shown on screen). Their task is to cross the line as close as possible to a target position, g . Unknown to the subjects, the actual position of the target differed from the visually presented one (± 1.46 cm in tablet coordinates). No visual feedback is provided after movement onset. Each subject did 100 trials.



We represent a movement with the point of line crossing, a one-dimensional control variable x .

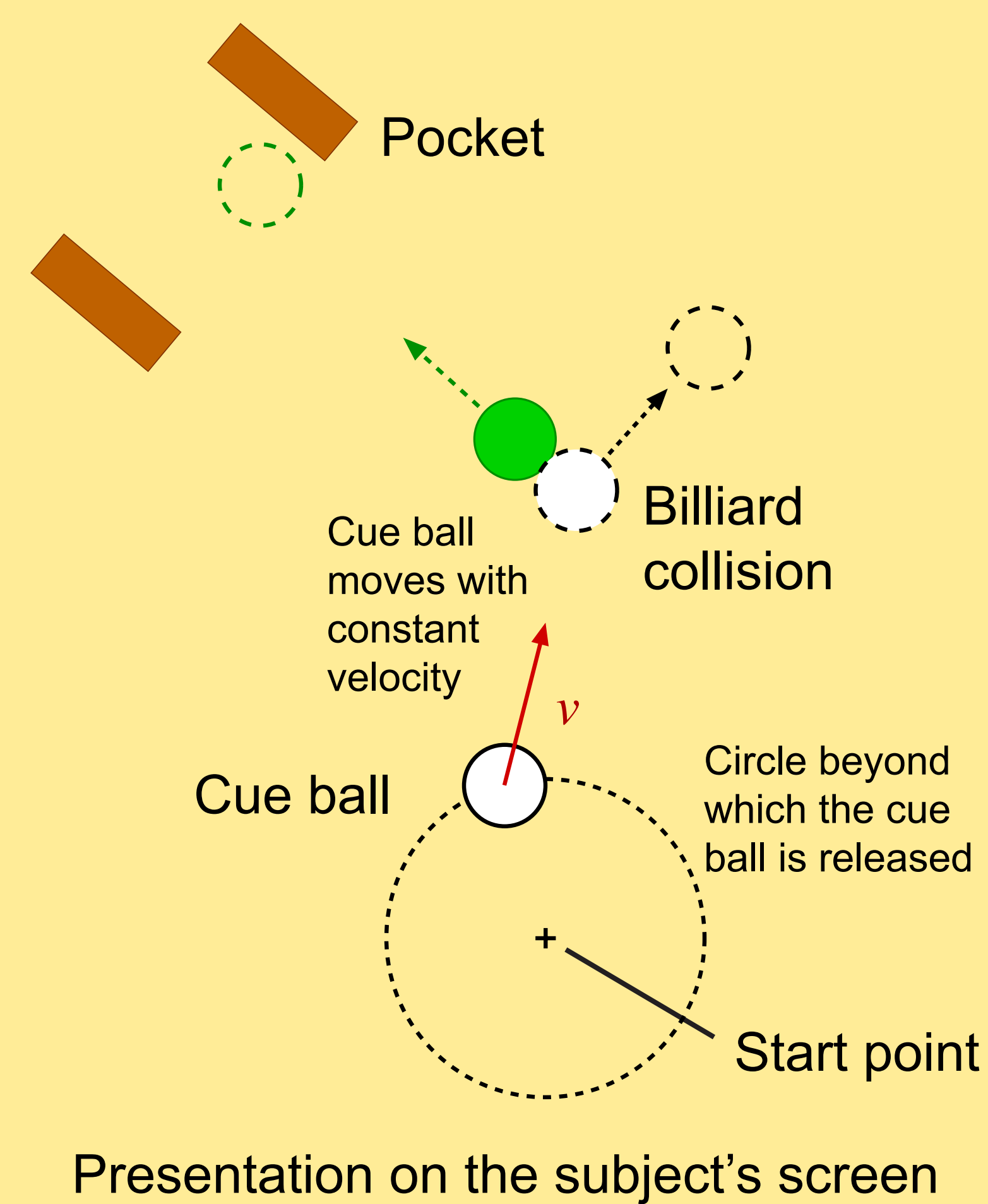
Experiment 2

The second experiment mimics pocket billiards. 8 naive, healthy, right-handed subjects participated.



The setup is the same as in experiment 1. The subjects' task is to hit with a cue ball a green ball such that the green ball hits the pocket. Initially, the cue ball is moved with the stylus until the ball crosses a circle shown on screen; then, the cue ball travels with constant speed.

