An Approach to Optimal Action Generation for a System that Interacts with the Environment*

Oleg Volkov

Revision 0.84 of 26 May 2013

Abstract

This paper describes an algorithm of operation of a system that performs step-by-step interaction with the environment by accepting and emitting signals. The goal of such operation is to get the maximum amount of spur (energy or some resource) from the environment. At key point of the algorithm a random number generator is used. The algorithm rests on the concept of a cycle in the graph of transitions between system states and on the concept of average spur increment per such cycle step. Case of system structure, when the system is comprised of two subsystems, the first of which identifies current environment state, and the second generates an optimal action based on that state, is also discussed. The idea of the algorithm can be used in control and human interaction systems that operate in real-time.

1 Algorithm Description

Each step $t \in \overline{1, n_{ev}}$ of interaction of the system with the environment is described by triplet $\left\langle s_{in}^{(t)}, \Delta E^{(t)}, s_{out}^{(t)} \right\rangle$ and implies accepting *informational signal* $s_{in}^{(t)} \in \overline{1, n_{in}}$ which is an approximation to environment state at step t, accepting spur (energy or some resource) increment $\Delta E^{(t)}$, and emitting *action signal* $s_{out}^{(t)} \in \overline{1, n_{out}}$ to the environment. The goal of system operation is to maximize $E = \sum_{t=1}^{n_{ev}} \Delta E^{(t)}$.

A system state is represented by an n-gram[1] of length N or of shorter length if the value of t is too small, and the number of events occurred is insufficient to form the n-gram of length N. Let $Sx_i \in \overline{1, n_{in}}$, $Ay_i \in \overline{1, n_{out}}$, then n-grams of form $\langle ...Sx_{N-3}, Ay_{N-2}, Sx_{N-1}, Ay_N \rangle$ will correspond to action execution states, and n-grams of form $\langle ...Ay_{N-3}, Sx_{N-2}, Ay_{N-1}, Sx_N \rangle$ will correspond to action choice states. A sequence of state transitions

$$\left\langle ...Sx_{N-1}, Ay_N \right\rangle \rightarrow \left\langle ...Sx_{N-1}, Ay_N, s_{in}^{(t)} \right\rangle \rightarrow \left\langle ...Sx_{N-1}, Ay_N, s_{in}^{(t)}, s_{out}^{(t)} \right\rangle$$

corresponds to single step t, and state $\left\langle ...Sx_{N-1}, Ay_N, s_{in}^{(t)} \right\rangle$ corresponds to substep $t + \frac{1}{2}$.

A cycle in the graph of transitions between n-grams, which represent action choice states, is the key concept used in the algorithm. A type of cycles is defined by pair $\langle h, z \rangle$, where action choice n-gram h is initial and final state of cycles, $z \in \overline{1, n_{out}}$ is a direction of cycles started from initial state h.

In memory space, which corresponds to each action choice n-gram h, condition vector

$$\left\langle \alpha^{(h)}, t_0^{(h)}, E_0^{(h)}, \psi^{(h)} \right\rangle$$

for an action choice state is stored. In this vector, $\alpha^{(h)} \in \overline{0, n_{out}}$ is an action signal emitted just before the last leaving node h of transition graph (or 0 if there was no such signal yet), which is used to determine the direction of a cycle at node h at the next return to that node, $t_0^{(h)}$ is the step of last leaving node h, $E_0^{(h)}$ is spur at step $t_0^{(h)}$, $\psi^{(h)}$ is a table of correspondence of possible action signals $z \in \overline{1, n_{out}}$ to cycle statistics vectors

$$\left\langle \nu^{(h,z)}, \omega^{(h,z)}, H^{(h,z)} \right\rangle$$

with zero vectors for initial values. In cycle statistics vectors, $\nu^{(h,z)}$ is the number of cycles of type $\langle h, z \rangle$ occurred since the beginning of system operation, $\omega^{(h,z)}$ is the total length of cycles of type $\langle h, z \rangle$ occurred, $H^{(h,z)}$ is the total spur increment over the cycles of type $\langle h, z \rangle$.

[1] An n-gram is a subsequence of length N of a sequence. In our case the sequence is the history of accepted and emitted signals $s_{in}^{(1)}, s_{out}^{(1)}, s_{in}^{(2)}, s_{out}^{(2)}, s_{in}^{(3)}, s_{out}^{(3)}, ...$ When $N = 2$, n-grams will be pairs of two neighbour elements.

