# Обучение и эволюция автономных адаптивных агентов*

Редько В.Г.[1)], Мосалов О.П. [2)], Прохоров Д.В. [3)]

[1)] Институт оптико-нейронных технологий РАН, Москва, redko@iont.ru
[2)]Московский физико-технический институт, olegmos_@mail.ru
[3)]Ford Research and Advanced Engineering, Ford Motor Company, Dearborn, U.S.A., dprokhor@ford.com

***Аннотация.*** Исследуется модель эволюции популяций самообучающихся агентов и анализируется взаимодействие между обучением и эволюцией. Рассматривается агент-брокер, который предсказывает изменения биржевого курса и использует эти предсказания для выбора действий. Система управления агента основана на нейросетевых адаптивных критиках. Проведено сравнение трех вариантов модели, в которых включены 1) либо обучение и эволюция одновременно, 2) либо отдельно эволюция, 3) либо отдельно обучение. Показано, что в данной модели может наблюдаться эффект Балдвина, т.е. первоначально приобретаемые навыки агентов в процессе эволюции становятся наследуемыми. Проведено сравнение поведения модельных агентов с поисковым поведением простейших животных.

# Learning and evolution of autonomous adaptive agents

Vladimir G. Red'ko[1)], Oleg P. Mosalov[2)], Danil V. Prokhorov[3)]

[1)] Institute of Optical Neural Technologies, Russian Academy of Science, redko@iont.ru
[2)] Moscow Institute of Physics and Technologies, olegmos_@mail.ru
[3)] Research and Advanced Engineering, Ford Motor Company, Dearborn, U.S.A., dprokhor@ford.com

**Abstract** We study a model of evolving populations of self-learning agents and analyze the interaction between learning and evolution. We consider an agent-broker that predicts stock price changes and uses its predictions for selecting actions. Each agent is equipped with a neural network adaptive critic design for behavioral adaptation. We discuss three cases in which either evolution or learning, or both, are active in our model. We show that the Baldwin effect can be observed in our model, viz., originally acquired adaptive policy of best agent-brokers becomes inherited over the course of the evolution. We also compare the behavioral tactics of our agents to the searching behavior of simple animals.

## 1. Введение. Автономные адаптивные агенты – новое направление в вычислительном интеллекте

Одно из новых и интересных направлений, которое развивается в последние годы в вычислительном интеллекте (Computational Intelligence), – исследование и применение автономных адаптивных агентов [1]. Такие агенты, подобно живым организмам, могут обладать собственными целями, собственными знаниями, формировать собственную политику поведения, выполнять те или иные действия, а также взаимодействовать с другими агентами. В связи с этим важно и интересно исследовать свойства автономных адаптивных агентов.

* Работа выполнена частичной поддержке программы Президиума РАН "Интеллектуальные компьютерные системы" (проект 2-45) и РФФИ (проект № 07-01-00180).

В настоящей работе исследуется модель эволюции популяции автономных адаптивных агентов, которые способны приобретать знания в процессе самообучения. Рассматриваемые агенты обучаются хорошо известным методом обучения с подкреплением (Reinforcement Learning) [2], в котором обучение происходит без учителя, а путем непосредственного взаимодействия агента с внешней средой. В зависимости от выполняемых агентом действий он получает подкрепления $r$ своих действий: поощрения ($r > 0$) или наказания ($r < 0$). При этом считается, что агент стремиться максимизировать суммарную награду $U$ (сумму значений $r$), которую он может получить в будущем.

Агенты имеют сравнительно простую «нервную систему», их система управления содержит два нейросетевых блока, которые предназначены 1) для прогноза будущих ситуаций **S** во внешней среде и 2) для оценки качества $V$(**S**) тех или иных ситуаций. Агенты оценивают качество $V$(**S**) какой-либо ситуации «субъективно», прогнозируя с помощью своей нейронной сети суммарную величину награды $U$, которую они могут получить в будущем, исходя из данной ситуации **S**.

Прогнозируя ситуации, в которые агент попадает при выполнении тех или иных действий, и оценки качества ситуаций, агент может выбирать действия таким образом, чтобы попадать в ситуации с наибольшими значениями качества. Тем самым агент будет стремиться получать награды и избегать наказаний.

Обучение агента состоит в том, что он, наблюдая ситуации, постепенно уточняет прогнозы ситуаций, а, получая те или иные поощрения и наказания, агент постепенно уточняет оценки качества ситуаций. Обучение происходит в результате модификации весов синапсов нейронных сетей агента.

Часть действий агент выбирает случайно, тем самым он имеет определенную поисковую активность и имеет возможность протестировать все новые и новые ситуации.

Детальнее схема обучения агентов излагается ниже.

Помимо обучения исследуемые в настоящей работе агенты эволюционируют, т.е. имеется популяция агентов и нейросетевые системы управления агентов «оптимизируются» еще и с помощью эволюции. А именно, предполагается, что начальные веса синапсов нейронных сетей, получаемые агентами при рождении, передаются от родителей к потомкам, испытывая при этом малые мутации. Причем, чем больше суммарная награда агента $U$, полученная им в течение жизни поколения, тем больше шансов имеет агент дать потомков в следующее поколение. А в течение жизни агенты еще и обучаются очерченным выше способом.

Исследование рассматриваемой модели проводится на примере агентов-брокеров, которые продают и покупают акции, стремясь максимизировать свой суммарный капитал.

Таким образом, нейросетевая система управления агентов адаптируется как за счет обучения, происходящего в течение жизни каждого поколения, так и в процессе эволюции, происходящей в течение ряда поколений. При этом критерий, по которому проводится та и другая адаптация, один и тот же – суммарная награда агента $U$.

Такая постановка сразу же вызывает ряд вопросов. Как будут взаимодействовать эти два вида адаптации: обучение и эволюция? Будут ли они способствовать друг другу или наоборот, препятствовать друг другу? Можно ли сопоставить поведение таких агентов с поведением простых животных?

Иллюстрируя эти вопросы, можно отметить, что такие исследования могли бы, в частности, пролить новый свет на весьма нетривиальный механизм ассимиляции приобретенных навыков, который был предложен более 100 лет назад Джеймсом Балдвином (1896 год) [3].

Согласно эффекту Балдвина приобретенные при обучении навыки организмов могут косвенно передаваться последующим поколениям. Эффект Балдвина работает в два этапа. На первом этапе некоторые эволюционирующие организмы (благодаря соответствующим мутациям) приобретают свойство обучиться некоторому полезному навыку. Приспособленность таких организмов увеличивается, следовательно, они распространяются по популяции. Но обучение имеет свои «накладные расходы», так как оно требует энергии и времени. Поэтому возможен второй этап (который называют генетической ассимиляцией): приобретенный полезный навык может быть «повторно изобретен» генетической эволюцией, в результате чего он записывается непосредственно в геном и становится наследуемым. Второй этап длится множество поколений, устойчивая окружающая среда и высокая корреляция между генотипом и фенотипом облегчают этот этап. Таким образом, полезный навык, который был первоначально приобретен, может стать наследуемым, хотя эволюция имеет дарвиновский характер.

Будет ли работать эффект Балдвина в развиваемой здесь модели? Забегая вперед, отметим, что эффект Балдвина в нашей модели действительно проявляется, но не всегда. Кроме того, в изложенной ниже модели возникает ряд нетривиальных особенностей взаимодействия обучения и эволюции: эволюция может «задавливать» обучение, несовершенное обучение может способствовать более эффективному эволюционному поиску, неправильный прогноз может использоваться для формирования вполне адаптивных действий и т.п. Мы также сопоставляем поведение наших агентов с поведением простых биологических организмов и находим режимы адаптации, в которых агенты способны адекватно использовать только общие закономерности изменений во внешней среде и игнорировать несущественные детали, аналогично тому, как это реально делают животные при поисковом поведении.