Reward feedback is provided implicitly as the outcome of the billiard collision. The direction tolerance of the green ball for hitting the pocket is 10° .



Subjects effectively choose two control variables: the location on the circle and the velocity direction of the cue ball. Both parameters are put into a two-dimensional variable x .

Computational modeling

Optimization strategies

We hypothesize four strategies:

1) Reward-weighted averaging (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)
$$\tilde{x}_{i+1} = \operatorname{argmax}_{\{x_i, x_{i-1}\}} R(x)$$

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

4) Hebbian-like learning (HL)
$$\tilde{x}_{i+1} = x_i + \eta (R_i - R_{i-1})(x_i - x_{i-1})$$

Movement noise

After executing a movement, subjects experience a movement error.

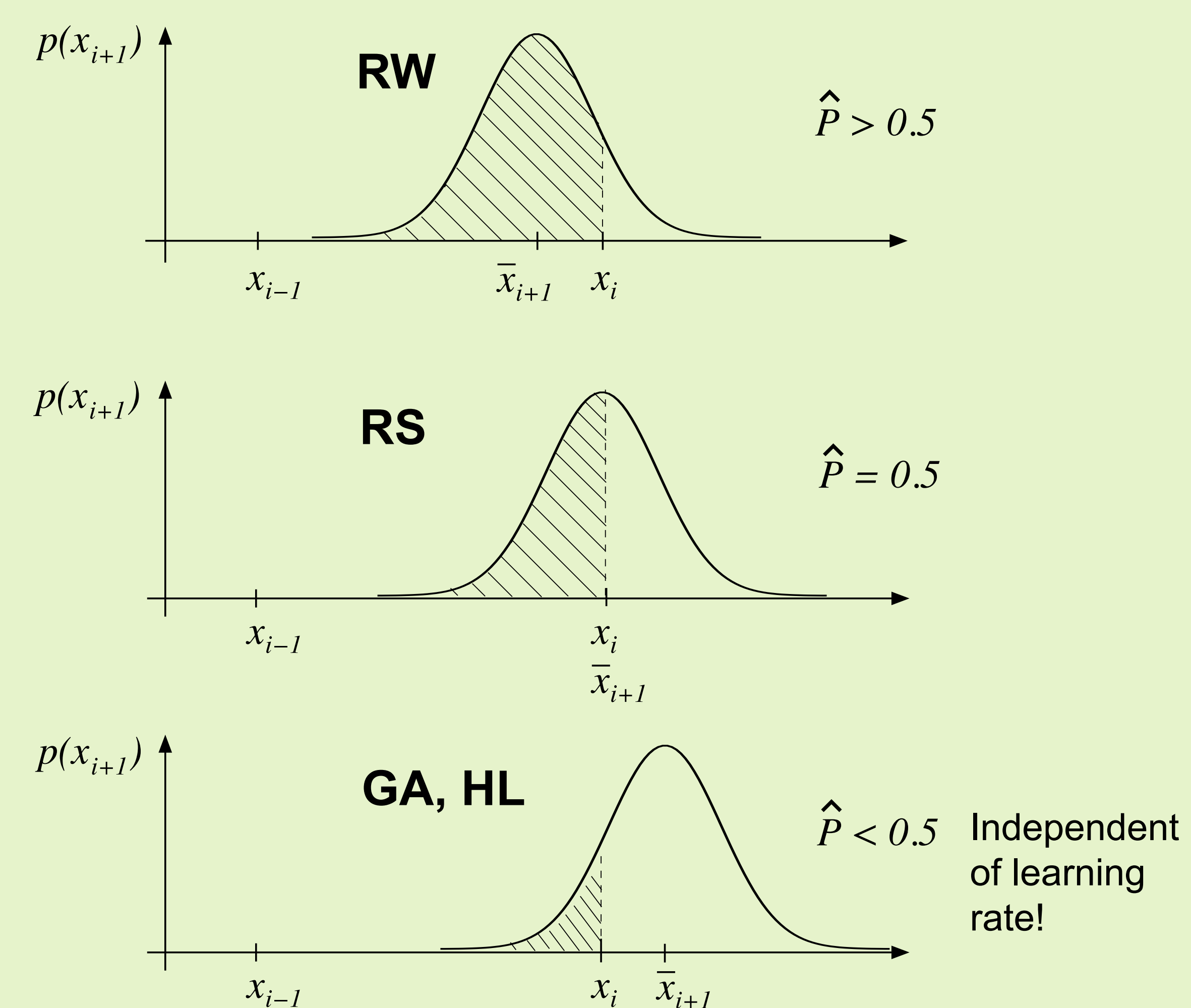
$$x_i = \tilde{x}_i + \text{noise}$$

Predictions

We compute the statistics of moving opposite to the gradient estimate.