When informational signal $s_{in}^{(t)}$ is accepted, transition to an action choice state that corresponds to n-gram h of length $\min\{2t-1, N\}$ is made. If $2t-1 < N$, then processing the informational signal will be finished. Otherwise, if $\alpha = \alpha^{(h)} \neq 0$, then update of components of cycle statistics vector will be performed for cycles of type $\langle h, \alpha \rangle$, which includes increment of the number of such cycles $\nu^{(h,\alpha)}$, increment of the total length $\omega^{(h,\alpha)}$ of cycles by the difference between current step t and step $t_0^{(h)}$ of cycle start, and increment of total spur augment $H^{(h,\alpha)}$ over the cycles by the difference between current spur $E^{(t)} = \sum_{i=1}^{t} \Delta E^{(i)}$ and spur $E_0^{(h)}$ at step $t_0^{(h)}$:

$$\begin{aligned} \nu^{(h,\alpha)} &:= \nu^{(h,\alpha)} + 1; \\ \omega^{(h,\alpha)} &:= \omega^{(h,\alpha)} + t - t_0^{(h)}; \\ H^{(h,\alpha)} &:= H^{(h,\alpha)} + E^{(t)} - E_0^{(h)}. \end{aligned}$$

If $2t-1 < N$, then for action signal $s_{out}^{(t)}$ a random signal will be chosen. In normal case, when $2t-1 \geq N$, action signal $z \in \overline{1, n_{out}}$ in state h is being chosen with probability proportional to

$$F(h, z, t) = \exp\left(\frac{H^{(h,z)}}{\omega^{(h,z)}} K(h, z, t)\right),$$

where $H^{(h,z)} / \omega^{(h,z)}$ is an average spur increment per step of cycles of type $\langle h, z \rangle$, and $K(h, z, t)$ is a multiplier, with greater value of which the probability of choice of action signal z that corresponds to greater value of $H^{(h,z)} / \omega^{(h,z)}$ increases. The use of an exponential form for the relative probability function, to which the Boltzmann distribution of probability of transition to one of possible system states corresponds, had shown to be efficient in solving global minimization tasks (e.g. see an algorithm described in [2]).

After emitting an action signal $s_{out}^{(t)}$ by the system, when $2t-1 \geq N$, the following assignments are made:

$$\begin{aligned} \alpha^{(h)} &:= s_{out}^{(t)}; \\ t_0^{(h)} &:= t; \\ E_0^{(h)} &:= E^{(t)}. \end{aligned}$$

Upon completion of action signal processing, transition to an action execution state is made and step index t is incremented.

Multiplier $K(h, z, t)$ can be defined in a variety of ways, e.g. following the approach used in simulated annealing algorithm, when the multiplier increases as t increases. Own way of definition of the multiplier is described in this paper.

The quantity being maximized is spur received by the system during its operating time or, in other words, an average spur increment per system operation step, equal to $E^{(t)} / t$, that verges towards an average spur increment per cycle step as t increases. Quantity $E^{(t)} / t$ will maximize if the frequency of repeating cycles of type $\langle h, z \rangle$ is as higher as the ratio of average spur increment per step of those cycles to quantity $\left|E^{(t)}\right| / t$ is greater. Quantity $t / \left|E^{(t)}\right|$ is a coefficient for translation of value $H^{(h,z)} / \omega^{(h,z)}$ to value w_z that does not depend on the scale of average spur increment per cycle step:

$$w_z = \frac{t H^{(h,z)}}{\left|E^{(t)}\right| \omega^{(h,z)}}.$$

In case of $w_z = 1$ when

$$\frac{H^{(h,z)}}{\omega^{(h,z)}} = \frac{\left|E^{(t)}\right|}{t}$$

for some z, and $w_i = 0$ when $H^{(h,i)} = 0$ for $i \in \overline{1, n_{out}} \setminus \{z\}$, let us perform choice of action z with probability p_c and choice of each action i with probability $(1 - p_c) / (n_{out} - 1)$. If action z in state h were chosen with probability p_c the previous time, then with probability $p_f = p_c^2$ an action chosen this time would be action z and would coincide with an action chosen in state h the previous time.

If the system had made a cycle of type $\langle h, z \rangle$ of length κ substeps and got into state h, then exact repeat of a sequence of signals, which correspond to substeps of that cycle, would bring the system into state h again. Let there is some signal sequence, then if we generate the second signal sequence of the same length, choosing with probability p_f at every position of the second sequence a signal, which is at that position in the first sequence, then both sequences will have common prefix exactly of length κ with probability $p_r(\kappa) = p_f^{\kappa}(1 - p_f)$, $\sum_{i=0}^{\infty} p_r(i) = 1$. If p_f is fixed, then the average length of common prefix will be equal to

$$\bar{\kappa} = \sum_{i=0}^{\infty} i p_r(i) = \frac{p_f}{1 - p_f},$$

whence it follows that $p_f = \frac{\bar{\kappa}}{\bar{\kappa}+1}$,

$$p_c = \sqrt{\frac{\bar{\kappa}}{\bar{\kappa}+1}}.$$

To probability p_c there corresponds a probability of falling a uniformly chosen random point into a segment of length x followed by $n_{out}-1$ segments of length $\exp 0 = 1$. By solving the equation

$$\frac{x}{x+n_{out}-1} = \sqrt{\frac{\bar{\kappa}}{\bar{\kappa}+1}},$$

we get

$$x = \sqrt{\bar{\kappa}}\left(\sqrt{\bar{\kappa}} + \sqrt{\bar{\kappa}+1}\right)(n_{out}-1).$$