Далее статья организована следующим образом. В разделе 2 кратко характеризуется метод обучения с подкреплением. Раздел 3 содержит формальное описание модели. В разделе 4 излагаются результаты компьютерного моделирования. Раздел 5 содержит выводы по работе.

## 2. Обучение с подкреплением

Теория обучения с подкреплением (Reinforcement Learning) была развита в работах Р. Саттона и Э. Барто (Массачусетский университет).

Общая схема обучения с подкреплением [2] показана на рис. 1.

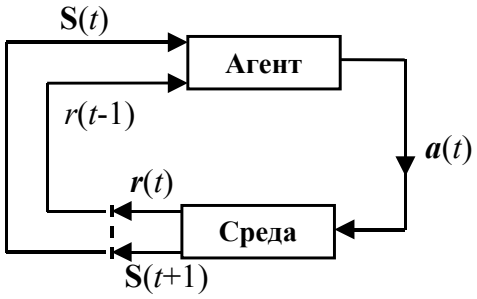


Рис. 1. Схема обучения с подкреплением.

Рассматривается агент, взаимодействующий с внешней средой. Время предполагается дискретным: $t = 1,2,\ldots$ В текущей ситуации агент $\mathbf{S}(t)$ выполняет действие $a(t)$, получает

подкрепление $r(t)$ и попадает в следующую ситуацию $\mathbf{S}(t+1)$. Подкрепление может быть положительным (награда) или отрицательным (наказание).

Цель агента – максимизировать суммарную награду, которую можно получить в будущем в течение длительного периода времени. Подразумевается, что агент может иметь свою внутреннюю "субъективную" оценку суммарной награды и в процессе обучения постоянно совершенствует эту оценку. Эта оценка определяется с учетом коэффициента забывания:

$$U(t) = \sum_{j=0}^{\infty} \gamma^j r(t+j), \qquad t = 1,2,..., \tag{1}$$

где $U(t)$ – оценка суммарной награды, ожидаемой после момента времени $t$, $\gamma$ – коэффициент забывания (дисконтный фактор), $0 < \gamma < 1$. Коэффициент забывания учитывает, что чем дальше агент «заглядывает» в будущее, тем меньше у него уверенность в оценке награды («рубль сегодня стоит больше, чем рубль завтра»).

В процессе обучения агент формирует *политику* (стратегию поведения). Политика определяет выбор (детерминированный или вероятностный) действия в зависимости от ситуации.

Метод обучения с подкреплением идейно связан с методом динамического программирования, и в том и другом случае общая оптимизация многошагового процесса принятия решения происходит путем упорядоченной процедуры одношаговых оптимизирующих итераций, причем оценки эффективности тех или иных решений, соответствующие предыдущим шагам процесса, переоцениваются с учетом знаний о возможных будущих шагах. Например, при решении задачи поиска оптимального маршрута в лабиринте от стартовой точки к определенной целевой точке сначала находится конечный участок маршрута, непосредственно приводящий к цели, а затем ищутся пути, приводящие к конечному участку, и т.д. В результате постепенно прокладывается трасса маршрута от его конца к началу. Обучение с подкреплением, адаптивные критики и подобные методы часто называют приближенным динамическим программированием [4].

Важное достоинство метода обучения с подкреплением – его простота. Т.е. агент получает от учителя или из внешней среды только сигналы подкрепления $r(t)$. Здесь учитель поступает с обучаемым объектом примитивно: "бьет кнутом" (если действия объекта ему не нравятся, $r(t) < 0$ ), либо "дает пряник" (в противоположном случае, $r(t) > 0$), не объясняя обучаемому объекту, как именно нужно действовать. Это радикально отличает этот метод от таких традиционных в теории нейронных сетей методах обучения, как метод обратного распространения ошибок, для которого учитель точно определяет, что должно быть на выходе нейронной сети при заданном входе.

Метод обучения с подкреплением был исследован рядом авторов (см. подробную библиографию в [2]) и был использован многочисленных приложениях. В частности, применения этого метода включают в себя:

- оптимизацию игры в триктрак (достигнут уровень мирового чемпиона);
- оптимизацию системы управления работы лифтов;
- формирование динамического распределения каналов для мобильных телефонов;
- оптимизацию расписания работ на производстве.

Отметим, что метод обучения с подкреплением может рассматриваться как развитие автоматной теории адаптивного поведения, разработанной в работах М.Л. Цетлина и его последователей [5].

Важная ветвь работ обучению с подкреплением разработки нейросетевых адаптивных критиков, в которых используются нейросетевые аппроксиматоры функций оценки качества функционирования агента [6,7]. Настоящая модель использует одну из простейших схем адаптивных критиков.

## 3. Описание модели

В работе исследуется модель эволюции популяции самообучающихся автономных агентов и анализируется взаимодействие между обучением и эволюцией. Система управления отдельного агента основана на нейросетевых адаптивных критиках [6,7]. Модель отрабатывается на примере агента-брокера.

**3.1. Схема агента-брокера.** Следуя [8], рассматриваем агента-брокера, который имеет ресурсы двух типов: деньги и акции; сумма этих ресурсов составляет капитал агента $C(t)$; доля акций в капитале равна $u(t)$. Внешняя среда определяется временным рядом $X(t)$, $t = 0,1,2,...$, $X(t)$ – курс акций на бирже в момент времени $t$ . Агент стремится увеличить свой капитал $C(t)$, изменяя значение $u(t)$. Динамика капитала определяется выражением [8]:

$$C(t+1) = C(t)\ \{1 + u(t+1)\ \Delta X(t+1) / X(t)\}\ [1 - J\ |u(t+1) - u(t)|], \qquad (2)$$

где $\Delta X(t+1) = X(t+1) - X(t)$ – текущее изменение курса акций, $J$ – параметр, учитывающий расходы агента на покупку/продажу акций. Следуя [9], используем логарифмическую шкалу для ресурса агента, $R(t) = \log C(t)$. Текущее подкрепление агента $r(t) = R(t+1) - R(t)$ равно:

$$r(t) = \log\ \{1 + u(t+1)\ \Delta X(t+1) / X(t)\} + \log\ [1 - J\ |u(t+1) - u(t)|]. \qquad (3)$$

Полагаем, что переменная $u$ может принимать только два значения $u = 0$ (весь капитал в деньгах) или $u = 1$ (весь капитал в акциях).

**3.2. Алгоритм обучения.** Система управления агента представляет собой простой адаптивный критик, состоящий из двух нейронных сетей (НС): Модель и Критик (рис. 2). Цель адаптивного критика – максимизировать функцию полезности $U(t)$, определяемую выражением (1).

Делая разумное предположение $\Delta X(t) << X(t)$, полагаем, что ситуация $\mathbf{S}(t)$, характеризующая состояние агента, зависит только от двух величин, $\Delta X(t)$ и $u(t)$: $\mathbf{S}(t) = \{\Delta X(t), u(t)\}$.

