Predicting Human Interactive Learning by Regret-Driven Neural Networks

Davide Marchiori, *et al.*

Science **319**, 1111 (2008);
DOI: 10.1126/science.1151185

Science (print ISSN 0036-8075; online ISSN 1095-9203) is published weekly, except the last week in December, by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005. Copyright 2008 by the American Association for the Advancement of Science; all rights reserved. The title *Science* is a registered trademark of AAAS.

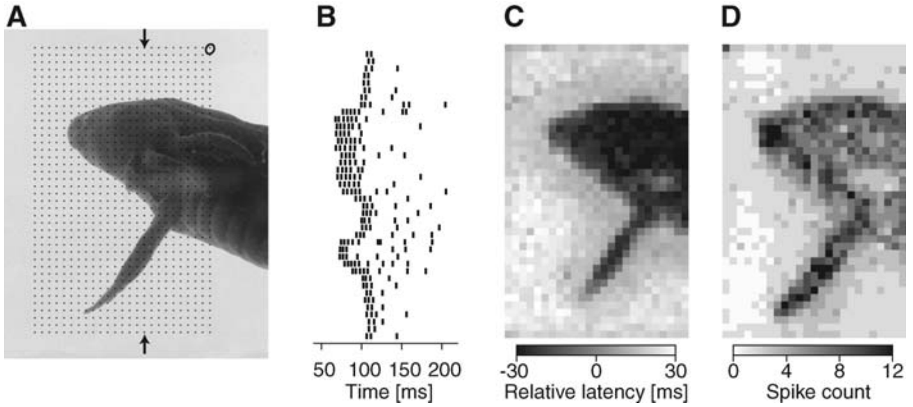

Fig. 4. Responses of a fast OFF ganglion cell to a flashed natural image. (For results from other cell types, see fig. S9.) (**A**) Photograph of a swimming salamander larva projected on the retina. The ellipse in the upper right corner shows a sample 1-SD outline of a ganglion cell receptive field. In each of 1000 presentations, the image was shifted slightly, and the grid of dots marks the resulting centers of the receptive field. Presentations were separated by gray illumination at the mean intensity of the photograph. The image onset produced luminance changes at most locations. (**B**) Spike trains of the ganglion cell for receptive-field locations along the column marked by the arrows in (A). (**C**) Gray-scale plot of the differential spike latency on single-trial presentations at the locations marked with dots in (A). The reference latency was chosen as the average value at all locations (*10*). (**D**) Corresponding gray-scale plot of the spike counts.

afferents (*30*), which is one possible readout mechanism for a latency code. Cortical neurons themselves carry substantial sensory information in their response latencies (*6*, *7*, *31*). Thus, it is conceivable that early aspects of sensory processing operate on the basis of the classification of spike latency patterns.