It is possible to get equality $F(h,z,t) = x$ when $w_z = 1$ and provide a condition that when w_z is changing, the value of function $F(h,z,t)$ changes in accordance with an exponential form chosen for that function if we define the function as $F(h,z,t) = B(h,z)^{w_z}$, where $B(h,z) = x$:

$$\begin{aligned} F(h,z,t) &= B(h,z)^{\frac{tH^{(h,z)}}{\left|E^{(t)}\right|\omega^{(h,z)}}} = \left(\sqrt{\bar{\kappa}}\left(\sqrt{\bar{\kappa}} + \sqrt{\bar{\kappa}+1}\right)(n_{out}-1)\right)^{\frac{tH^{(h,z)}}{\left|E^{(t)}\right|\omega^{(h,z)}}}; \\ K(h,z,t) &= \frac{t}{\left|E^{(t)}\right|}\ln B(h,z) = \frac{t}{\left|E^{(t)}\right|}\ln\left(\sqrt{\bar{\kappa}}\left(\sqrt{\bar{\kappa}} + \sqrt{\bar{\kappa}+1}\right)(n_{out}-1)\right). \end{aligned}$$

Here

$$\bar{\kappa} = \frac{2\omega^{(h,z)}}{\nu^{(h,z)}}$$

is the average number of substeps in cycles of type $\langle h,z\rangle$. If $\nu^{(h,z)} = 0$ or $\omega^{(h,z)} = 0$ or $E^{(t)} = 0$, let us assume that $F(h,z,t) = 1$ and $K(h,z,t) = 0$.

2 Example of System Operation

Let us see into system operation according to the algorithm by an example. Let $n_{in} = 5$, $n_{out} = 5$, $N = 2$, and the sequence of system operation steps is described in Table 1.

t	$s_{in}^{(t)}$	$E^{(t)}$	Action choice state	$s_{out}^{(t)}$	Action execution state
1	1	0	$\langle S1\rangle$	2	$\langle S1, A2\rangle$
2	2	0	$\langle A2, S2\rangle$	3	$\langle S2, A3\rangle$
3	3	0	$\langle A3, S3\rangle$	5	$\langle S3, A5\rangle$
4	4	1	$\langle A5, S4\rangle$	3	$\langle S4, A3\rangle$
5	3	1	$\langle A3, S3\rangle$	4	$\langle S3, A4\rangle$
6	5	2	$\langle A4, S5\rangle$	1	$\langle S5, A1\rangle$
7	1	2	$\langle A1, S1\rangle$	2	$\langle S1, A2\rangle$
8	2	2	$\langle A2, S2\rangle$	3	$\langle S2, A3\rangle$

Table 1: example of system operation steps sequence

In Figure 1, action choice states (encircled by thick lines) and action execution states (encircled by thin lines), which occur in that sequence of steps, are shown, and directions of transitions between states are shown (thick arrows indicate transitions from action choice states). Values near arrows indicate increments of system spur. The initial state is $\langle S1\rangle$, the final state is $\langle S2, A3\rangle$.

Let us assume that after completion of example sequence of steps, at step $t = 9$ the system had accepted informational signal $s_{in}^{(9)} = 3$, to which spur $E^{(t)} = E^{(9)} = 2$ corresponds, and the system changed its state from the final one $\langle S2, A3\rangle$ to action choice state $h = \langle A3, S3\rangle$. Current value of $\alpha = \alpha^{(h)} = \alpha^{(\langle A3,S3\rangle)}$ is 4, because the last transition from state $\langle A3, S3\rangle$ was a transition to state $\langle S3, A4\rangle$ by action signal $s_{out}^{(5)} = 4$. Current values of variables $t_0^{(h)}$ and $E_0^{(h)}$, which were stored during processing action signal $s_{out}^{(5)}$, are: $t_0^{(\langle A3,S3\rangle)} = 5$, $E_0^{(\langle A3,S3\rangle)} = E^{(5)} = 1$. New values $\nu^{(h,\alpha)} = \nu^{(\langle A3,S3\rangle,4)}$, $\omega^{(h,\alpha)} = \omega^{(\langle A3,S3\rangle,4)}$, and $H^{(h,\alpha)} = H^{(\langle A3,S3\rangle,4)}$ are computed in the following way on the basis of previous zero values:

$$\begin{aligned} \nu^{(h,\alpha)} &:= \nu^{(\langle A3,S3\rangle,4)} + 1 = 0 + 1 = 1; \\ \omega^{(h,\alpha)} &:= \omega^{(\langle A3,S3\rangle,4)} + t - t_0^{(\langle A3,S3\rangle)} = 0 + 9 - 5 = 4; \\ H^{(h,\alpha)} &:= H^{(\langle A3,S3\rangle,4)} + E^{(9)} - E_0^{(\langle A3,S3\rangle)} = 0 + 2 - 1 = 1. \end{aligned}$$