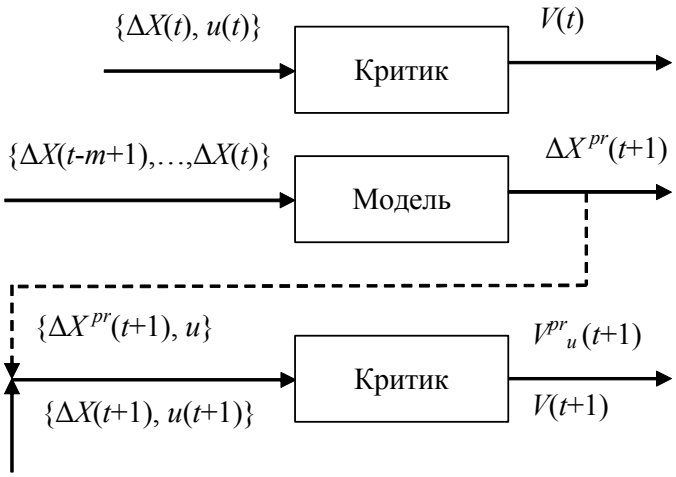


Рис. 2. Схема системы управления агента. НС Критика показана для двух последовательных тактов времени. Модель предназначена для прогнозирования изменения курса временного ряда. Критик предназначен для оценки качества ситуаций $V(\mathbf{S})$ для текущей ситуации $\mathbf{S}(t) = \{\Delta X(t), u(t)\}$, для ситуации в следующий такт времени $\mathbf{S}(t+1)$ и для предсказываемых ситуаций для обоих возможных действий $\mathbf{S^{pr}}_u(t+1) = \{\Delta X^{pr}(t+1), u\}$, $u = 0$ либо $u = 1$.

Модель предназначена для прогнозирования изменения курса временного ряда. На вход Модели подается $m$ предыдущих значений изменения курса $\Delta X(t-m+1)$, … , $\Delta X(t)$, на выходе формируется прогноз изменения курса в следующий такт времени $\Delta X^{pr}(t+1)$. Модель представляет собой двухслойную НС, работа которой описывается формулами:

$$\mathbf{x^M} = \{\Delta X(t-m+1),\ldots,\Delta X(t)\}, \qquad y^M_j = \text{th}\,(\textstyle\sum_i w^M_{ij}\, x^M_i), \quad \Delta X^{pr}(t+1) = \textstyle\sum_j v^M_j\, y^M_j,$$

где $\mathbf{x^M}$ – входной вектор, $\mathbf{y^M}$ – вектор выходов нейронов скрытого слоя, $w^M_{ij}$ и $v^M_j$ – веса синапсов НС.

Критик предназначен для оценки качества ситуаций $V(\mathbf{S})$, а именно, оценки функции полезности $U(t)$ (см. формулу (3)) для агента, находящегося в рассматриваемой ситуации $\mathbf{S}$. Критик представляет собой двухслойную НС, работа которой описывается формулами:

$$\mathbf{x^C} = \mathbf{S}(t) = \{\Delta X(t), u(t)\}, \qquad y^C_j = \text{th}\,(\textstyle\sum_i w^C_{ij}\, x^C_i), \quad V(t) = V(\mathbf{S}(t)) = \textstyle\sum_j v^C_j\, y^C_j,$$

где $\mathbf{x^C}$ – входной вектор, $\mathbf{y^C}$ – вектор выходов нейронов скрытого слоя, $w^C_{ij}$ и $v^C_j$ – веса синапсов НС.

Каждый момент времени $t$ выполняются следующие операции:

1) Модель предсказывает следующее изменение временного ряда $\Delta X^{pr}(t+1)$.

2) Критик оценивает величину $V$ для текущей ситуации $V(t) = V(\mathbf{S}(t))$ и для предсказываемых ситуаций для обоих возможных действий $V^{pr}_u(t+1) = V(\mathbf{S^{pr}}_u(t+1))$, где $\mathbf{S^{pr}}_u(t+1) = \{\Delta X^{pr}(t+1), u\}$, $u = 0$ либо $u = 1$.

3) Применяется $\varepsilon$-жадное правило [2]: действие, соответствующее максимальному значению $V^{pr}_u(t+1)$ выбирается с вероятностью $1 - \varepsilon$, и альтернативное действие выбирается с вероятностью $\varepsilon$ ($0 < \varepsilon << 1$). Выбор действия есть выбор величины $u(t+1)$: перевести весь капитал в деньги, $u(t+1) = 0$; либо в акции, $u(t+1) = 1$.

4) Выбранное действие $u(t+1)$ выполняется. Происходит переход к моменту времени $t+1$. Подсчитывается подкрепление $r(t)$ согласно (3). Наблюдаемое значение $\Delta X(t+1)$ сравнивается с предсказанием $\Delta X^{pr}(t+1)$. Веса НС Модели подстраиваются так, чтобы минимизировать ошибку предсказания методом обратного распространения ошибки. Скорость обучения Модели равна $\alpha_M > 0$.

5) Критик подсчитывает $V(t+1) = V(\mathbf{S}(t+1))$; $\mathbf{S}(t+1) = \{\Delta X(t+1), u(t+1)\}$. Рассчитывается ошибка временной разности:

$$\delta(t) = r(t) + \gamma\, V(t+1) - V(t)\,. \qquad (4)$$

Величина $\delta(t)$ характеризует ошибку в оценке $V(t) = V(\mathbf{S}(t))$ – суммарной награды, которую можно получить, исходя из состояния $\mathbf{S}(t)$. Ошибка $\delta(t)$ рассчитывается с учетом текущей награды $r(t)$ и оценки суммарной награды $V(\mathbf{S}(t+\tau))$, которую можно получить, исходя из следующего состояния $\mathbf{S}(t+\tau)$.

6) Веса НС Критика подстраиваются так, чтобы минимизировать величину $\delta(t)$, это обучение осуществляется градиентным методом, аналогично методу обратного распространения ошибки. Скорость обучения Критика равна $\alpha_C > 0$.

Смысл обучения Модели – уточнение прогнозов будущих ситуаций.

Смысл обучения Критика состоит в том, чтобы итеративно уточнять оценку качества ситуаций $V(\mathbf{S}(t))$ в соответствии с поступающими подкреплениями.

**3.3. Схема эволюции.** Эволюционирующая популяция состоит из $n$ агентов. Каждый агент имеет ресурс $R(t)$, который изменяется в соответствии с подкреплениями агента: $R(t+1) = R(t) + r(t)$, где $r(t)$ определено в (3).

Эволюция происходит в течение ряда поколений, $n_g$=1,2,…, $N_g$. Продолжительность каждого поколения $n_g$ равна $T$ тактов времени ($T$ – длительность жизни агента). В начале каждого поколения начальный ресурс каждого агента равен нулю, т.е., $R(T(n_g\text{-}1)+1) = 0$.

Начальные веса синапсов обоих НС (Модели и Критика) формируют геном агента $\mathbf{G}=\{\mathbf{W}_{M0},\mathbf{W}_{C0}\}$. Геном $\mathbf{G}$ задается в момент рождения агента и не меняется в течение его жизни. В противоположность этому текущие веса синапсов НС $\mathbf{W}_M$ и $\mathbf{W}_C$ подстраиваются в течение жизни агента путем обучения, описанного в п. 3.2.

В конце каждого поколения определяется агент, имеющий максимальный ресурс $R_{max}$ ($n_g$) (лучший агент поколения $n_g$). Этот лучший агент порождает $n$ потомков, которые составляют новое ($n_g$+1)-ое поколение. Геномы потомков $\mathbf{G}$ отличаются от генома родителя небольшими мутациями.

Более конкретно, в начале каждого нового ($n_g$+1)-го поколения мы полагаем для каждого агента $G_i(n_g+1) = G_{best,\,i}\,(n_g) + \mathrm{rand}_i$, $\mathbf{W}_0(n_g+1) = \mathbf{G}(n_g+1)$, где $\mathbf{G}_{\mathbf{best}}(n_g)$ – геном лучшего агента предыдущего $n_g$-го поколения и $\mathrm{rand}_i$ – это $N(0, P^2_{mut})$, т.е., нормально распределенная случайная величина с нулевым средним и стандартным отклонением $P_{mut}$ (интенсивность мутаций), которая добавляется к каждому весу.

Таким образом, геном $\mathbf{G}$ (начальные веса синапсов, получаемые при рождении) изменяется только посредством эволюции, в то время как текущие веса синапсов $\mathbf{W}$ дополнительно к этому подстраиваются посредством обучения, изложенным в п. 3.2. При этом в момент рождения агента $\mathbf{W} = \mathbf{W}_0 = \mathbf{G}$.

