Collective Intelligence for Optimization (Summary)

David H. Wolpert
NASA Ames Research Center
Moffett Field, CA 94035
dhw@ptolemy.arc.nasa.gov

Kagan Tumer
NASA Ames Research Center
Moffett Field, CA 94035
kagan@ptolemy.arc.nasa.gov

November 4, 1999

A "COllective INtelligence" (COIN) is a distributed set of interacting reinforcement learning (RL) algorithms designed so that their collective behavior optimizes a global utility function. One can cast a COIN as a multi agent system (MAS) where:

i) there is little to no centralized communication or control;

ii) each agent runs a 'greedy' Reinforcement Learning (RL) algorithm, in an attempt to increase its own utility;

iii) there is a well-specified global objective function that rates the full system.

Rather than use a conventional modeling approach (e.g., model the system dynamics, and hand-tune agents to cooperate), we aim to solve the COIN design problem implicitly, via the "adaptive" character of th e RL algorithms of each of the agents. This approach introduces an entirely new, profound design problem: Assuming the RL algorithms are able to achieve high rewards, what reward functions for the individual agents will, when pursued by those agents, result in high world utility? In other words, what reward functi ons will best ensure that we do not have phenomena like the tragedy of the commons , Braess's paradox, or the liquidity trap?

An example of a naturally occurring COIN is a capitalist economy. One can declare 'world utility' to be a time average of the Gross Domestic Product (GDP) of the country in question. The reward functions for the human agents can then be the achievements of their personal goals (usually involving personal wealth to some degree). To achieve high world utility in any COIN it is necessary to avoid phenomena like the Tragedy of the Commons (TOC), in which individual avarice works to lower global utility. One way to avoid such phenomena is by modifying the agents' utility functions. In the context of capitalist economies, this can be done via punitive legislation. A real world example of an attempt to make just such a modification was the creation of anti-trust regulations designed to prevent monopolistic practices.

In designing a COIN we have more freedom than anti-trust regulators though, in that there is no base-line "organic" local utility function over which we must superimpose legislation-like incentives. Rather, the entire "psychology" of the individual agents is at our disposal, when designing a COIN. This obviates the need in designed COINs for honesty-elicitation ('incentive compatible') mechanisms, like auctions, which form a central component of computational economics.

We have explored a mathematical framework for COIN design, and investigated the (successful) application of that framework in several domains, e.g., optimizing the packet throughput of a telecommunications network [10]. Here we present a summary of a different investigation of our COIN methodology involving a variant of Arthur's "El Farol" bar problem [1, 2, 4, 7, 8, 3], a problem which first arose in economics (see [11, 9] for details). In the bar problem, at each time step each agent independently predicts, based on its previous experience, whether a bar will be too crowded to be "enjoyable" at that time step. The agent then uses this prediction to decide whether attending the bar or not will maximize its local utility. The global objective in this problem is to keep the bar as close to capacity as possible. In this problem the "greedy" nature of the agents can readily thwart the optimization of the world utility.

In the problem we investigated, there are N agents, each picking one of seven nights to attend a bar the following week, a process that is then repeated. In each week, each agent's pick is determined by its predictions of the associated rewards it would receive. Each such prediction in turn is based solely upon the rewards received by the agent in those preceding weeks in which it made that pick.

The world utility is $G(\underline{\zeta}) = \sum_t R_G(\underline{\zeta}_{,t})$, where $R_G(\underline{\zeta}_{,t}) \equiv \sum_{k=1}^{7} \phi_k(x_k(\underline{\zeta}, t))$, $x_k(\underline{\zeta}, t)$ is the total attendance on night k at week t, $\phi_k(y) \equiv \alpha_k y \exp(-y/c)$; and c and the $\{\alpha_k\}$ are real-valued parameters. Intuitively, this G is the sum of the "world rewards" for each night in each week. Our choice of $\phi_k(.)$ means that when too few agents attend some night in some week, the bar suffers from lack of activity and therefore the world reward is low. Conversely, when there are too many agents the bar is overcrowded and the reward is again low.

Each agent η has a 7-dimensional vector representing its estimate of the reward it would receive for attending each night of the week. At the end of each week, the component of this vector corresponding to the night just attended is proportionally adjusted towards the actual reward just received. At the beginning of the succeeding week, to trade off exploration and exploitation, η picks the night to attend randomly using a Boltzmann distribution with 7 energies $\epsilon_i(\eta)$ given by the components of η's estimated rewards vector, and with a temperature decaying in time. This learning algorithm is similar to Claus and Boutilier's independent learner algorithm [5].

We considered three agent reward functions, using the same learning parameters (learning rate, Boltzmann temperature, decay rates, etc.) for each. The first reward function was $\gamma_\eta = G \; \forall \eta$, i.e., agent η's reward function equals R_G. The other two reward functions are: $R_{UD;\eta} \equiv \phi_{d_\eta}(x_{d_\eta}(\underline{\zeta}, t))/x_{d_\eta}$, $R_{WL;\eta}(\underline{\zeta}_{,t}) \equiv R_G - R_G(CL_\eta)$, where d_η is the night picked by η, and CL_η is a function derived from COIN theory.