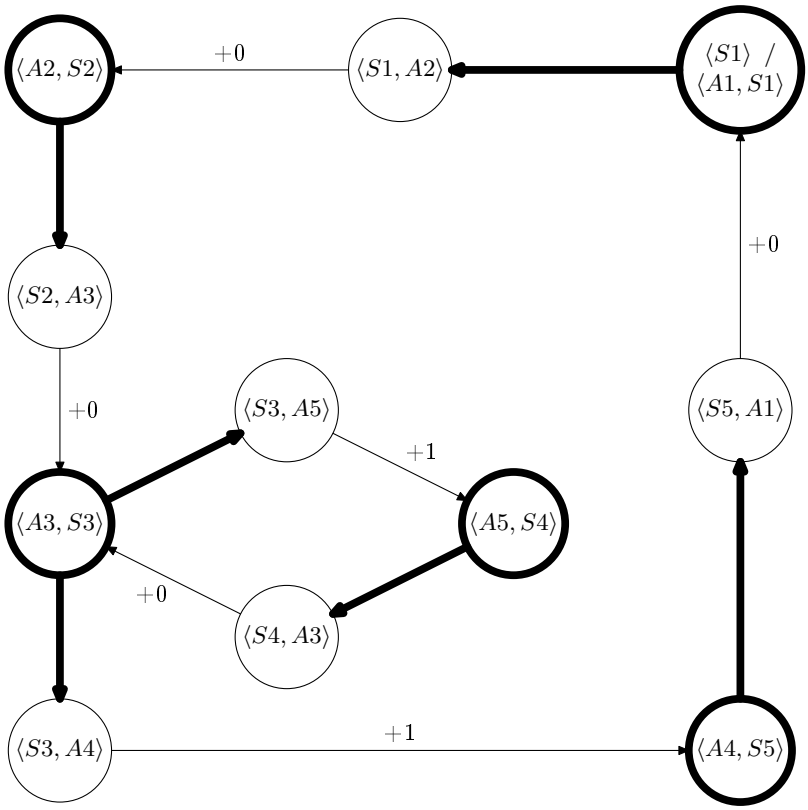

Figure 1: example of transition graph between system states

Because

$$\frac{t}{E^{(t)}} = \frac{9}{2},\ \frac{H^{(\langle A3,S3\rangle,4)}}{\omega^{(\langle A3,S3\rangle,4)}} = \frac{1}{4},\ \frac{H^{(\langle A3,S3\rangle,5)}}{\omega^{(\langle A3,S3\rangle,5)}} = \frac{1}{2},\ \bar{\kappa}_4 = \frac{2\omega^{(\langle A3,S3\rangle,4)}}{\nu^{(\langle A3,S3\rangle,4)}} = 8,\ \bar{\kappa}_5 = \frac{2\omega^{(\langle A3,S3\rangle,5)}}{\nu^{(\langle A3,S3\rangle,5)}} = 4,$$

$$\begin{aligned}\sqrt{\bar{\kappa}_4}\left(\sqrt{\bar{\kappa}_4}+\sqrt{\bar{\kappa}_4+1}\right)(n_{out}-1) &= \quad 8\left(4+3\sqrt{2}\right) \quad \approx 65.9411,\\ \sqrt{\bar{\kappa}_5}\left(\sqrt{\bar{\kappa}_5}+\sqrt{\bar{\kappa}_5+1}\right)(n_{out}-1) &= \quad 8\left(2+\sqrt{5}\right) \quad \approx 33.8885,\end{aligned}$$

we get $F\left(\langle A3,S3\rangle,4,9\right) \approx 65.9411^{\frac{9}{2}\cdot\frac{1}{4}} \approx 111.314$, $F\left(\langle A3,S3\rangle,5,9\right) \approx 33.8885^{\frac{9}{2}\cdot\frac{1}{2}} \approx 2770.88$, $F\left(\langle A3,S3\rangle,z,9\right) = 1$ for $z \in \{1,2,3\}$.

3 The Composition of Systems

The system is oriented to a situation when informational signals that arrive at its input at every step of interaction with the environment are approximations to environment states. Since in practice informational signals usually do not correspond to environment states, but correspond to transitions of the environment from one state to another, this greatly reduces efficiency of system operation. To increase the efficiency, homogeneous system can be divided into two subsystems, the first of which identifies current environment state, and the second generates an optimal action based on that state.

