ELSEVIER

Neurocomputing 44–46 (2002) 951 – 956

NEUROCOMPUTING

www.elsevier.com/locate/neucom

Evolution of reinforcement learning in foraging bees: a simple explanation for risk averse behavior

Yael Niv[a,*], Daphna Joel[a], Isaac Meilijson[b], Eytan Ruppin[b]

[a] *Department of Psychology, Tel-Aviv University, Tel-Aviv, Israel*
[b] *School of Mathematical Sciences, Tel-Aviv University, Tel-Aviv, Israel*

Abstract

Reinforcement learning is a fundamental process by which organisms learn to achieve goals from their interactions with the environment. We use evolutionary computation techniques to derive (near-)optimal neuronal learning rules in a simple neural network model of decision-making in simulated bumblebees foraging for nectar. The resulting bees exhibit efficient reinforcement learning. The evolved synaptic plasticity dynamics give rise to varying exploration/exploitation levels and to the well-documented foraging strategy of risk aversion. This behavior is shown to emerge directly from optimal reinforcement learning, providing a biologically founded, parsimonious and novel explanation of risk-averse behavior. © 2002 Published by Elsevier Science B.V.

Keywords: Reinforcement learning; Bumble bees; Evolutionary computation; Risk aversion; Exploration/exploitation tradeoff

1. Introduction

Behavioral research indicates that reinforcement learning (RL) is a fundamental means by which experience changes behavior and by which both vertebrates and invertebrates learn to achieve goals from their interactions with the environment [9]. In RL, learning is contingent upon a scalar reinforcement signal which provides evaluative information about how good an action is in a certain situation, without providing an instructive cue to the most rewarding behavior. RL has been studied excessively in the

* Corresponding author.

E-mail addresses: yaeln@cns.tau.ac.il (Y. Niv), djoel@post.tau.ac.il (D. Joel), isaco@math.tau.ac.il (I. Meilijson), ruppin@math.tau.ac.il (E. Ruppin).

PII: S0925-2312(02)00496-4

ethology of foraging bumble-bees. Real [7] showed that when foraging for nectar in a field of blue and yellow artificial flowers yielding different amounts of nectar, bumblebees exhibited efficient RL, rapidly switching their preference for flower type when reward contingencies were switched between the flowers. The bees also manifested risk averse behavior, showing a strong preference for landing on constant rewarding flowers, as opposed to variably rewarding flowers yielding the same mean reward. Risk-averse behavior has also been demonstrated in other animals [2], and has traditionally been accounted for by hypothesizing a nonlinear subjective "utility function" for reward [8].

RL has attracted ample attention in computational neuroscience, yet a fundamental question regarding the underlying mechanism has not been sufficiently addressed, namely, *what are the optimal learning rules for maximizing reward in RL?* In this paper, we use evolutionary computation techniques to derive the optimal neuronal learning rules that give rise to efficient RL in uncertain environments. We further investigate the *behavioral strategies which emerge as a result of optimal RL.*

In a previous neural network (NN) model, Montague et al. [5] simulated bee foraging in a 3D arena of flowers, based on a neurocontroller modelled after an identified interneuron in the honeybee suboesophogeal ganglion [1]. While this model replicated Real's foraging results and provided a basic and simple NN architecture to solve RL tasks, many aspects of the model, first and foremost the handcrafted synaptic learning rule, were arbitrarily specified and their optimality questionable. Towards this end, we use a generalized and parameterized version of this model in order to determine the optimal synaptic learning rules for RL (with respect to maximizing nectar intake) using a genetic algorithm [4]. We define a general framework for evolving learning rules, which encompasses all heterosynaptic Hebbian learning rules, along with other characteristics of the learning dynamics, such as learning dependencies between modules.

2. The model

A simulated bee flies in a 3D arena, above a patch of 60 × 60 randomly scattered blue and yellow flowers. In each trial the bee descends from height 10, advancing in steps of 1 unit in any downward heading direction. The bee views the world through a cyclopean eye (10° cone view), and in each timestep decides whether to maintain the current heading direction or to change direction randomly, based on its visual inputs. Upon landing, the bee consumes the nectar in the chosen flower and another trial begins. A bee's life consists of 100 trials. *The evolutionary goal (the fitness criterion) is to maximize nectar intake.*

In the NN controlling the bee's flight (Fig. 1a), three modules contribute their input via modifiable synaptic weights to a linear neuron P, whose continuous-valued output is

$$P(t) = R(t) + \sum_{i \in \text{regular}} W_i X_i(t) + \sum_{i \in \text{differential}} W_i [X_i(t) - X_i(t-1)]. \tag{1}$$

The regular input module reports the percentage of the bee's field of view filled with yellow $[X_y(t)]$, blue $[X_b(t)]$ and neutral $[X_n(t)]$. The differential input module reports

