Distributed On-line Bayesian Search

Alfredo Garcia
University of Virginia
agarcia@virginia.edu

Enrique Campos
George Washington University
ecamposn@gwu.edu

Chenyang Li
University of Virginia

Abstract

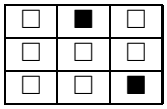
In this paper, we outline the basis for a new distributed Bayesian search scheme in which all Bayesian decision makers recognize the same performance objective but do not possess the ability to communicate with each other. Coordination among the players is achieved indirectly by tracking search performance. This scheme is ideal for unmaned vehicles searching in extreme environments where extensive bilateral communication between actuators is not feasible. Preliminary results suggests our new search scheme is scalable and can easily be adapted for tracking moving targets.

1 Introduction

Rather than (ex-ante) specifying rules that ensure the ability to react to changes in a coordinated manner, the distributed Bayesian search scheme presented in this paper makes Bayesian decision makers grope for optimality by dynamically experimenting with courses of action that are aimed at improving the system's historical performance. The evaluation of system's performance, which is available to all decision makers through summary statistics and/or highly aggregated measurements is obtained through online sampling. Our preliminary results indicate that coordinated response emerges endogenously as a result of "learning", i.e. convergence to a certain joint action profile. Two important features support the scalability of our proposed scheme. First, the approach does not require bilateral communication amongst decision makers but simply the general dissemination & updating of *aggregate* statistics of system performance. Secondly, the updating of decisions can be made in parallel and/or asynchronously.

2 Illustrative Example

Let us start by introducing our ideas through a simple example. Suppose two underwater unmanned vehicles were tasked to find and destroy two known underwater mines placed in stationary (and different) locations in a 3×3 area. For example, they may be located in the upper-center and lower-right regions, as shown below:



Let $X = \{1, 2, 3\} \times \{1, 2, 3\}$ represent the set of possible coordinates. Also, let S denote the set of feasible locations for the two targets, i.e. $S = \{s = (s_1, s_2) \mid s_1, s_2 \in X\}$. Upon perfect or imperfect knowledge of target locations, the agents are to decide (independently) which location to probe. We denote by $a = (a_1, a_2)$ the probe choices made by the agents (where $a_i \in X$). If s_1^* and s_2^* denote the target locations, the loss function is

$$L(a; s) = \begin{cases} 0 & a_i = s_i^* \text{ and } a_j = s_j^* \\ 1 & a_i = s_i^* \text{ and } a_j \neq s_j^* \\ 2 & a_i \neq s_i^* \text{ and } a_j \neq s_j^* \end{cases}$$

2.1 A Coordination Game

Let us first assume that the target locations are perfectly known to the two agents. Without the ability to communicate, the choice of location becomes a "coordination" game: i.e. the choices (a_1, a_2) and (a_1', a_2') where $a_1 = a_2' = (1, 2)$ and $a_2 = a_1' = (3, 3)$ are both Nash equilibria; however, without bilateral communication, players are unable to determine which of these two outcomes should be implemented. In fact, failure to coordinate may lead to outcomes such as (a_1'', a_2'') where $a_1'' = a_2'' = (3, 3)$ in which one of the targets is not destroyed. As the number of targets and/or agents increases, the inability to communicate results in a higher likelihood of coordination failures.

Let us now consider the more interesting case in which target locations are not known. Let $\mu^0(x)$ be the a priori probability that a target is located at $x \in X$. Let α and β denote the conditional probability of obtaining "false positives" and "false negatives", respectively. Let us denote by Z_1 and Z_2 the probe results for the choice of locations

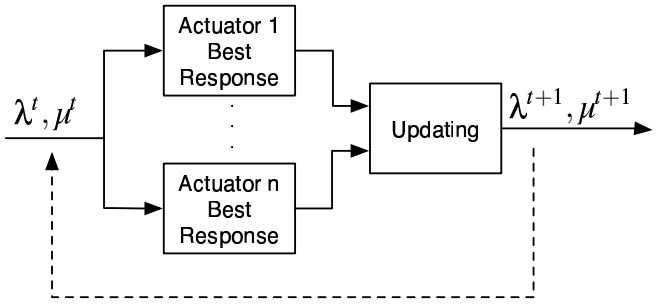


Figure 1. Schematic for Learning Algorithm.