Each step of interaction of the first subsystem with the environment and with the second subsystem is described by triplet $\left\langle \left(s_{out}^{(t-1)}-1\right)n_{in}+s_{in}^{(t)}, \Delta E^{(t)}, s_{st}^{(t)} \right\rangle$ and implies accepting signal $\left(s_{out}^{(t-1)}-1\right)n_{in}+s_{in}^{(t)} \in \overline{1, n_{in}\cdot n_{out}}$ composed of an action signal emitted by the system at the previous step (let us assume that $s_{out}^{(0)}=1$) and of an informational signal accepted from the environment at current step and emitting *environment state signal* $s_{st}^{(t)} \in \overline{1, n_{st}}$ to the second subsystem. To support finite automaton logic, when emitted signal $s_{st}^{(t)}$ is based on signals $s_{st}^{(t-1)}$ and $\left(s_{out}^{(t-1)}-1\right)n_{in}+s_{in}^{(t)}$, condition $N=2$ needs to be true.

Each step of interaction of the second subsystem with the first subsystem and the environment is described by triplet $\left\langle s_{st}^{(t)}, \Delta E^{(t)}, s_{out}^{(t)} \right\rangle$ and implies accepting an environment state signal from the first subsystem and emitting a system action signal to the environment.

The principle of operation of the subsystems is the algorithm described, with a number of changes made.

Changes common to both subsystems are changes directed to increase the coherence of subsystems operation. Let at some step current action choice state of the first subsystem was state h_1, current action choice state of the second subsystem has become state h_2, then the whole system state is pair $\langle h_1, h_2 \rangle$. The system can make a cycle of length l and get into state $\langle h_1, h_2 \rangle$ again if both subsystems emit signal sequences of length l that have caused the system to change its state from the previous state $\langle h_1, h_2 \rangle$ to the current one $\langle h_1, h_2 \rangle$. If during the process of transition of the system from the previous state $\langle h_1, h_2 \rangle$ to the current one $\langle h_1, h_2 \rangle$ the first subsystem got into state h_1 each time after emitting l_1 signals, and the second subsystem got into state h_2 each time after emitting l_2 signals, then l will be the least common multiple of numbers l_1 and l_2. Emitting action signals by the subsystems is made on the basis of cycle length of the whole system with an object of approximate repeating that cycle.

As numbers l_1 and l_2 can be fractional, for approximate cycle length of the whole system, which does not depend on additional parameters, the product of those numbers can be taken. If the first subsystem is in action choice state h_1, and an action signal being chosen is z, then to compute cycle length of the whole system, for approximation to the value of l_1 average length of cycles of type $\langle h_1, z \rangle$ can be taken, and for approximation to the value of l_2 average cycle length of the second subsystem can be taken. If the average length of cycles of type $\langle h_1, z \rangle$ for the first subsystem is equal to κ_1 substeps, the average cycle length of the second subsystem is equal to κ_2 substeps, and an action signal chosen by the first subsystem is z, then approximately after $\bar{\kappa} = \kappa_1 \kappa_2$ substeps the system will be found in state $\langle h_1, h_2 \rangle$ again. Similar method of computing cycle period of the whole system is used by the second subsystem.

Thus, to the list of assignments made during transition to an action choice state when $2t - 1 \geq N$, increments of the number ν_c and of the total length ω_c of cycles processed are added (at the beginning of subsystem operation $\nu_c = 0$ and $\omega_c = 0$), and the full list of assignments takes the following form:

$$\begin{aligned}
\nu^{(h,\alpha)} &:= \nu^{(h,\alpha)} + 1; \\
\omega^{(h,\alpha)} &:= \omega^{(h,\alpha)} + t - t_0^{(h)}; \\
H^{(h,\alpha)} &:= H^{(h,\alpha)} + E^{(t)} - E_0^{(h)}; \\
\nu_c &:= \nu_c + 1; \\
\omega_c &:= \omega_c + t - t_0^{(h)}.
\end{aligned}$$

For the formula of relative probability of action choice (when $\nu_c > 0$) we should now take

$$\bar{\kappa} = \frac{2\omega^{(h,z)}}{\nu^{(h,z)}} \cdot \frac{2\omega_c}{\nu_c},$$

where ω_c and ν_c pertain to another subsystem.

Besides the changes described, which are common to both subsystems, additional changes were made for the algorithm of operation of the environment state identification subsystem, directed to increase the product of probabilities of occurrences of triplets $\left\langle s_{st}^{(t-1)}, \left(s_{out}^{(t-1)} - 1\right) n_{in} + s_{in}^{(t)}, s_{st}^{(t)} \right\rangle$ in the history of accepted and emitted signals, which correspond to state transitions $s_{st}^{(t-1)} \to s_{st}^{(t)}$, and therefore to increase the probability of a state sequence for an input signal sequence.