## 4. Результаты моделирования

**4.1. Общие особенности адаптивного поиска.** Изложенная модель была реализована в виде компьютерной программы. В наших вычислительных экспериментах мы использовали два варианта временного ряда:
1) синусоида:

$$X(t) = 0{,}5(1 + \sin(2\pi t/20)) + 1 , \qquad (5)$$

2) стохастический временной ряд, использованный в [8]:

$$X(t) = \exp(p(t)/1200), \quad p(t) = p(t\text{-}1) + \beta(t\text{-}1) + k\,\lambda(t), \qquad \beta(t) = \alpha\beta(t\text{-}1) + \mu(t), \qquad (6)$$

где $\lambda(t)$ и $\mu(t)$ – два нормальных процесса с нулевым средним и единичной дисперсией, $\alpha$ = 0,9, $k$ = 0,3.

Некоторые параметры модели имели одно и то же значение для всех экспериментов: фактор забывания $\gamma$ = 0,9; количество входов НС Модели $m$ = 10; количество нейронов в скрытых слоях НС Модели и Критика $N_{hM} = N_{hC}$ = 10; скорость обучения Модели и Критика $\alpha_M = \alpha_C$ = 0,01; параметр $\varepsilon$-жадного правила $\varepsilon$ = 0,05; интенсивность мутаций $P_{mut}$ = 0,1; расходы агента на покупку/продажу акций $J$ = 0. Остальные параметры (продолжительность поколения $T$ и численность популяции $n$) принимали разные значения в разных экспериментах, см. ниже.

Мы анализировали следующие варианты рассматриваемой модели:
- Случай L (чистое обучение); в этом случае рассматривался отдельный агент, который обучался методом временной разности, см. п. 1.2;
- Случай E (чистая эволюция), т.е. рассматривается эволюционирующая популяция без обучения;
- Случай LE (эволюция + обучение), т.е. полная модель, изложенная выше.

Было проведено сравнение ресурса, приобретаемого агентами за 200 временных тактов для этих трех способов адаптации. Для случаев E и LE бралось $T$ = 200 ($T$ – продолжительность поколения) и регистрировалось максимальное значение ресурса в популяции $R_{max}(n_g)$ в конце каждого поколения. В случае L (чистое обучение) рассматривался только один агент, ресурс которого для удобства сравнения со случаями E и LE обнулялся каждые $T$ = 200 тактов времени: $R(T(n_g\text{-}1)+1) = 0$. В этом случае индекс $n_g$ увеличивался на единицу после каждых $T$ временных тактов, и полагалось $R_{max}(n_g) = R(T\,n_g)$.

Графики $R_{max}(n_g)$ для синусоиды (5) показаны на рис. 3. Чтобы исключить уменьшение значения $R_{max}(n_g)$ из-за случайного выбора действий при применении $\varepsilon$-жадного правила для случаев LE и L, полагалось $\varepsilon$ = 0 после $n_g$ = 100 для случая LE и после $n_g$ = 2000 для случая L (на рис. 3 видно резкое увеличение $R_{max}(n_g)$ после $n_g$ = 100 и $n_g$ = 2000 для соответствующих случаев). Результаты усреднены по 1000 экспериментам; $n$ = 10, $T$ = 200.

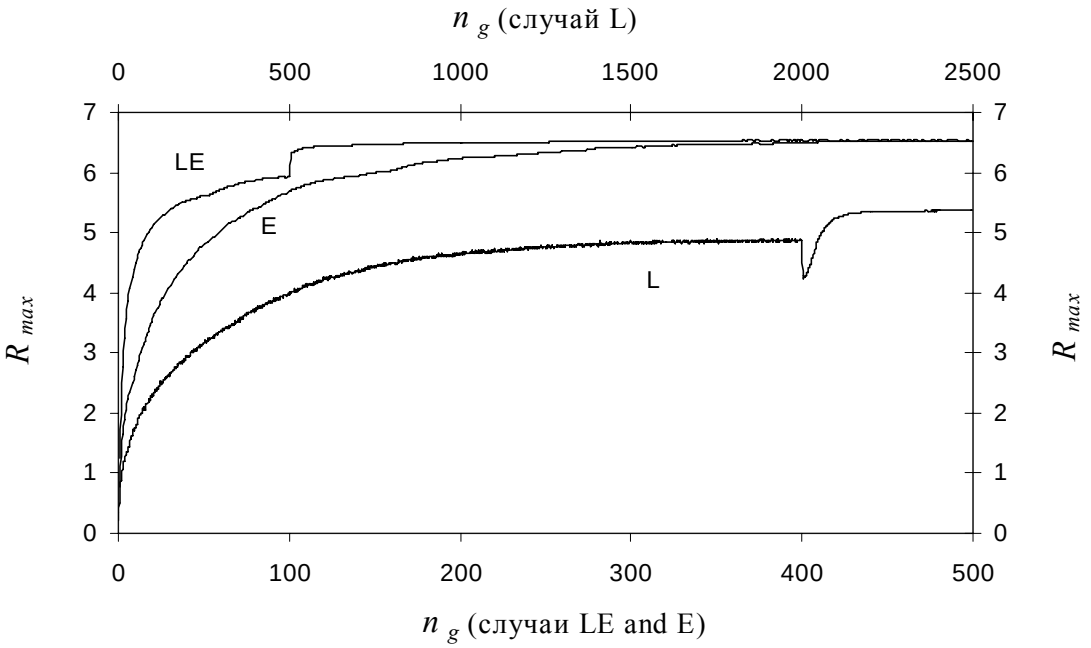


Рис. 3. Зависимости $R_{max}(n_g)$ . Кривая LE соответствует случаю эволюции, объединенной с обучением, кривая E – случаю чистой эволюции, кривая L – случаю чистого обучения. Временная шкала для случаев LE и E (номер поколения $n_g$) представлена снизу, для случая L (индекс $n_g$) – сверху. Моделирование проведено для синусоиды, кривые усреднены по 1000 экспериментам; $n = 10$, $T = 200$.

Рис. 3 показывает, что обучение, объединенное с эволюцией (случай LE), и чистая эволюция (случай E) дают одно и то же значение конечного ресурса $R_{max}(500) = 6{,}5$. Однако эволюция и обучение вместе обеспечивают нахождение больших значений $R_{max}$ быстрее, чем эволюция отдельно – существует симбиотическое взаимодействие между обучением и эволюцией.

Из (2) следует, что существует оптимальная стратегия поведения агента (в настоящей работе пренебрегаем затратами на покупку/продажу акций, т.е. всюду полагаем $J = 0$): вкладывать весь капитал в акции ($u(t+1) = 1$) при росте курса ($\Delta X(t+1) > 0$), вкладывать весь капитал в деньги ($u(t+1) = 0$) при падении курса ($\Delta X(t+1) < 0$).

Анализ экспериментов, представленных на рис. 3, показывает, что в случаях LE (обучение + эволюция), и E (чистая эволюция) такая оптимальная стратегия находится. Это соответствует асимптотическому значению ресурса $R_{max}(500) = 6{,}5$.

В случае L (чистое обучение) асимптотическое значение ресурса ($R_{max}(2500) = 5{,}4$) существенно меньше. Анализ экспериментов для этого случая показывает, что одно обучение обеспечивает нахождение только следующей «субоптимальной» стратегии поведения: агент держит капитал в акциях при росте и при слабом падении курса и переводит капитал в деньги при сильном падении курса. Та же тенденция к явному предпочтению вкладывать капитал в акции при чистом обучении наблюдается и для экспериментов на стохастическом ряде (6).