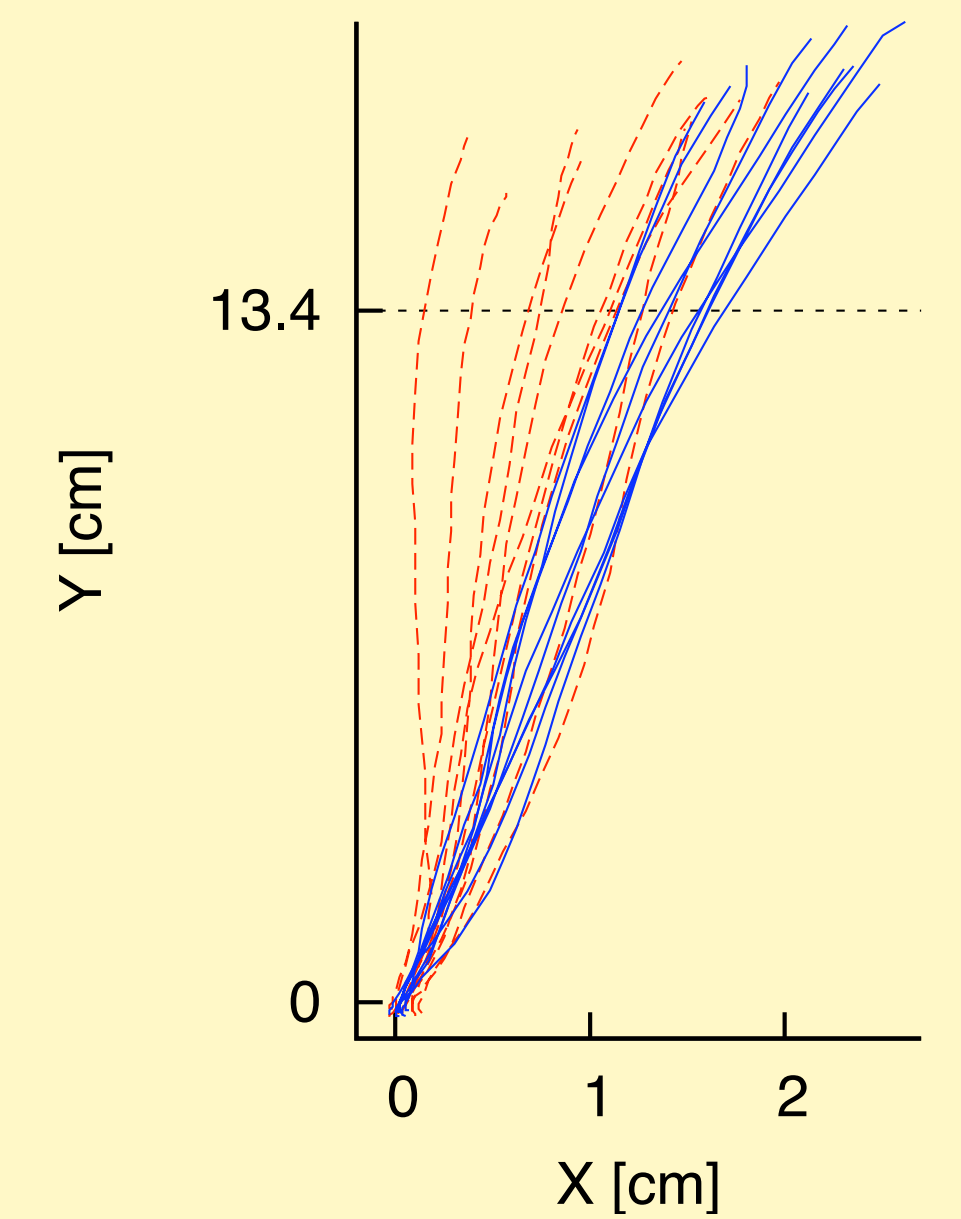
$$\hat{p} = p((x_{i+1} - x_i)^T (x_i - x_{i-1}) < 0 | R_i > R_{i-1})$$

The following graph illustrates this probability (shaded area) for each of the three strategies.