To the effect, component $E_1^{(h)}$, which holds the logarithm of product of probabilities of triplets occurred in the history of accepted and emitted signals before step $t_0^{(h)}$ inclusive, was added to condition vectors for action choice states, and component $H_1^{(h,z)}$ with initial value 0, which holds the logarithm of product of probabilities of triplets occurred over cycles of type $\langle h, z \rangle$, was added to cycle statistics vectors. Let us point to the fact that the use of the logarithm of augment value transforms spur augment from multiplicative to additive one and therefore suitable for application in the algorithm. Condition vectors for action choice states take form

$$\left\langle \alpha^{(h)}, t_0^{(h)}, E_0^{(h)}, E_1^{(h)}, \psi^{(h)} \right\rangle.$$

Cycle statistics vectors take form

$$\left\langle \nu^{(h,z)}, \omega^{(h,z)}, H^{(h,z)}, H_1^{(h,z)} \right\rangle.$$

To the list of assignments made during a transition to an action choice state when $2t - 1 \geq N$, increment of logarithm E_m of product of probabilities of triplets occurred in the history of accepted and emitted signals since the beginning of subsystem operation (when $E_m = 0$) and increment of $H_1^{(h,\alpha)}$ by the difference between E_m and $E_1^{(h)}$ were added. The value of E_m is incremented by the logarithm of probability of triplet occurrence, computed as the ratio of the number of triplet occurrences $\nu^{(h,\alpha)}$ before step $t_0^{(h)}$ inclusive to the value of $t_0^{(h)} - N/2$. The full list of assignments made during a transition to an action choice state takes the following form:

$$\begin{aligned}
\nu^{(h,\alpha)} &:= \nu^{(h,\alpha)} + 1; \\
\omega^{(h,\alpha)} &:= \omega^{(h,\alpha)} + t - t_0^{(h)};
\end{aligned}$$

$$
\begin{aligned}
H^{(h,\alpha)} &:= H^{(h,\alpha)} + E^{(t)} - E_0^{(h)};\\
\nu_c &:= \nu_c + 1;\\
\omega_c &:= \omega_c + t - t_0^{(h)};\\
E_m &:= E_m + \ln \frac{2\nu^{(h,\alpha)}}{2t_0^{(h)} - N};\\
H_1^{(h,\alpha)} &:= H_1^{(h,\alpha)} + E_m - E_1^{(h)}.
\end{aligned}
$$

When two types of spur are used, let us take for the common relative probability of action choice the product of relative probability of action choice calculated for the first type of spur and relative probability of action choice calculated for the second type of spur. Following the argumentation used when composing the formula of relative probability F of action choice, let us introduce an exponential multiplier into the formula, which increases as the product of probabilities of triplets occurred over cycles of type $\langle h, z\rangle$ increases. That multiplier goes to the summand of exponent, and the formula takes the form

$$
F_2(h,z,t) = B(h,z)^{\frac{tH^{(h,z)}}{\left|E^{(t)}\right|\omega^{(h,z)}}} \cdot B(h,z)^{\frac{tH_1^{(h,z)}}{\left|E_m\right|\omega^{(h,z)}}} = B(h,z)^{\frac{t}{\omega^{(h,z)}}\left(\frac{H^{(h,z)}}{\left|E^{(t)}\right|} + \frac{H_1^{(h,z)}}{\left|E_m\right|}\right)}.
$$

After emitting environment state signal $s_{st}^{(t)}$ by the subsystem, when $2t-1 \geq N$, additional assignment $E_1^{(h)} := E_m$ is made, and the full list of assignments takes form

$$
\begin{aligned}
\alpha^{(h)} &:= s_{st}^{(t)};\\
t_0^{(h)} &:= t;\\
E_0^{(h)} &:= E^{(t)};\\
E_1^{(h)} &:= E_m.
\end{aligned}
$$

4 Testing

System operation testing was performed using random deterministic finite automatons with various numbers of input and output signals and states. Action signals of the system were the input signals of automaton. Testing was performed in modes when informational signals of the system were:

1) automaton states, $N = 1$;

2) automaton output signals, $N = 2$;

3) automaton output signals; the system was comprised of the environment state identification subsystem with $N = 2$ and of the optimal action generation subsystem with $N = 1$.

System spur increment equal to 1 was performed in case of emitting output signal 1 by the automaton. Automatons generated did not have transitions to the same state triggered by some input signal with simultaneous emitting output signal 1. For comparison with value *earned* of spur received by the system during interaction with the automaton according to the algorithm, value *random* of spur received by the system as a result of random interaction with the automaton (when action signals were equiprobably chosen from the set of allowed signals) was used.

For each automaton generated, the maximum amount of spur *maximal*, which the system would receive during the most optimal interaction with the automaton, was determined. To do that, search of a cycle, continuous repeating of which gives the maximum amount of spur to the system, was performed in an automaton state graph, which had been provided to be connected one. To make that search feasible, a random automaton with connected graph was first simplified by replacing transitions from (source) states to other states with transitions to source states until graph connectivity is preserved, and if resulting automaton had more than 100 cycles, then the attempt to generate an automaton acceptable for the test was repeated.