$a = (a_1, a_2)$, where $Z_i = 1$ if a "positive" result is obtained and $Z_i = 0$, otherwise. If $x \neq a_1$ and $x \neq a_2$ then $\mu^1(x) = \mu^0(x)$. Otherwise, if $a_i \neq a_j$ and $x = a_i$ then

$$\mu^1(x) = \begin{cases} \frac{(1-\beta)\mu^0(x)}{\alpha(1-\mu^0(x))+(1-\beta)\mu^0(x)} & Z_i = 1 \\ \frac{\beta\mu^0(x)}{(1-\alpha)(1-\mu^0(x))+\beta\mu^0(x)} & Z_i = 0 \end{cases} \quad (1)$$

Finally, if $x = a_1 = a_2$ then

$$\mu^1(x) = \begin{cases} \frac{(1-\beta)^2\mu^0(x)}{\alpha^2(1-\mu^0(x))+(1-\beta)^2\mu^0(x)} & Z_1 = Z_2 = 1 \\ \frac{\beta(1-\beta)\mu^0(x)}{\alpha(1-\alpha)(1-\mu^0(x))+\beta(1-\beta)\mu^0(x)} & Z_i = 1 \text{ and } Z_j = 0 \\ \frac{\beta^2\mu^0(x)}{(1-\alpha)^2(1-\mu^0(x))+\beta^2\mu^0(x)} & Z_1 = Z_2 = 0 \end{cases} \quad (2)$$

2.2 A Learning Algorithm

The basic structure of the algorithm we propose is illustrated in Figure 1 and is based on the following two principles:

- *(Decentralized and Parallel Decision-making)* Each agent (i.e. actuator) computes at each iteration the best location to probe, say $a_i^t \in X$, vis-a-vis the empirical frequency of past probe choices. In other words, each player aims at minimizing expected (Bayes) loss assuming the next iteration (random) probes by other players are distributed according to $\lambda_{-i}^t = \lambda^t - \lambda_i^t$, where $\lambda^t(x)$ denotes the frequency distribution that location $x \in X$ is probed by all the agents, $\lambda_i^t(x)$ is agent i's empirical frequency of probing location $x \in X$ alone and the current beliefs on target location are given by μ^t. Formally,

$$a_i^t \in \arg\min_{a_i} \sum_{s \in S} \sum_{a_{-i} \in X} L(a_i, a_{-i}; s) \lambda_{-i}^t(a_{-i}) \mu^t(s) \quad (1)$$

- *(Sampling)* Once a sample is obtained, the updated probability distribution μ^{t+1} is constructed as described in (1) and (2) above and λ^t, λ_i^t are updated as follows:

$$\lambda^{t+1}(x) = \lambda^t(x) + \frac{1}{t+1}(\mathbf{1}_{\{x=a_i^t\}} - \lambda^t(x))$$
$$\lambda_i^{t+1}(x) = \lambda_i^t(x) + \frac{1}{t+1}(\mathbf{1}_{\{x=a_i^t, x \neq a_j^t\}} - \lambda_i^t(x))$$

where

$$\mathbf{1}_{\{x=a_i^t\}} = \begin{cases} 1 & x = a_i^t \\ 0 & \text{otherwise} \end{cases}$$
$$\mathbf{1}_{\{x=a_i^t, x \neq a_j^t\}} = \begin{cases} 1 & x = a_i^t \text{ and } x \neq a_j^t \\ 0 & \text{otherwise} \end{cases}$$

The network architecture implicitly represented in Figure 1 is fluid because agents only need to know how often a given location has been probed in the past (i.e. $\lambda^t(x)$, $x \in X$) *regardless of the identity of the agent(s) that executed the probes*, and the updated Bayesian beliefs (i.e. $\mu^t(x)$, $x \in X$). Thus, agents do not need to know the makeup of the group so new agents can enter the network and others can exit (e.g., in a low fuel state), thus providing a flexible and adaptable network for dynamic mission objectives. While bilateral communication is not required in this scheme, agents must be able to access at all times a (possibly, distributed) repository of the "state" of the system (i.e. the values of $\lambda^t(x)$ and $\mu^t(x)$, $x \in X$)

From a theoretical standpoint, the algorithm closely resembles a widely-studied game-theoretic learning algorithm known as *fictitious play* (see [7], [1] and the more recent work reported in [8] and [5]). In the fictitious play algorithm, players compute at each step, their best action or "reply" based on the assumption that other players' actions follow a probability distribution in agreement with the historical frequency of their past action choices. However, joint online sampling has the effect of introducing *correlation* into the players' decision processes. This constitutes a major point of departure from *fictitious play* where players *ignore* correlation and consequently, *do not react to the system's historical performance*. An early illustration of the application to the problem of dynamic traffic routing of the learning algorithm here discussed, is given in Garcia et. al. [2].