References and Notes

1. M. F. Land, *J. Comp. Physiol. A* **185**, 341 (1999).
2. C. Werner, W. Himstedt, *Zool. Jahrb. Abt. Allg. Zool. Physiol. Tiere* **89**, 359 (1985).
3. S. Thorpe, D. Fize, C. Marlot, *Nature* **381**, 520 (1996).
4. M. Meister, M. J. Berry, *Neuron* **22**, 435 (1999).
5. V. J. Uzzell, E. J. Chichilnisky, *J. Neurophysiol.* **92**, 780 (2004).
6. T. J. Gawne, T. W. Kjaer, B. J. Richmond, *J. Neurophysiol.* **76**, 1356 (1996).
7. D. S. Reich, F. Mechler, J. D. Victor, *J. Neurophysiol.* **85**, 1039 (2001).
8. M. Greschner, A. Thiel, J. Kretzberg, J. Ammermüller, *J. Neurophysiol.* **96**, 2845 (2006).
9. N. B. Sawtell, A. Williams, C. C. Bell, *Curr. Opin. Neurobiol.* **15**, 437 (2005).
10. S. M. Chase, E. D. Young, *Proc. Natl. Acad. Sci. U.S.A.* **104**, 5175 (2007).
11. J. J. Hopfield, *Nature* **376**, 33 (1995).
12. E. J. Chichilnisky, *Network* **12**, 199 (2001).
13. J. Keat, P. Reinagel, R. C. Reid, M. Meister, *Neuron* **30**, 803 (2001).
14. Materials and methods are available as supporting material on *Science* Online.
15. W. A. Hare, W. G. Owen, *J. Neurophysiol.* **76**, 2005 (1996).
16. S. A. Baccus, M. Meister, *Neuron* **36**, 909 (2002).
17. J. D. Victor, R. M. Shapley, *J. Gen. Physiol.* **74**, 671 (1979).
18. J. B. Demb, K. Zaghloul, L. Haarsma, P. Sterling, *J. Neurosci.* **21**, 7447 (2001).
19. M. N. Geffen, S. E. de Vries, M. Meister, *PLoS Biol.* **5**, e65 (2007).
20. F. S. Werblin, J. E. Dowling, *J. Neurophysiol.* **32**, 339 (1969).
21. F. M. de Monasterio, *J. Neurophysiol.* **41**, 1435 (1978).
22. F. R. Amthor, E. S. Takahashi, C. W. Oyster, *J. Comp. Neurol.* **280**, 97 (1989).
23. D. A. Burkhardt, P. K. Fahey, M. Sikora, *Vis. Neurosci.* **15**, 219 (1998).
24. J. L. Coombs, D. Van Der List, L. M. Chalupa, *J. Comp. Neurol.* **503**, 803 (2007).
25. J. F. Ashmore, D. R. Copenhagen, *Nature* **288**, 84 (1980).
26. X. L. Yang, *Prog. Neurobiol.* **73**, 127 (2004).
27. M. D. Menz, R. D. Freeman, *Nat. Neurosci.* **6**, 59 (2003).
28. B. Roska, F. Werblin, *Nat. Neurosci.* **6**, 600 (2003).
29. P. Kara, R. C. Reid, *J. Neurosci.* **23**, 8547 (2003).
30. W. M. Usrey, J. M. Alonso, R. C. Reid, *J. Neurosci.* **20**, 5461 (2000).
31. S. Panzeri, R. S. Petersen, S. R. Schultz, M. Lebedev, M. E. Diamond, *Neuron* **29**, 769 (2001).

32. We thank F. Engert and members of the Meister laboratory for advice. This work was supported by grants from the National Eye Institute (M.M.) and the Human Frontier Science Program Organization (T.G.).

Supporting Online Material

www.sciencemag.org/cgi/content/full/319/5866/1108/DC1
Materials and Methods
Figs. S1 to S10
References

23 August 2007; accepted 11 December 2007
10.1126/science.1149639

Predicting Human Interactive Learning by Regret-Driven Neural Networks

Davide Marchiori[1] and Massimo Warglien[2]*

Much of human learning in a social context has an interactive nature: What an individual learns is affected by what other individuals are learning at the same time. Games represent a widely accepted paradigm for representing interactive decision-making. We explored the potential value of neural networks for modeling and predicting human interactive learning in repeated games. We found that even very simple learning networks, driven by regret-based feedback, accurately predict observed human behavior in different experiments on 21 games with unique equilibria in mixed strategies. Introducing regret in the feedback dramatically improved the performance of the neural network. We show that regret-based models provide better predictions of learning than established economic models.

The surge of interest in the neural bases of economic behavior (*1*–*3*) prompts the question of how well neural networks can model human interactive decision-making (*4*). This question implies two issues: the choice of the network architecture and the selection of input information to the network that has to be both economically and neurophysiologically motivated.

Interactive learning differs from individual learning in that, given *n* agents, each agent adapts to behaviors that are modified by the concurrent learning of the other *n*–1 agents. It has an obvious relevance in economic contexts, but (more generally) much of human learning that occurs in social contexts has an interactive nature. Experimental game theory has provided a large set of laboratory data on human interactive learning in repeated games (*5*), often contradicting the predictions of standard game theory. The need for models of interactive learning in games arises from the difficulties of ordinary game-solution concepts to explain both the trajectories and the long-run stationary state of experimentally observed human behavior in repeated games. Games with unique equilibria in mixed strategies are an especially interesting case, because Nash equilibrium not only fails to approximate behavior in early rounds but also is often a poor predictor of the stable behavior emerging in the long run.