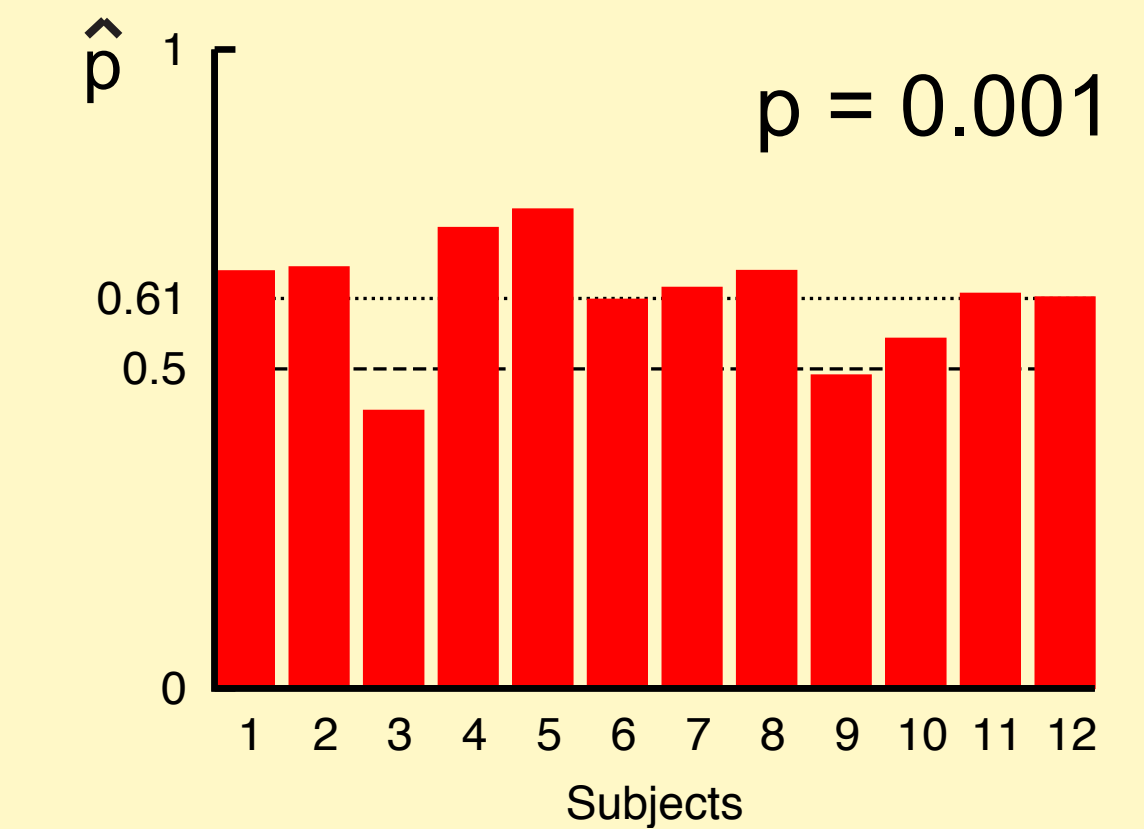
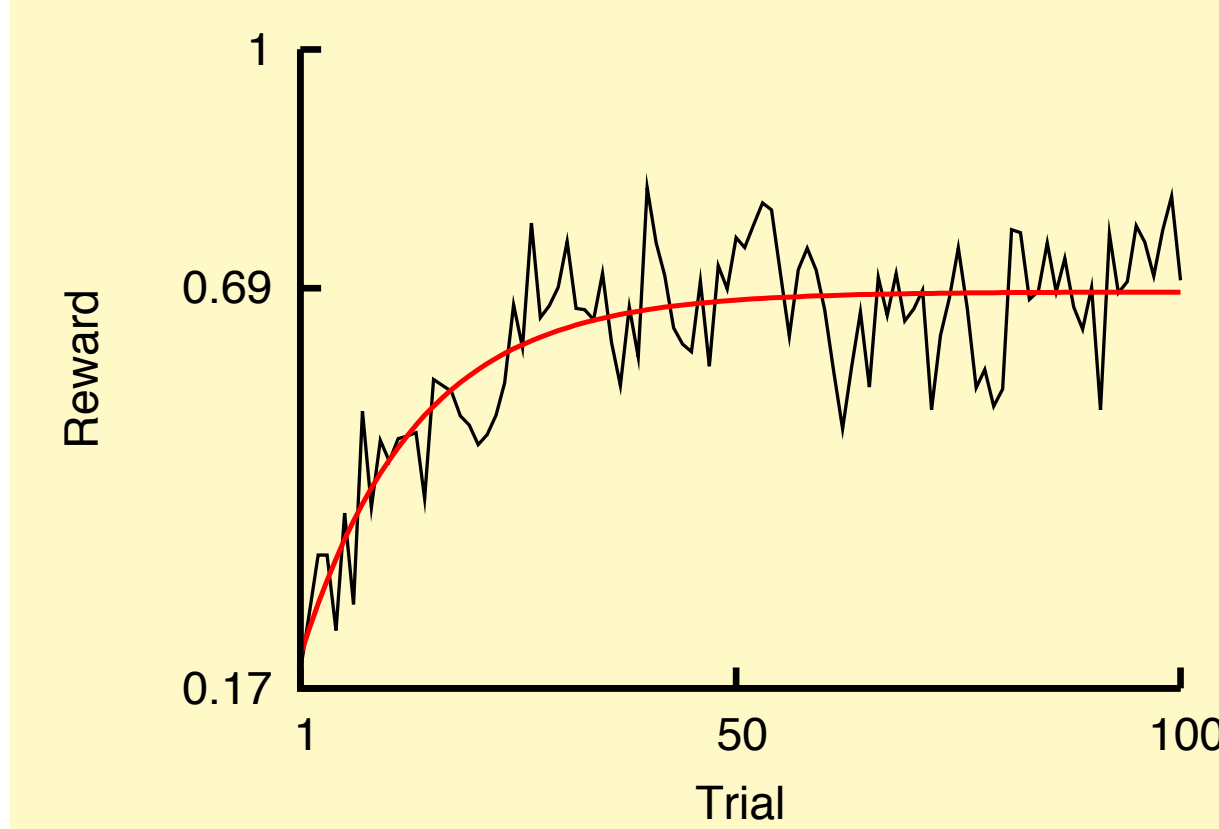