Итак, результаты, представленные на рис. 3, демонстрируют, что хотя обучение в настоящей модели и несовершенно, оно способствует более быстрому нахождению оптимальной стратегии поведения по сравнению со случаем чистой эволюции (см. графики LE и E на рис. 3).

Интересная особенность процесса поиска оптимального решения продемонстрирована на рис. 4. Этот рисунок показывает график $R_{max}(n_g)$ наряду со стандартным отклонением $\sigma(n_g)$ для случая LE. Значения $\sigma(n_g)$ характеризуют разброс значений $R_{max}(n_g)$ для различных реализаций моделируемых процессов. Рис. 4 показывает, что рост $R_{max}(n_g)$ сопровождается и ростом разброса значений $R_{max}$ : кривая $\sigma(n_g)$ имеет максимум в области быстрого роста $R_{max}(n_g)$. Эта особенность имеет общий характер: кривые $\sigma(n_g)$ имеют аналогичные максимумы и для случаев L и E.

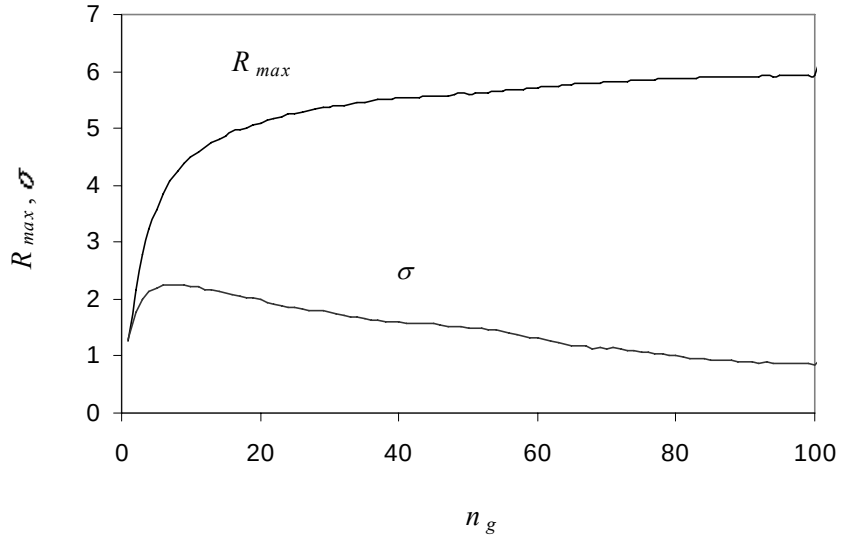


Рис. 4. Зависимости максимального по популяции ресурса $R_{max}(n_g)$, приобретаемого в течение поколения, и стандартного отклонения $\sigma(n_g)$ величины $R_{max}(n_g)$ от номера поколения $n_g$ . Случай LE. Значения $\sigma(n_g)$ характеризуют разброс значений $R_{max}(n_g)$ для различных реализаций моделируемых процессов. Моделирование проведено для синусоиды, кривые усреднены по 1000 экспериментам; $n = 10$, $T = 200$.

Стоит отметить, что эта особенность – увеличение числа случайных вариантов возможных решений на активной стадии поиска – сходна с явлением генерализации при выработке условного рефлекса [10]. При выработке условного рефлекса на стадии генерализации также происходит интенсификация случайной поисковой активности: реакция возникает не только на условный стимул, но на различные подобные ему (дифференцировочные) раздражители. И только затем происходит специализация, при которой реакция на дифференцировочные стимулы постепенно ослабевает и сохраняется только реакция на условный стимул. При выработке условного рефлекса зависимость условной реакции от числа экспериментов подобна кривой $R_{max}(n_g)$ на рис. 4. Увеличение интенсивности случайного поиска при генерализации сходно с ростом величины $\sigma(n_g)$ в области быстрого увеличения величины $R_{max}(n_g)$.

**4.2. Особенности обучения (чистое обучение без эволюции).** Рис. 3 демонстрирует, что рассмотренная простая форма обучения при данной структуре НС (см. п.3.2) несовершенна, так как она может привести лишь к «субоптимальной» стратегии поведения, даже если обучение происходит в течение большого числа поколений. Асимптотическое значение $R_{max}$ для синусоиды составляет только $R_{max}$= 5,4 (см. кривую L на рис.3), что значительно меньше асимптотического значения $R_{max}$= 6,5, соответствующего оптимальной стратегии (кривые LE и E на рис.3). Анализ выбираемых обучающимся агентом действий показывает, что чистое обучение способно найти лишь следующую «субоптимальную» стратегию: агент покупает акции, когда их цена растет или слегка падает и продает акции, когда их цена падает значительно. Такое поведение агента для синусоидального и стохастического временного ряда показано на рис. 5 и 6, соответственно.

Таким образом, в случае чистого обучения агент явно предпочитает хранить свой капитал в акциях, а не в деньгах.

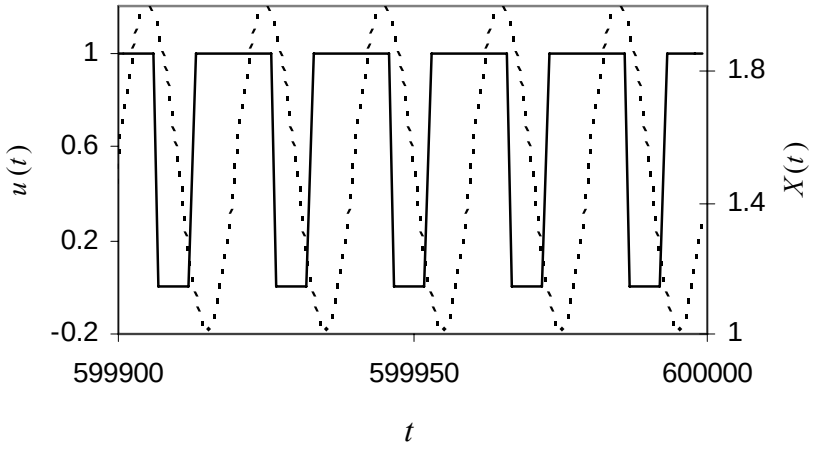


Рис. 5. Динамика поведения обучающегося агента для синусоиды (5). Действия агента характеризуются величиной $u(t)$ (сплошная линия): при $u = 0$ весь капитал переведен в деньги, при $u = 1$ весь капитал переведен в акции. Временной ряд $X(t)$ показан пунктирной линией. Случай чистого обучения.

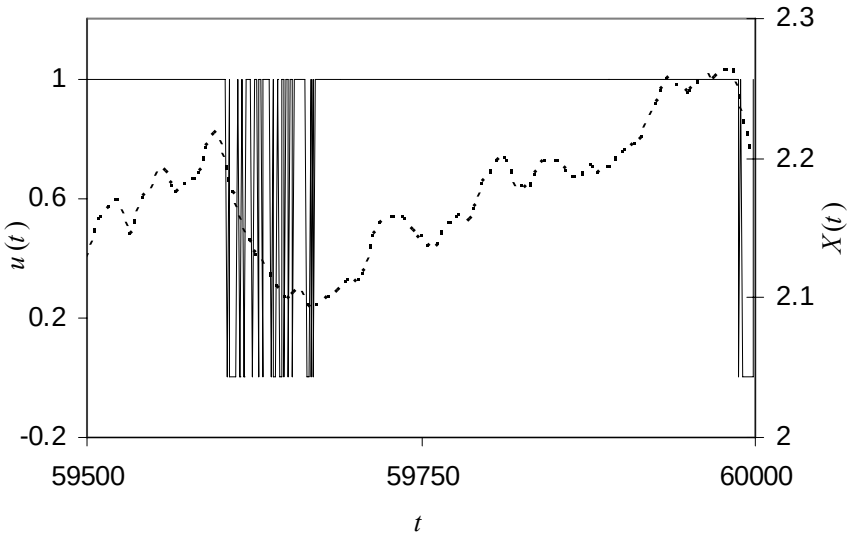


Рис. 6. Динамика поведения обучающегося агента для стохастического ряда (5). Действия агента характеризуются величиной $u(t)$ (сплошная линия): при $u = 0$ весь капитал переведен в деньги, при $u = 1$ весь капитал переведен в акции. Временной ряд $X(t)$ показан пунктирной линией. Случай чистого обучения.