System operation efficiency was evaluated on the basis of total values *earned*, *random*, and *maximal*, accumulated as a result of 200 passes of the algorithm; for each pass new random automaton was used. To evaluate efficiency, values

$$
\begin{aligned}
efr &= \frac{earned}{random} \cdot 100\%;\\
efa &= \frac{earned - random}{maximal - random} \cdot 100\%.
\end{aligned}
$$

were computed. Let us denote efr as relative efficiency and efa as actual efficiency.

The results of testing algorithm operation for homogeneous systems are represented in Table 2, for systems comprised of two subsystems—in Table 3. The plots of dependency of efa on n_{ev} for some testing modes for systems comprised of two subsystems are represented in Figures 2-4. Each point on a plot corresponds to a value of efa evaluated on basis of 100 algorithm passes, and thick line is regression.

5 Conclusions

Algorithm testing results can be considered as a departure in comparison of various algorithms of interaction of a system with the environment represented by a deterministic finite automaton. Using various interpretations for the concept of spur to get the maximum amount of which the system is directed to, such algorithms can be adapted to solve various tasks of optimal behavior search. There can be several spur types used simultaneously in the algorithm described.

The environment state identification subsystem, which operates according to the algorithm, solves the task of approximate identification of current environment state on the basis of signals sent to the environment and received from the environment. The model of environment states is built by taking into account the goal of the system to get the maximum amount of spur from the environment. The advantage of the algorithm is that the process of system training goes simultaneously with the process of generating actions by the system, and that each step of interaction with the environment is performed by the system for a short fixed time.

The idea of the algorithm can be used in control and human interaction systems that operate in real-time.

Acknowledgements

I wish to express thanks to my chief at Unicorn Ltd. (Kiev, Ukraine) Maxim Navrotsky, Ph.D., for a number of valuable remarks when he critically reviewed the preliminary version of this paper.

References

[1] Yoav Freund, Michael Kearns, Dana Ron, Ronitt Rubinfeld, Robert E. Schapire, and Linda Sellie. Efficient Learning of Typical Finite Automata from Random Walks. *In Proceedings of the 24th Annual ACM Symposium on Theory of Computing*, 315–324, 1993.

[2] Aaron F. Stanton, Richard E. Bleil, and Sabre Kais. A New Approach to Global Minimization. *Journal of Computational Chemistry*, Vol. 18, No. 4, 594–599, 1997.

[3] Bradley J. Clement. Learning Harmonic Progression Using Markov Models. *Genetic Algorithms in Search, Optimization, and Machine Learning*, 1998.

Num. of automaton states	n_{in}	n_{out}	n_{ev}	efr, %	efa, %
Informational signals type: automaton states; $N = 1$					
10	10	10	5,000	849.8	77.9
10	10	10	10,000	851.5	78.2
20	20	20	5,000	1,608.6	69.3
20	20	20	10,000	1,697.4	70.9
30	30	30	15,000	2,503.5	66.9
30	30	30	30,000	2,557.1	70.1
40	40	40	35,000	3,594.1	68.7
40	40	40	70,000	3,735.6	69.7
50	50	50	65,000	4,799.3	68.7
50	50	50	130,000	4,922.0	69.1
Informational signals type: automaton output signals; $N = 2$					
10	10	10	5,000	293.9	20.2
10	10	10	10,000	313.4	22.2
20	20	20	5,000	379.4	12.8
20	20	20	10,000	415.2	14.0
30	30	30	15,000	475.3	10.4
30	30	30	30,000	519.4	12.0
40	40	40	35,000	577.7	9.4
40	40	40	70,000	607.1	9.7
50	50	50	65,000	695.6	8.7
50	50	50	130,000	704.7	8.7

Table 2: results of testing algorithm operation for homogeneous systems

Num. of automaton states	n_{in}	n_{out}	n_{ev}	n_{st}	efr, %	efa, %
10	10	10	10,000	10	445.3	36.3
10	10	10	40,000	20	618.9	52.2
20	20	20	160,000	20	647.8	24.7
20	20	20	640,000	40	1,109.2	46.0
30	30	30	810,000	30	723.7	17.5
30	30	30	3,240,000	60	1,448.7	36.1
40	40	40	2,560,000	40	844.1	14.7
40	40	40	10,240,000	80	1,861.8	33.8
50	50	50	6,250,000	50	1,048.6	14.2
50	50	50	25,000,000	100	2,106.0	29.4