Results

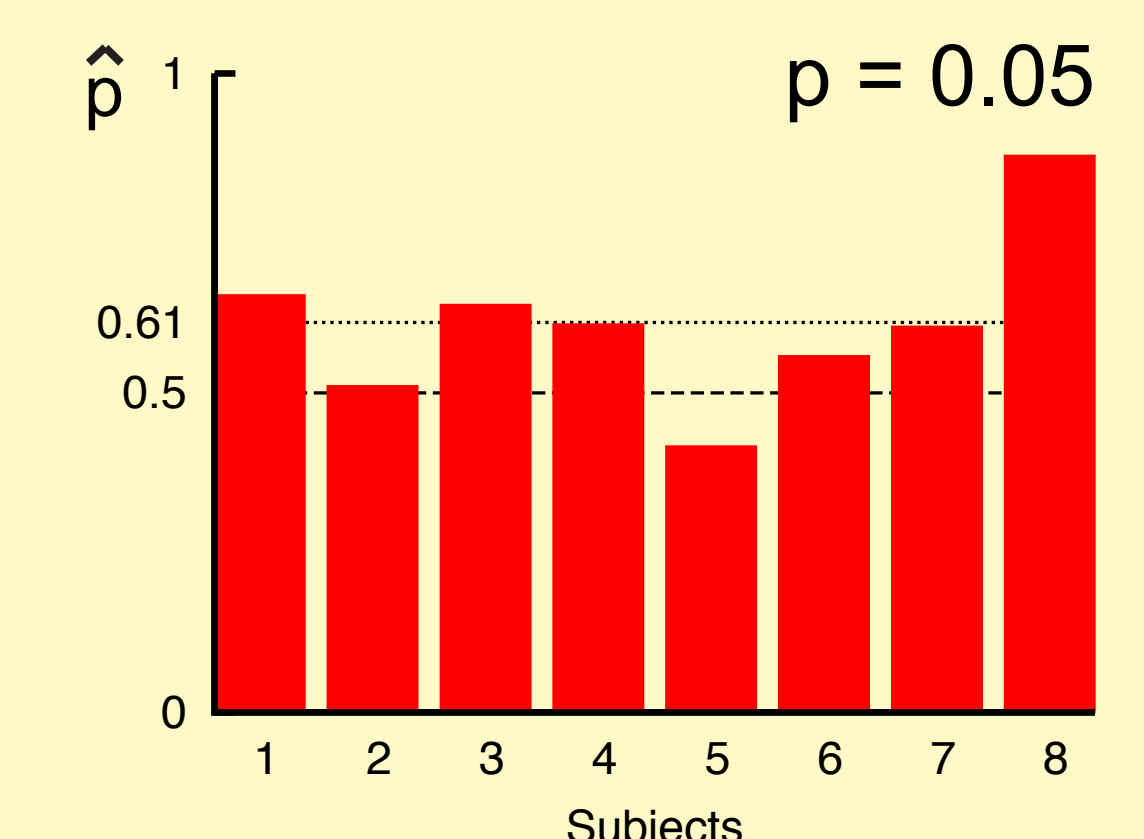
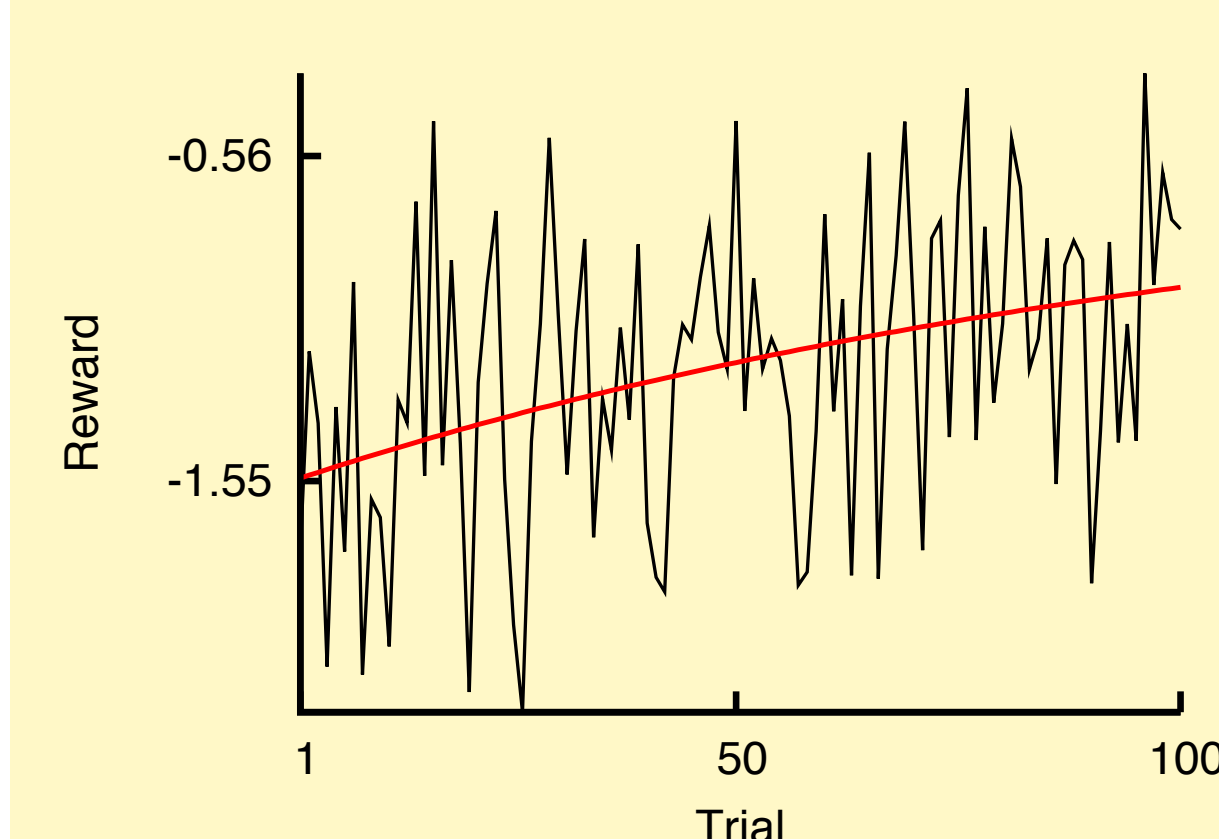
Subjects adapted their movements given only reward feedback. For the first experiment, the graph on the right shows sample movements of one subject, showing the first 20 (red) and the last 20 trials (blue). The following graphs show the average reward and, for each subject, the ratio \hat{p} .



Experiment 1



Experiment 2



The results are consistent with the predictions only for RW.

Conclusions

To the best of our knowledge, we report the first experiments showing learning from continuous reward feedback that was given at the end of a movement.

We were able to develop metrics that distinguish underlying optimization strategies from behavioral data.

Surprisingly, gradient ascent - as an optimization strategy - was incompatible with the results, while reward-weighted averaging predicted them.

This work was funded in part by DFG grant HO 3887/1-1 to HH.