**4.2. Взаимодействие между обучением и эволюцией. Эффект Балдвина.** Как показано на рис. 3 (кривая E) для синусоидального временного ряда чистая эволюция способна найти оптимальную стратегию во всех экспериментах. В случае стохастического временного ряда оптимальная стратегия также может быть найдена при помощи только эволюции, но лишь в некоторых экспериментах. Например, при $N_g = 300$ и $T = 200$ эволюция смогла найти оптимальную стратегию в восьми из 10 экспериментов. Типичный пример оптимальной стратегии поведения представлен на рис. 7.

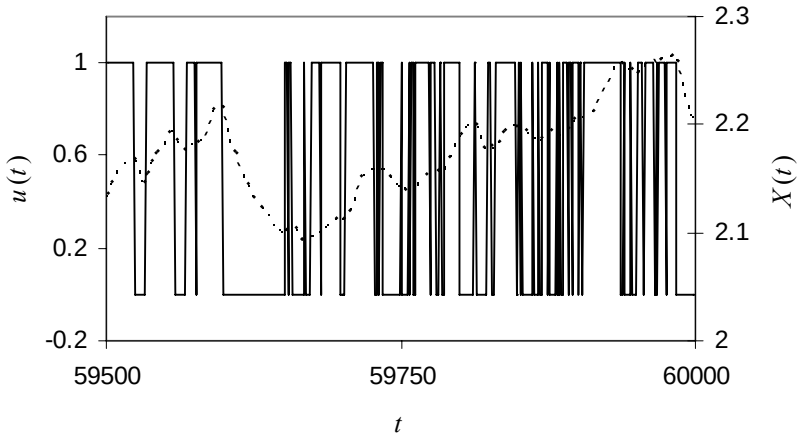


Рис. 7. Стратегия поведения лучшего агента в популяции. Действия агента характеризуются величиной $u(t)$ (сплошная линия): при $u = 0$ весь капитал переведен в деньги, при $u = 1$ весь капитал переведен в акции. Временной ряд $X(t)$ показан пунктирной линией. $n = 10$, $T = 200$. Случай чистой эволюции. Стратегия поведения агента практически оптимальна: агент покупает/продает акции при росте/падении курса акций.

Рис. 3 также демонстрирует, что поиск оптимальной стратегии посредством только эволюции происходит медленнее, чем при эволюции, объединенной с обучением (см. кривые E и LE на этом рисунке). Хотя обучение в нашей модели само по себе не оптимально, оно помогает эволюции находить лучшие стратегии.

Если длительность поколения $T$ была достаточно большой (1000 и более тактов времени), то для случая LE часто наблюдалось и более явное влияние обучения на эволюционный процесс. В первых поколениях эволюционного процесса существенный рост ресурса агентов наблюдался не с самого начала поколения, а спустя 200-300 тактов, т.е. агенты явно обучались в течение своей жизни находить более или менее приемлемую стратегию поведения, и только после смены ряда поколений рост ресурса начинался с самого начала поколения. Это можно интерпретировать как проявление известного эффекта Балдвина: исходно приобретаемый навык в течение ряда поколений становился наследуемым [3,11,12]. Этот эффект наблюдался в ряде экспериментов, один из которых представлен на рис. 8.

Для этого эксперимента было проанализировано, как изменяется значение ресурса наилучшего агента в популяции $R_{max}(t)$ в течение первых пяти поколений. Расчет был проведен для синусоидального ряда (5). Рис. 8 показывает, что в течение первых двух поколений значительный рост ресурса лучшего в популяции агента начинается только после задержки 100-300 тактов времени; т.е., очевидно, что агент оптимизирует свою стратегию поведения при помощи обучения. От поколения к поколению агент находит хорошую стратегию поведения все раньше и раньше. К пятому поколению лучший агент «знает» хорошую стратегию поведения с самого рождения, и обучение не приводит к существенному улучшению стратегии. Таким образом, рис. 8 показывает, что стратегия, изначально приобретаемая посредством обучения, становится наследуемой (эффект Балдвина).

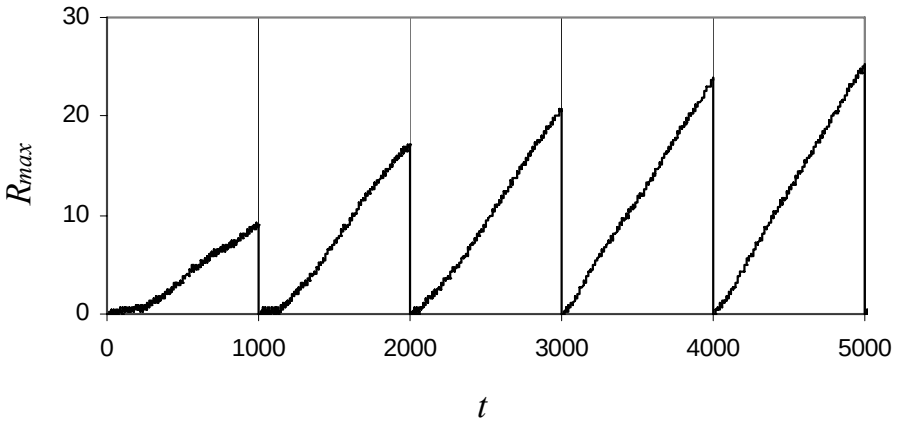


Рис. 8. Зависимость ресурса лучшего в популяции агента $R_{max}$ от времени $t$ для первых пяти поколений. Случай LE (эволюция, объединенная с обучением); размер популяции $n$ = 10, длительность поколения $T$ = 1000. Моменты смены поколений показаны вертикальными линиями. Для первых двух поколений есть явная задержка в 100-300 тактов времени в росте ресурса агента. К пятому поколению лучший агент «знает» хорошую стратегию поведения с самого рождения, т.е. стратегия, изначально приобретаемая посредством обучения, становится наследуемой.

Мы проанализировали различные наборы параметров модели и выяснили, что эффект Балдвина стабильно проявляется, если продолжительность поколения $T$ составляет 1000 и более тактов времени, что обеспечивает достаточно эффективное обучение в течение жизни агента.

**4.3. Особенности предсказания Модели. Практика не есть критерий истины.** Система управления каждого агента включает в себя нейронную сеть Модели, предназначенную для предсказания изменения значения $\Delta X$ ($t$+1) временного ряда в следующий такт времени $t$+1. Мы проанализировали работу Модели и обнаружили очень интересную особенность. Нейронная сеть Модели может давать неверные предсказания, однако агент, тем не менее, может использовать эти предсказания для принятия верных решений. Например, рис. 9 показывает предсказываемые изменения $\Delta X^{pr}(t+1)$ и реальные изменения $\Delta X(t+1)$ стохастического временного ряда в случае чистой эволюции (случай E). Предсказания нейронной сети Модели достаточно хорошо совпадают по форме с кривой $\Delta X$. Однако, предсказанные значения $\Delta X^{pr}(t+1)$ отличаются примерно в 25 раз от значений $\Delta X(t+1)$.

На рис. 10 приведен другой пример особенностей предсказания нейронной сети Модели в случае LE (эволюция, объединенная с обучением). Этот пример показывает, что предсказания нейронной сети Модели могут отличаться от реальных данных не только масштабом, но и знаком.