Until now, two main modeling strategies have been used with some success in trying to fit and predict how humans learn in repeated games in a laboratory setting. One modeling strategy extends a classical paradigm of learning theory (i.e., rein-

[1] Interdepartmental Center for Research Training in Economics and Management (CIFREM), University of Trento, Italy. [2] Advanced School of Economics and Department of Business Economics and Management, Ca' Foscari University, Venezia, Italy.

*To whom correspondence should be addressed. E-mail: warglien@unive.it

forcement learning) (6–8) to games. The second strategy builds hybrid models that blend reinforcement learning with modeling the evolution of a player's beliefs about other players' moves: The relative weight of both learning processes depends on parameters that can be tuned, in turn, by experience (*9*, *10*). More recently, a model emphasizing post-decision regret for foregone payoffs as the driver of learning has also been proposed (*11*). Interest in neuroeconomics suggests that a different modeling strategy might be explored, with the use of neural networks as models of human interactive behavior. We chose to keep the model as simple as possible, using (despite its well-known computational limitations) one of the most elementary learning neural network architectures: the simple (one-layered) analog perceptron (*12*, *13*). At the same time, we modified the feedback process to take into account some elementary economic considerations (in accordance with both theoretical insights and empirical evidence). Our basic assumption was that learning is driven by a sort of "ex-post" rationalizing process (*14*): Individuals modify their behavior by looking backward to what might have been their best move, once they know what the other individual's move was. They adjust in the direction of such an ex-post best response. Furthermore, we hypothesized that the intensity of such directional change is proportional to a measure of regret: how much they have missed by not playing such move (*15*, *16*). This is consistent with recent neuroscience research on individual decision-making, showing that regret affects learning and that both neurophysiological and behavioral responses to the experience of regret are correlated to its amplitude (*17*, *18*).

Our model maps the structure of a strategic game onto a neural network in a very straightforward way, by having an input node x_j corresponding to each payoff in the game matrix and by also including the opponent's payoffs and an output node y_i for each action available to a player *k* (Fig. 1). The input information is coded by having each input node take the value of the corresponding payoff in the current game; the output node activation is computed by summing up inputs to each output node weighted by the value of the incoming connections w_{ij} and transforming the summation via the hyperbolic tangent (tanh) activation function

$$y_i = \tanh\left(\beta \times \sum_j w_{ij} x_j\right) \qquad (1)$$

where β is the parameter tuning the steepness of the tanh function.

The activation values of the output nodes can be interpreted as propensities to play an action and are transformed into actual probabilities of play by normalization.

Thus far, this model is a very conventional, simple analog perceptron, where learning is modeled, as usual, as adaptive updating of the connections' weights. We adopted a variant of the Hopfield update rule (*12*, *13*), which provides a more direct probabilistic interpretation of the classical perceptron learning procedure

$$w_{ij}^t = w_{ij}^{t-1} + \Delta w_{ij} \qquad (2)$$

given the action *m* chosen by player *k*, a_m^k

$$\Delta w_{ij} = \lambda^2 \times [t_i(a^{-k}) - y_i] \times R^k(a_m^k, a^{-k}) \times x_j \qquad (3)$$

where $t_i(a^{-k})$ is the ex-post best response of player *k* to the other players actions a^{-k}; y_i is its propensity to play action *i*; $R^k(\cdot)$ is the regret given the action a_m^k and other players' actions a^{-k}; x_j is the strength of the input to the node and can be interpreted as payoff saliency; and λ is the learning rate. Regret is computed as the difference between the actual payoff received by a player *k* and the maximum payoff obtainable, given other players' actions. Thus, the psychological intuition underlying Eq. 3 is that connection weight adjustment is driven by a series of factors that can be summarized as adjustment = learning rate × distance from ex-post best response × regret × input saliency.