Table 3: results of testing algorithm operation for subsystem compositions

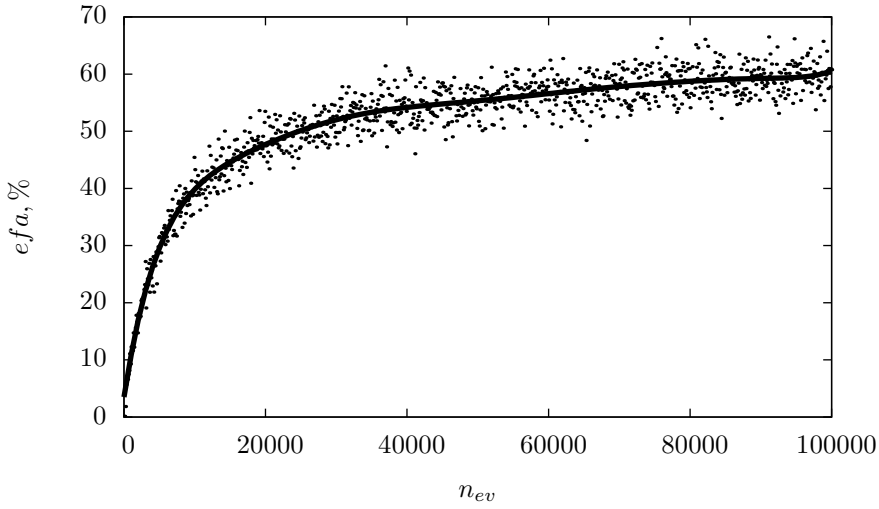

Figure 2: automaton with 10 states, $n_{in} = 10$, $n_{out} = 10$, $n_{st} = 20$

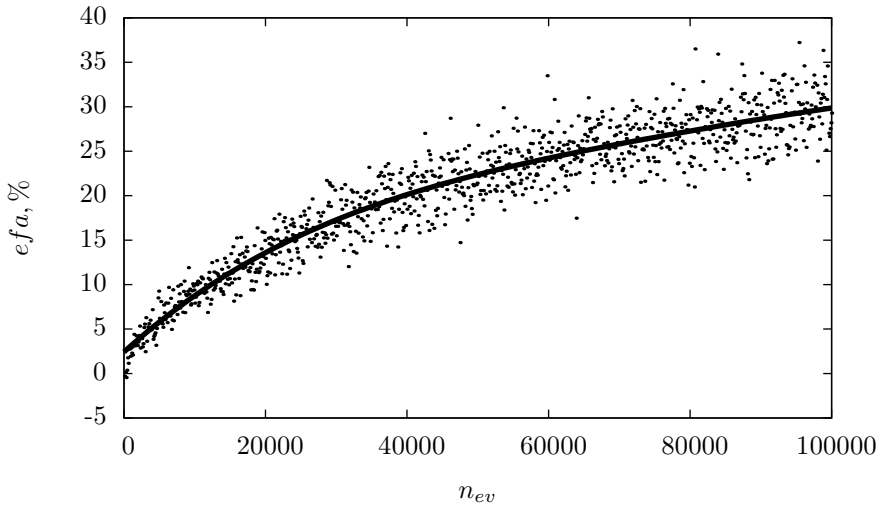

Figure 3: automaton with 20 states, $n_{in} = 20$, $n_{out} = 20$, $n_{st} = 40$

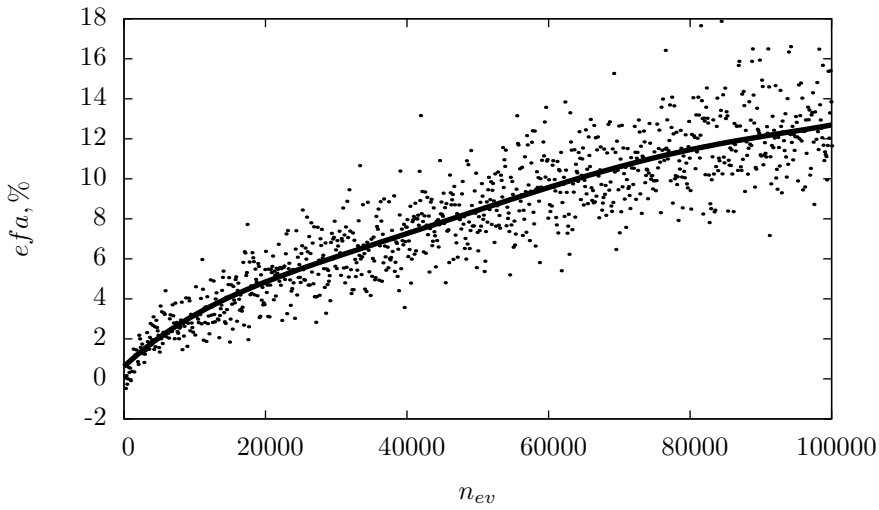

Figure 4: automaton with 30 states, $n_{in} = 30$, $n_{out} = 30$, $n_{st} = 60$