The R_{UD} reward is a "natural" reward function to use; each night's total reward is uniformly divided among the agents attending that night. R_G is the "team game" reward function that has been investigated in the MAS community [6]; every agent gets the world reward as its reward signal. R_{WL} is the reward function recommended by COIN theory.

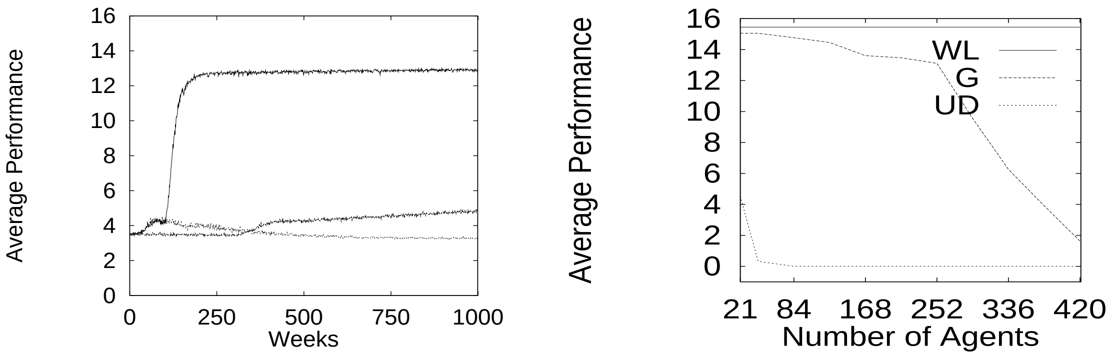

Figure 1: Average world reward convergence and scaling properties. In both plots the top curve is R_{WL}, middle is R_G, and bottom is R_{UD}.

The left-hand figure in Figure 1 graphs world reward value as a function of time, averaged over 50 runs, for all three reward functions. The naive choice of R_{UD} actually leads to *deterioration* of performance with time. Performance with R_G eventually converges to the global optimum. Systems using R_{WL} also converged to optimal performance, but far faster (30 times as quickly). This slow convergence of systems using R_G is a result of the reward signal being "diluted" by the large number of agents in the system. The right-hand figure in Figure 1 shows how performance at $t = 2000$ scales with N. Systems using R_{UD} perform poorly regardless of N. Systems using R_G perform well when N is low. As N increases however, it becomes increasingly difficult for the agents to extract the information they need from R_G.

In conclusion, the COIN framework summarized in this article addresses large distributed computational optimization tasks from a novel perspective, one that works much better than the other systems we investigated.

References

[1] W. B. Arthur. Complexity in economic theory: Inductive reasoning and bounded rationality. *The American Economic Review*, 84(2):406–411, May 1994.

[2] G. Caldarelli, M. Marsili, and Y. C. Zhang. A prototype model of stock exchange. *Europhys. Letters*, 40:479–484, 1997.

[3] A. Cavagna, Garrahan J. P., I. Giardina, and D. Sherrington. A thermal model for adaptive competition in a market. preprint cond-mat/9903415 v.3, July 1999.

[4] D. Challet and Y. C. Zhang. On the minority game: Analytical and numerical studies. *Physica A*, 256:514, 1998.

[5] C. Claus and C. Boutilier. The dynamics of reinforcement learning cooperative multiagent systems. In *Proceedings of the Fifteenth National Conference on Artificial Intelligence*, pages 746–752, June 1998.

[6] R. H. Crites and A. G. Barto. Improving elevator performance using reinforcement learning. In D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo, editors, *Advances in Neural Information Processing Systems - 8*, pages 1017–1023. MIT Press, 1996.

[7] M. A. R. de Cara, O. Pla, and F. Guinea. Competition, efficiency and collective behavior in the "El Farol" bar model. preprint cond-mat/9811162 (to appear in European Physics Journal B), November 1998.

[8] W. A. Sethares and A. M. Bell. An adaptive solution to the El Farol problem. In *Proceedings. of the Thirty-Sixth Annual Allerton Conference on Communication, Control, and Computing*, Allerton, IL, 1998. (Invited).

[9] D. H. Wolpert and K. Tumer. An Introduction to Collective Intelligence. In J. M. Bradshaw, editor, *Handbook of Agent technology*. AAAI Press/MIT Press, 1999. to appear.

[10] D. H. Wolpert, K. Tumer, and J. Frank. Using collective intelligence to route internet traffic. In *Advances in Neural Information Processing Systems - 11*. MIT Press, 1999.

[11] D. H. Wolpert, K. Wheeler, and K. Tumer. General principles of learning-based multi-agent systems. In *Proceedings of the Third International Conference of Autonomous Agents*, pages 77–83, 1999.