Хотя предсказания Модели могут быть неверными количественно, мы полагаем, что правильность их формы или правильность после линейных преобразований (например, изменения знака) приводит к тому, что Модель является полезной для адаптивного поведения. Эти предсказания эффективно используются системой управления агентов для нахождения оптимальной поведения: стратегия поведения агентов для обоих приведенных примеров работы Модели была подобна стратегии, представленной на рис. 7.

По-видимому, наблюдаемое увеличение значений $\Delta X^{pr}$ нейронной сетью Модели полезно для работы нейронной сети Критика, так как реальные значения $\Delta X(t+1)$ слишком малы (порядка 0,001). Таким образом, нейронная сеть Модели может не только предсказывать значения $\Delta X^{pr}(t+1)$, но также осуществлять полезные преобразования этих значений.

Эти особенности работы нейронной сети Модели обусловлены доминирующей ролью эволюции над обучением при оптимизации системы управления агентов. На самом деле, из-за малой длительности поколений ($T$ = 200) в нашем моделировании, веса синапсов нейронных сетей изменяются большей частью за счет эволюционных мутаций. Такой процесс делает предпочтительными такие системы управления, которые устойчивы в эволюционном смысле. Кроме того, важно подчеркнуть, что задача, которую «решает» эволюция в настоящей модели, значительно проще, чем та задача, которую решает обучение. Эволюции достаточно обеспечить выбор действий (покупать или продавать), приводящий к награде. А схема обучения предусматривает довольно сложную процедуру прогноза ситуации **S,** оценки качества прогнозируемых ситуаций, итеративного формирования оценок качества ситуаций $V$(**S**) и выбора действия на основе этих оценок. То есть эволюция идет к нужному результату более прямым путем, а так как задача агентов проста, то эволюция в определенной степени «задавливает» довольно сложный механизм обучения. Тем не менее, есть определенная синергия во взаимодействии обучения и эволюции: обучение ускоряет процесс поиска оптимальной стратегии поведения.

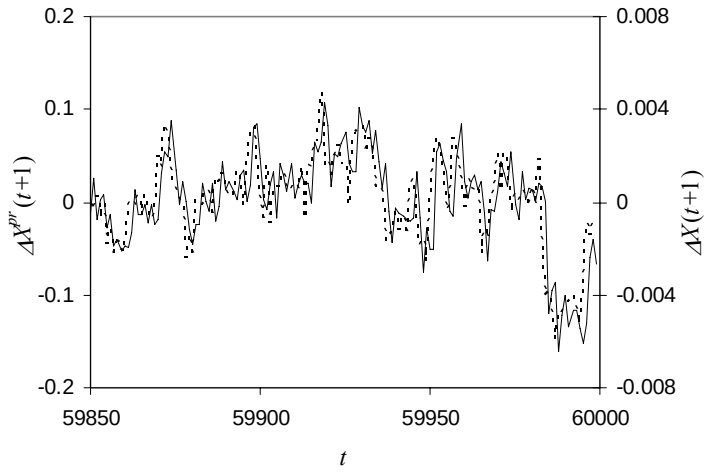


Рис. 9. Предсказываемые $\Delta X^{pr}(t+1)$ (пунктирная линия) и реальные изменения $\Delta X(t+1)$ (сплошная линия) стохастического временного ряда. Случай чистой эволюции. $n$ = 10, $T$ = 200. Хотя обе кривые имеют сходную форму, по величине $\Delta X^{pr}$ ($t$+1) и $\Delta X$ ($t$+1) радикально различаются.

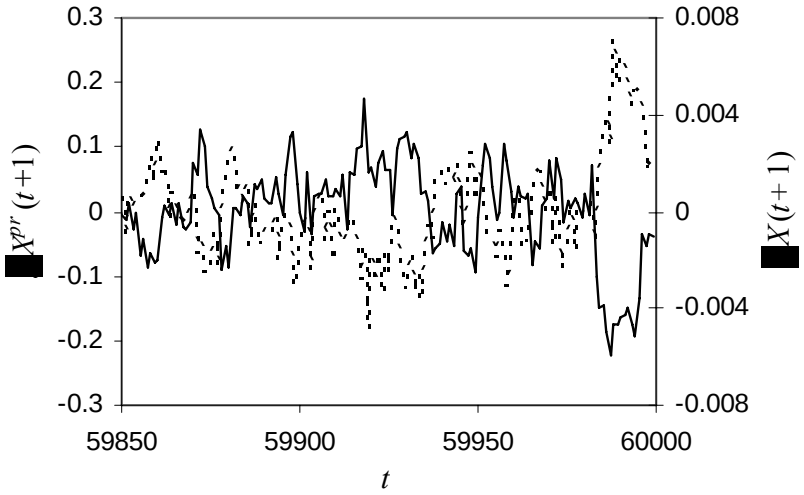


Рис. 10. Предсказываемые $\Delta X^{pr}(t+1)$ (пунктирная линия) и реальные изменения $\Delta X(t+1)$ (сплошная линия) стохастического временного ряда. Случай эволюции, объединенной с обучением. $n$ = 10, $T$ = 200. Кривые $\Delta X^{pr}$ ($t$+1) и $\Delta X$ ($t$+1) различаются как по величине, так и знаком.

**2.4. Сравнение с поведением простейших животных.** Исследуемые агенты имеют две поведенческие тактики (продавать или покупать акции) и выбирают действия, переключаясь между этими тактиками. Можно сопоставить особенности этого поведения с переключением между двумя тактиками при поисковом поведении простейших животных. Например, некоторые виды личинок ручейников используют аналогичные тактики [13,14]. Личинки живут на речном дне и носят на себе «домик» – трубку из песка и других частиц, которые они собирают на дне водоемов. Личинки строят свои домики из твердых частичек разной величины. Они могут использовать маленькие или большие песчинки [13]. Большие песчинки распределены случайно, но обычно встречаются группами. Используя большие песчинки, личинка может построить домик гораздо быстрее и эффективнее, чем используя маленькие, и, естественно, предпочитает использовать большие частицы. Личинка использует две тактики: 1) тестирование частиц вокруг себя и использование выбранных частиц, 2) поиск нового места для сбора частиц. Исследование поведения личинок обнаруживает инерцию в переключении с первой тактики на вторую [13,14]. Если личинка находит большую частицу, она продолжает тестировать частицы, пока не найдет несколько маленьких, и только после нескольких неудачных попыток найти новую большую частицу, переходит ко второй тактике. Во время поиска нового места личинка время от времени тестирует частицы, которые попадаются на ее пути. Она может переключиться со второй тактики на первую, если найдет большую частицу; при этом переключении также может проявляться инерция. Таким образом, переключение между тактиками имеет характер случайного поиска с явным эффектом инерции. Процесс инерционного переключения позволяет животному использовать только общие крупномасштабные свойства окружающего мира, и игнорировать мелкие случайные детали.

В наших компьютерных экспериментах поведение агента-брокера, подобное поведению животных с инерционным переключением между двумя тактиками, наблюдалось, когда система управления агента оптимизировалась с помощью чистой эволюции при достаточно большой численности популяции. То есть фактически происходила оптимизация методом случайного поиска в достаточно большой области возможных решений. Рис. 11 показывает фрагмент стратегии поведения агента, найденной на ранней стадии эволюции в большой популяции, $n = 100$. Эта стратегия агента подобна описанному выше поведению животных с инерционным переключением между двумя тактиками. Стратегия переключения между $u = 0$ и $u = 1$ представляет собой реакцию только на общие изменения в окружающей среде (агент игнорирует мелкие флуктуации в изменении курса акций). Кроме того, переключение явно обладает свойством инерционности.