As compared with Hopfield's perceptron rule, the main difference of this variant is that the error feedback is multiplied by the regret size. One version of the model (henceforth PB1) squeezes the number of parameters to one by equating β to λ. This choice is justified because the effects of both parameters on the adjustment process are highly correlated (computer simulations have confirmed that this parameter trimming implies no substantial

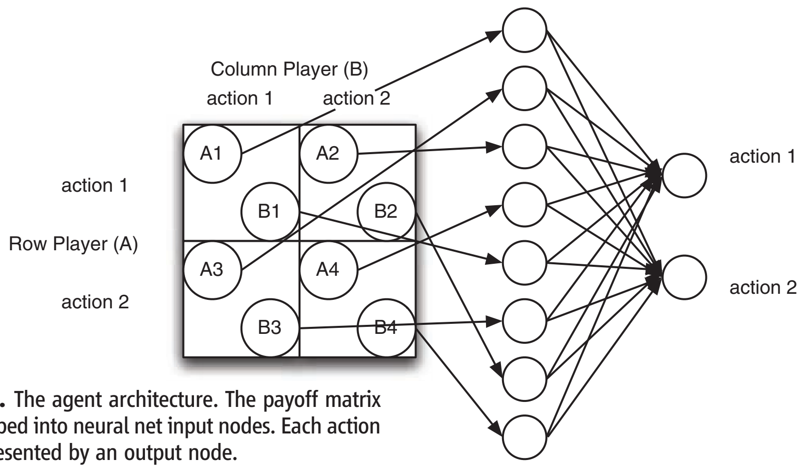

Fig. 1. The agent architecture. The payoff matrix is mapped into neural net input nodes. Each action is represented by an output node.

Average MSD scores (SD)	Average AICc scores (SD)									
2.45 (1.96)	–37.790 (18.13)	**NNET2**								
2.55 (1.78)	–41.380 (15.55)	**Random**	1, 15 17, 4							
1.65 (1.48)	–42.019 (15.93)	**REL**	13, 8 16, 6	15, 5 11, 10						
2.74 (2.85)	–42.313 (13.51)	**Nash Eq.**	8, 13 12, 9	10, 11 10, 11	8, 13 8, 13					
1.83 (1.09)	–43.464 (17.74)	**BR**	13, 8 16, 5	14, 6 11, 10	6, 14 13, 8	10, 9 11, 10				
1.07 (1.25)	–51.725 (18.33)	**stEWA**	15, 6 18, 3	16, 4 18, 2	17, 4 20, 1	14, 7 15, 6	18, 3 16, 5			
0.78 (0.66)	–52.550 (16.97)	**PB1**	16, 1 21, 0	21, 0 19, 2	17, 3 18, 2	18, 3 15, 6	20, 1 17, 4	14, 6 14, 7		
0.50 (0.35)	–52.681 (16.92)	**NFP**	20, 0 20, 1	21, 0 18, 3	19, 2 16, 5	18, 3 15, 6	20, 1 18, 3	18, 3 12, 9	15, 5 11, 10	
0.86 (0.75)	–54.660 (17.14)	**PB0**	18, 2 21, 0	20, 1 20, 1	16, 5 19, 2	17, 4 15, 6	19, 1 20, 1	12, 9 14, 7	8, 12 15, 6	4, 17 12, 9
			NNET2	**Random**	**REL**	**Nash Eq.**	**BR**	**stEWA**	**PB1**	**NFP**

Fig. 2. Results. The first and second columns indicate average MSD and average AICc scores, respectively, over all 21 games. The third column and bottom row indicate model names. Cells in the remaining columns contain two sets of paired values: The top pair indicates the number of tasks for which the MSD score of the "row" model was significantly better or worse (the first and second numbers in each pair, respectively) than that of the "column" model, and the bottom pair shows those relationships for AICc scores. As a result of ties, the sums of each pair in a cell may be less than 21.

losses in descriptive and predictive performances). We also introduced a second model (PB0), which has no free parameters as a result of a simple form of meta-learning (*10*, *19*), allowing the endogenous determination of the learning rate λ. Once more, regret plays a central role: We assumed that the learning rate is tuned by a cumulative regret function that increases λ as the current ratio of experienced regret to the maximum possible regret exceeds the average and decreases λ in the opposite case

$$\lambda_t^k = \frac{\sum_t R_t^k}{\sum_t \max(R_t^k)} \qquad (4)$$

where t is the number of iterations, R_t^k is the regret that is actually experienced by player k at round t, and $\max(R_t^k)$ is the maximum possible regret that player k could experience at time t (of course, in a repeated game, such value is constant).