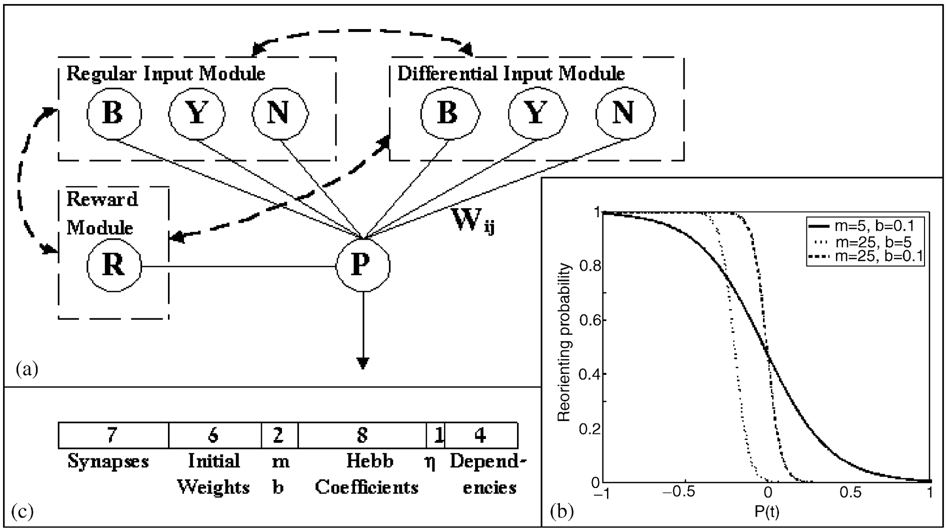

Fig. 1. (a) The bee's neural network controller. The weights $W_i(t)$ of the regular and differential modules are modifiable. (b) The bee's action function. Probability of reorienting direction of flight as a function of $P(t)$ for different values of evolvable parameters m, b. (c) The "genome" sequence of the simulated bee.

temporal differences of these percentages $[X_i(t) - X_i(t-1)]$. The reward module reports the actual amount of nectar received from a flower $[R(t)]$ in the nectar-consuming timestep (in which it is also assumed that there is no new input $[X_i(t) = 0]$), and zero during flight. The bee's action is determined according to the output $P(t)$ using Montague et al.'s probabilistic action function [5] shown in Fig. 1b. Note that *we do not incorporate any nonlinear utility function* with respect to the reward.

During the bee's "lifetime" the synaptic weights of the regular and differential modules are modified via a heterosynaptic Hebb learning rule of the form:

$$\Delta W_i(t) = \eta[AX_i(t)P(t) + BX_i(t) + CP(t) + D], \tag{2}$$

where η is a global learning rate parameter, $X_i(t)$ and $P(t)$ are the pre-synaptic and the post-synaptic values, respectively, W_i their connection weight, and A–D are real-valued evolvable parameters. In addition, learning in one module can be dependent on another module (dashed arrows in Fig. 1a), such that if module M depends on module N, M's synaptic weights will be updated according to Eq. (2) only if module N's neurons have fired. Thus the bee's "brain" is capable of *a non-trivial neuromodulatory gating of synaptic plasticity.*

The simulated bee's "genome" (Fig. 1c) consisted of a string of 28 genes, each representing a parameter governing the network architecture or learning dynamics. Using a genetic algorithm [4], the bees were "evolved" and the optimal gene values determined. A first generation of bees was produced by randomly generating 100 genome strings. Each bee performed 100 trials independently and received a fitness score according to the average amount of nectar gathered per trial. To form the next generation, 50 pairs of parents were chosen (with returns) with a bee's fitness specifying the probability of it being chosen as a parent. Each two parents gave birth to two offspring, which inherited their parents' genome after performing recombination and adding random mutations. The offspring were once again tested in the flower field. This process continued for a large number of generations. In each generation one of the two flower types was randomly assigned as a constant-yielding flower (containing 0.7 μl nectar), and the other a variable-yielding flower (1 μl nectar in $\frac{1}{5}$th of the flowers and zero otherwise). Reward contingencies were switched between the flower types in a randomly chosen trial during each bee's lifetime.

3. Results: evolution of RL

About half of the evolutionary runs were successful runs, in which reward-dependent choice behavior was evolved. These runs reveal that an interaction between eight genes governing the network structure and learning dependencies is essential for producing efficient learning in the bee's uncertain environment. We find that in our framework only a network architecture similar to that used by Montague et al. [5] can produce above-random foraging behavior, supporting their choice as an optimal one. However, *our optimized networks utilize a heterosynaptic learning rule different from that used by Montague et al., giving rise to several important behavioral strategies.*

In order to understand the evolved learning rule, we examined the foraging behavior of individual bees from the last generation of successful runs. In general, the bees manifest efficient RL, showing a marked preference for the high-mean rewarding flower, with a rapid transition of preferences after the reward contingencies are switched between the flower types. However, we find that there are individual differences between the bees in their degree of exploitation of the high-rewarding flowers versus exploration of the other flowers (Fig. 2). This phenomenon results from an *interesting relationship between the micro-level Hebb rule coefficients and the exploration/exploitation trade-off characteristic of the macro-level behavior.*