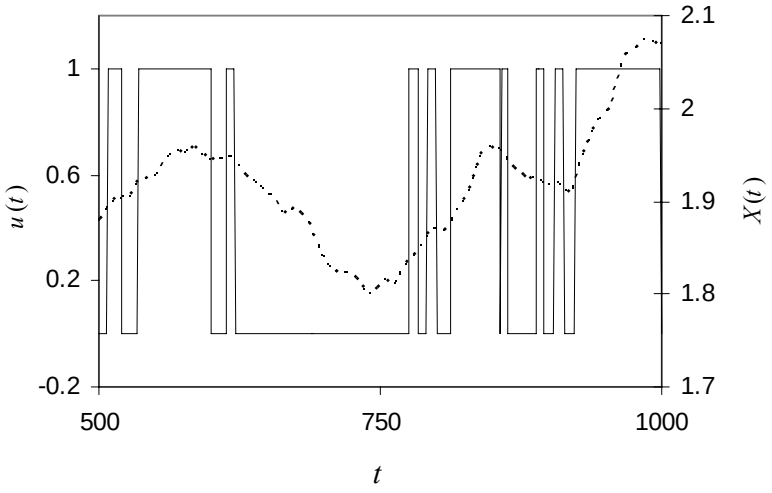


Рис. 11. Стратегия поведения лучшего агента в популяции. Действия агента характеризуются величиной $u(t)$ (сплошная линия): при $u = 0$ весь капитал переведен в деньги, при $u = 1$ весь капитал переведен в акции. Временной ряд $X(t)$ показан пунктирной линией. $n = 100$, $T = 200$. Стратегия агента подобна поведению животных с инерционным переключением между двумя тактиками. Агент игнорирует мелкие флуктуации динамики курса акций, переключение между при выборе действия $u = 0$ и выбором действия $u = 1$ обладает свойством инерционности.

## 5. Заключение

Итак, построена модель эволюции автономных агентов, система управления которых основана на нейросетевых адаптивных критиках. Проведено исследование взаимодействия между обучением и эволюцией в популяциях самообучающихся агентов-брокеров. Проанализировано три варианта модели, в которых веса синапсов нейронных сетей настраивались 1) обучением, 2) эволюцией или 3) комбинацией обучения и эволюции.

Показано, что оптимальная стратегия обеспечивается в случаях чистой эволюции или комбинации обучения и эволюции. Чистое обучение не обеспечивает нахождения оптимальной стратегии. Тем не менее, хотя обучение в настоящей модели и несовершенно, оно способствует более быстрому нахождению оптимальной стратегии поведения для случая комбинации обучения и эволюции по сравнению со случаем чистой эволюции.

При достаточно большой длительности жизни агентов часто наблюдалось и более явное влияние обучения на эволюционный процесс. В первых поколениях эволюционного процесса агенты явно обучались находить удачную стратегию поведения в течение своей жизни, а после смены ряда поколений такая стратегия была у агентов с самого рождения. То есть исходно приобретаемый навык в течение ряда поколений становился наследуемым, что можно интерпретировать как проявление известного эффекта Болдуина.

Система управления агентов содержала две нейронные сети: Модель и Критик. Модель предназначена для прогнозирования изменения курса временного ряда. Критик предназначен для оценки качества ситуаций. Моделирование продемонстрировало, что в случае комбинации обучения и эволюции предсказание Модели может быть количественно неверным (Модель предсказывает правильно лишь форму изменений временного ряда, причем, возможно, с неправильным знаком), тем не менее, на основе этого неверного прогноза формируется оптимальная стратегия поведения агента. По-видимому, эта особенность работы системы управления агента обусловлена тем, что задача, которую

«решает» эволюция в настоящей модели значительно проще, чем та задача, которую решает обучение. Эволюции достаточно обеспечить выбор действий (покупать или продавать), приводящий к награде. А схема обучения предусматривает значительно более сложную процедуру прогноза ситуации, оценки качества прогнозируемых ситуаций, итеративного формирования оценок качества ситуаций и выбора действия на основе этих оценок. То есть эволюция идет к нужному результату более прямым путем, а так как задача агентов проста, то эволюция в определенной степени «задавливает» сложный механизм обучения.

Проведено сравнение поведение агента-брокера с поисковым поведением простых животных. Исследования поискового поведения животных показывают, что часто в процессе поиска проявляются инерционные эффекты при переключении между разными поведенческими тактиками. Такая инерция помогает животному адаптивно реагировать только на общие изменения в окружающей среде. В нашей модели подобное инерционное поведение формируется в процессе эволюционного поиска на ранних стадиях эволюции при условии достаточного большого размера популяции.

Опыт работы с настоящей моделью показывает важность вопроса о том, какие системы управления автономных агентов являются эволюционно устойчивыми. Под эволюционной устойчивостью мы подразумеваем свойство фенотипа (и соответствующего ему генотипа) становиться практически нечувствительным к мутациям. В частности, наше моделирование демонстрирует, что сложные нейросетевые схемы обучения могут быть эволюционно нестабильны, если процесс обучения неустойчив относительно к возмущениям весов синапсов нейронных сетей.

Авторы благодарны В.А. Непомнящих за ряд консультаций по принципам поискового поведения простых животных.

**Литература:**

1. Тарасов В.Б. От многоагентных систем к интеллектуальным организациям: философия, психология, информатика. М.: Эдиториал УРСС, 2002. 352 с.
2. Sutton R., Barto A. Reinforcement Learning: An Introduction. – Cambridge: MIT Press, 1998. See also: http://www.cs.ualberta.ca/~sutton/book/the-book.html
3. Baldwin J.M. A new factor in evolution // American Naturalist, 1896, vol. 30, pp. 441-451.
4. Learning and Approximate Dynamic Programming: Scaling Up to the Real World (Edited by J. Si, A. Barto, W. Powell, D. Wunsch), IEEE Press and John Wiley & Sons, 2004.
5. Цетлин М.Л. Исследования по теории автоматов и моделирование биологических систем. – М.: Наука, 1969. 316 с.
6. Редько В.Г., Прохоров Д.В. Нейросетевые адаптивные критики // Научная сессия МИФИ-2004. VI Всероссийская научно-техническая конференция "Нейроинформатика-2004". Сборник научных трудов. Часть 2. М.: МИФИ, 2004. С.77-84.
7. Prokhorov D.V. , Wunsch D.C. Adaptive critic designs // IEEE Transactions on Neural Networks, 1997, vol.8, pp.997-1007.
8. Prokhorov D., Puskorius G., Feldkamp L. Dynamical neural networks for control // In J. Kolen and S. Kremer (Eds.) A field guide to dynamical recurrent networks. NY: IEEE Press, 2001, pp. 257-289.
9. Moody J., Wu L., Liao Y., Saffel M. Performance function and reinforcement learning for trading systems and portfolios // Journal of Forecasting, 1998, vol.17, pp. 441-470.
10. Котляр Б.И., Шульговский В.В. Физиология центральной нервной системы. М.: Изд-во МГУ. 1979. 342 с.
11. Turney P., Whitley D., Anderson R. (Eds.). Evolution, Learning, and Instinct: 100 Years of the Baldwin Effect // Special Issue of Evolutionary Computation on the Baldwin Effect, V.4, N.3, 1996.

12. Weber B.H., Depew D.J. (Eds.) Evolution and learning: The Baldwin effect reconsidered. MA: MIT Press, 2003.
13. Nepomnyashchikh V.A. Selection behaviour in caddis fly larvae // In R. Pfeifer et al (Eds.) From Animals to Animats 5: Proceedings of the Fifth International Conference of the Society for Adaptive Behavior. Cambridge, MA: MIT Press, 1998, pp.155-160.
14. Непомнящих В.А. Как животные решают плохо формализуемые задачи поиска // Синергетика и психология. Тексты. Выпуск 3. Когнитивные процессы (под ред. Аршинова В.И., Трофимовой И.Н., Шендяпина В.М.) М.: Издательство Когнито-центр, 2004. С. 197-209.