To test the descriptive and predictive accuracy of our model, we considered experiments on 21 different games with unique equilibria in mixed strategies (*8*, *20*–*25*). Over more than 50 years, these experiments have been conducted by researchers other than the authors of this paper. The games have several actions, which range from two to five, that are available to each player. Of those games, 17 are constant sum games, whereas the remaining 4 are games in which players have no incentive to favor the other player: In other words, in each experiment, players had to learn strategies of conflict. In order to let learning processes unfold, we selected experiments with the constraint that there should be a minimum of 100 iterations of the stage game.

Our focus on such a class of games was motivated by multiple considerations. Because such games have unique equilibria, game theory lends a unique prediction of agents' behavior, providing a nonequivocal benchmark. Furthermore, all the 21 games that we considered have nondegenerate solutions: In equilibrium, subjects have to randomize their behavior. This is a source of cognitive complexity, despite the apparent simplicity of the game structures. Nash equilibrium turns out to be a poor predictor of observed behavior in such games (in many of them, it performs even worse than a "random behavior" prediction). Finally, this is the class of games on which the largest set of experiments with sufficient iterations is available.

We compared our model with different breeds of models (*7*): in particular, we took the Nash equilibrium, blind random behavior, and three of the most established learning models in the behavioral game-theory literature [i.e., the Basic Reinforcement Learning (BR) model (*7*), the Reinforcement Learning (REL) model by Erev *et al.* (*8*), and the self-tuning Experience-Weighted Attraction (stEWA) model (*10*)] as competitors, as well as the recent Normalized Fictitious Play (NFP) model (*8*, *11*). To single out the value added by introducing a regret term in the perceptron feedback, we further compared our model with the corresponding one-layer analog perceptron (NNET2) that uses the ordinary error feedback measure (dropping the regret term from Eq. 3) and has independent λ and β free parameters.

Given the availability of 21 different experimental conditions, it is appropriate to use—for each single condition (game)—the other 20 games to calibrate free parameters and predict behavior in the given condition (*6*, *7*, *26*). This provides 21 different predictions: one for each game. We used Mean Square Deviation (MSD) as a measure of goodness-of-fit for calibration and prediction (*27*). Although PB0 has no free parameters to estimate, PB1, reinforcement learning, and stEWA have one free parameter to estimate, and the REL model and NFP have two free parameters: Thus, the models to be compared differ both in functional form and in the number of free parameters. Among generally accepted criteria for comparing models with different complexity, we used the corrected Akaike's Information Criterion (AICc) (adjusted for sample size), which is the method according the lowest penalty to excess parameters.

Figure 2 shows the results of our analysis. The regret-based perceptrons PB1 and PB0 had the second- and third-best average MSD scores on all 21 prediction tasks, and in most conditions, were predicting better than other models, with the only exception of the other regret-based model, NFP (which was the best performer in terms of MSD). Notably, the no-parameter PB0 preserved much of the one-parameter PB1 performance, with average MSD scores much smaller not only than other nonparametric models but also than the parametric ones, with the obvious exception of PB1 and NFP. Once the number of free parameters was taken into account, however, the no-parameter perceptron PB0 had the lowest AICc score and compared favorably to all other models in most of the games. Thus, no matter how one measures performance, regret-based models always fared better than the other models, although PB0 gained an advantage from its great parsimony (*28*).

Furthermore, the PB0 and PB1 models clearly outperform the traditional NNET2 analog perceptron, demonstrating the determinant role played by the introduction of regret as a source of feedback for learning.

Another important advantage for models such as PB1 and PB0 comes from the nature of the learning tasks that can be modeled. Most human interactive learning happens in contexts where tasks do not repeat themselves identically over time, as in the experiments considered here. Generalization from examples and the learning of conditional behavior (different responses to different inputs) are natural features of human behavior. Standard models of economic learning (including the recent NFP) do not capture such features, because there is no way that they can model dependence of behavior from the perception of different game structures. On the contrary, even simple neural networks, such as those investigated here, can easily model generalization and conditional behavior and thus are open to the investigation of more realistic interactive learning tasks.