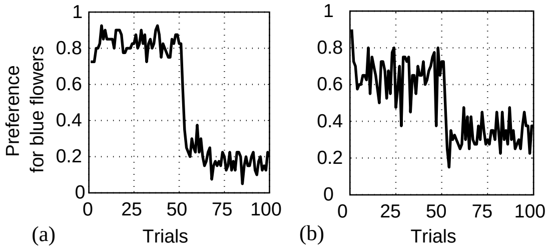

Fig. 2. Preference for blue flowers for two different bees from the last generation of different successful runs, averaged over 40 test bouts, each consisting of 100 trials. Blue is the initial constant-rewarding high-mean flower. Reward contingencies were switched between flower types at trial 50.

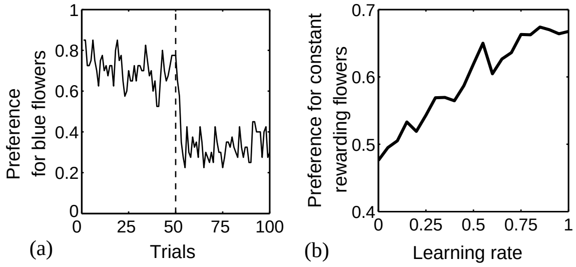

Fig. 3. (a) Risk aversion—preference for blue flowers in 100 test trials averaged over 40 previously evolved bees. Blue flowers contain $\frac{1}{2}$ μl nectar, yellow flowers contain 1 μl in half the flowers, contingencies switched at trial 50. (b) Risk aversion is ordered by learning rate. Percentage of visits to constant-rewarding flowers in 50 trials averaged over 40 bees, tested with a clamped learning rate.

According to the dependencies evolved, learning (synaptic updating) occurs primarily in the differential module, and only upon landing. We can thus analyze the effect of the heterosynaptic learning rule on the differential module's synaptic weights. This analysis reveals that as a result from the heterosynaptic nature of the learning rule, positive C and D values result in "spontaneous" strengthening of competing synapses, leading to an exploration-inclined bee. On the other hand, negative C and D values will result in exploitation-inclined behavior.

4. Emergence of risk aversion

A prominent foraging strategy exhibited by the evolved bees is risk-aversion. Fig. 3a shows the choice behavior of previously evolved bees, tested in a new environment where the mean rewards of the two flower types are identical. Although the situation does not call for any flower preference, the bees prefer the constant-rewarding flower. In contradistinction to the conventional explanations of risk aversion common in the fields of economics, our model does not include a non-linear utility function. *What hence brings about risk-averse behavior?* We argue that this behavior is a direct consequence of Hebbian learning dynamics in a n-armed-bandit-like RL situation.

In essence, due to the learning process, the bee makes its decisions based on finite time-windows, and does not compute the long-term mean reward obtained from each flower. This is even more pronounced with high learning rates, as after landing on an empty variable-rewarding type flower, the bee updates the reward expectation from this flower type (i.e. updates the corresponding synaptic weight according to the evolved heterosynaptic Hebb update rule) to near zero. As a result, the bee prefers the constantly rewarding flower, from which it constantly receives a reward. As long as the bee continues to choose the constant-rewarding flower, it will not update the expectation from the variable-rewarding flower, which will remain near zero. As has been suggested by March [3], Fig. 3b shows that higher learning rates lead to the more risk aversion. Corroborating these simulated results, in [6] we prove analytically that risk aversion is indeed a direct consequence of Hebbian learning dynamics in two-armed-bandit RL situation, and that risk aversion is ordered by learning rate.

5. Conclusions

In summary, we have presented a novel model of evolved reinforcement learning agents, which enables the concomitant study of both their macro (behavioral) and micro (dynamics) levels. We have shown that the evolved (near-) optimal synaptic learning rules control the tradeoff between exploration and exploitation seen in foraging behavior. We have further shown that optimal reinforcement learning can directly explain complex behaviors such as risk aversion, without need for additional assumptions.

References

[1] M. Hammer, Nature 366 (1993) 59–63.
[2] A. Kacelnik, M. Bateson, Amer. Zool. 36 (1996) 402–434.
[3] J.G. March, Psychol. Rev. 103 (2) (1996) 309–319.
[4] T. Mitchell, Machine Learning, McGraw Hill, New York, 1997.
[5] P.R. Montague, P. Dayan, C. Person, T.J. Sejnowski, Nature 377 (1995) 725–728.
[6] Y. Niv, D. Joel, I. Meilijson, E. Ruppin, Adaptive Behavior (2002), in press.
[7] L.A. Real, Science 30 (253) (1991) 980–985.
[8] P.D. Smallwood, Amer. Zool. 36 (1996) 392–401.
[9] R.S. Sutton, A.G. Barto, Reinforcement Learning: An Introduction, MIT Press, Cambridge, MA, 1998.