References and Notes

1. C. F. Camerer, *Science* **300**, 1673 (2003).
2. M. Kosfeld *et al.*, *Nature* **435**, 673 (2005).
3. N. D. Daw, J. P. O'Doherty, P. Dayan, B. Seymour, R. J. Dolan, *Nature* **441**, 876 (2006).
4. D. Sgroi, D. J. Zizzo, *Physica A* **375**, 717 (2007).
5. C. F. Camerer, *Behavioral Game Theory* (Princeton Univ. Press, Princeton, NJ, 2003).
6. A. Roth, I. Erev, *Games Econ. Behav.* **8**, 164 (1995).
7. I. Erev, A. Roth, *Am. Econ. Rev.* **88**, 848 (1998).
8. I. Erev, A. Roth, L. Slonim, G. Barron, *Econ. Theory* **33**, 29 (2007).
9. C. F. Camerer, T. H. Ho, *Econometrica* **67**, 827 (1999).
10. T. H. Ho, C. F. Camerer, J. K. Chong, *J. Econ. Theory* **133**, 177 (2007).
11. E. Ert, I. Erev, *J. Behav. Decision Making* **20**, 305 (2007).
12. J. Hertz, A. Krogh, R. G. Palmer, *Introduction to the Theory of Neural Computation* (Addison-Wesley, Redwood City, CA, 1991).
13. J. J. Hopfield, *Proc. Natl. Acad. Sci. U.S.A.* **84**, 8429 (1987).
14. R. Selten, R. Stöcker, *J. Econ. Behav. Organ.* **7**, 47 (1986).
15. R. Selten, K. Abbink, R. Cox, *Exp. Econ.* **8**, 5 (2005).
16. S. Hart, A. Mas-Collel, *Econometrica* **68**, 1127 (2000).
17. N. Camille *et al.*, *Science* **304**, 1167 (2004).
18. G. Coricelli *et al.*, *Nat. Neurosci.* **8**, 1255 (2005).
19. D. Lee, *Nature* **441**, 822 (2006).
20. P. Suppes, R. C. Atkinson, *Markov Learning Models for Multiperson Interactions* (Stanford Univ. Press, Stanford, CA, 1960).
21. D. Malcolm, B. Lieberman, *Psychon. Sci.* **12**, 373 (1965).
22. B. O'Neill, *Proc. Natl. Acad. Sci. U.S.A.* **84**, 2106 (1987).
23. A. Rapoport, R. B. Boebel, *Games Econ. Behav.* **4**, 261 (1992).
24. J. Ochs, *Games Econ. Behav.* **10**, 202 (1995).
25. J. Avrahami, W. Guth, Y. Kareev, *Theory Decision* **59**, 255 (2005).
26. J. R. Busemeyer, Y. M. Wang, *J. Math. Psychol.* **44**, 171 (2000).
27. R. Selten, *Exp. Econ.* **1**, 43 (1998).
28. In a recent comparison of different learning models (*7*), NFP was tested on 10 of the games included in our data set, estimating its parameters on a different set of data. In this case, NFP did not perform better than the other models, including REL and reinforcement learning, confirming that (with two free parameters) the choice of the data set on which parameters are estimated may be a source of instability of the predictive performance of the model.
29. I. Erev, J. Ochs, and B. O'Neill kindly provided to us their original data sets. Discussions with R. Selten were a major source of inspiration for the project. P. Pellizzari, C. Gilbert, the anonymous reviewers and the editor helped us to substantially improve the paper. Ministero dell'Università e della Ricerca projects Fondo per gli Investimenti della Ricerca di Base RBNE03A9A7 and Progetti di Ricerca di Interesse Nazionale 2005139342 provided financial support for the research.

Supporting Online Material

www.sciencemag.org/cgi/content/full/319/5866/1111/DC1
Materials and Methods
Figs. S1.1 to S1.10 and S2.1 to S2.37
Tables S1 to S7
References

1 October 2007; accepted 11 January 2008
10.1126/science.1151185