2.3 Convergence

Monderer and Shapley [6] proved convergence of fictitious play (in relative frequencies) to the set of Nash equilibria in games where players share a common objective (also referred to in the literature, as games of identical interests). However, in the case where players' decisions are correlated the Monderer and Shapley's result does not apply. As

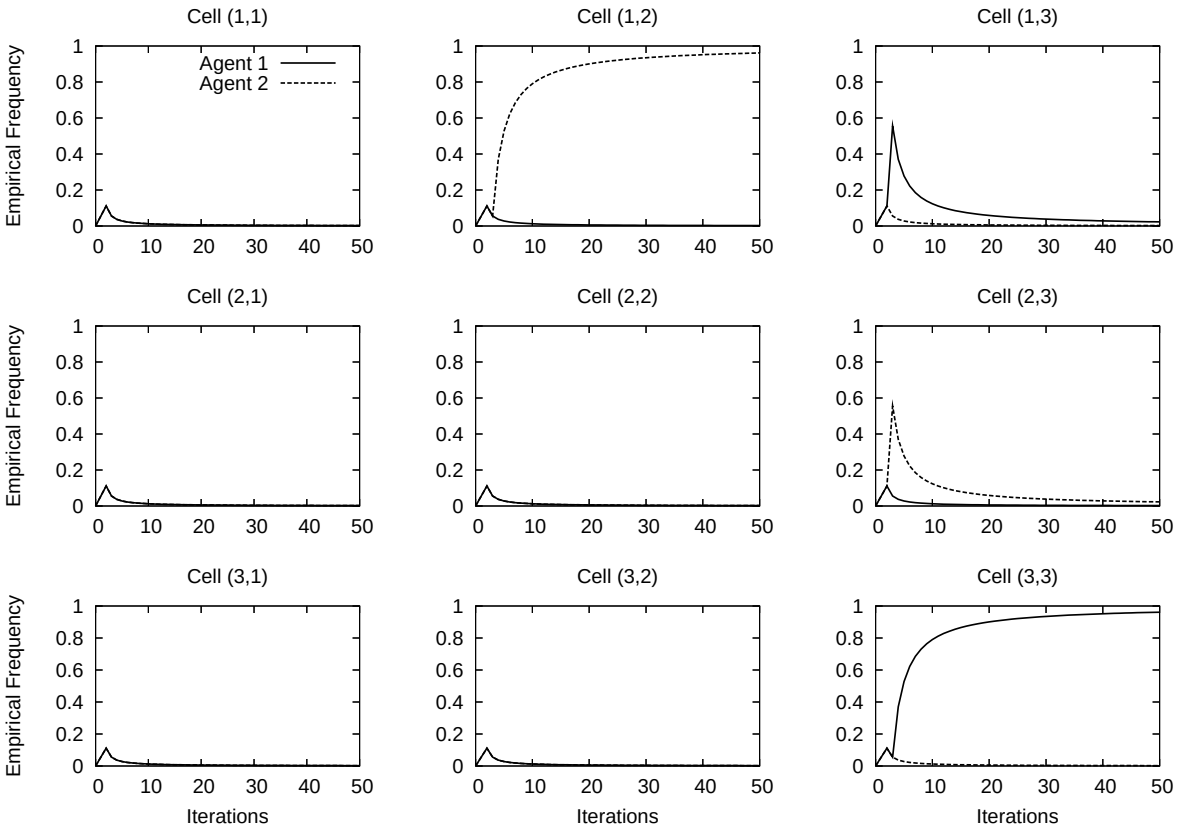


Figure 2. Algorithm's output (i.e. $\lambda^t(x), x \in X$) for illustrative example.

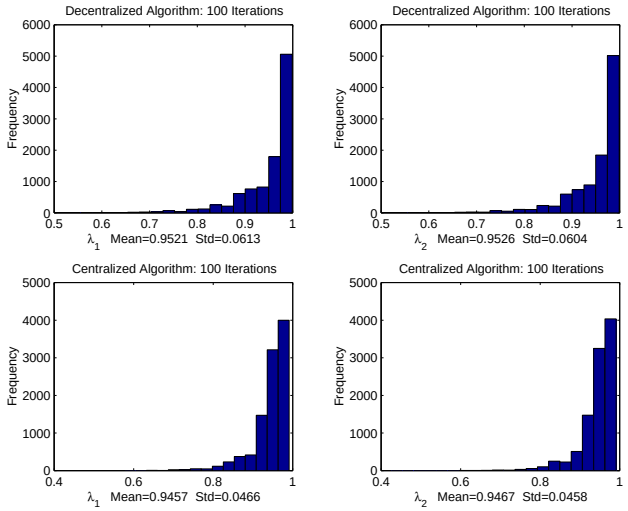


Figure 3. Algorithm's performance compared with centralized optimization (10,000 simulations).

mentioned above, this is precisely the case here since joint online sampling introduces *correlation* into the players' decision processes. Recently, we have been able to prove convergence of the class of algorithms represented in Figure 1 (see, [3]).

In Figure 2, a sample realization of the learning algorithm is depicted. Note how the players achieve coordination after a few iterations (agent 1 settles for the target located on $(3, 3)$ while agent 2 settles for the target at $(1, 2)$).

Figure 3 shows the comparison results for our decentralized decision-making algorithm and the centralized one. In this specific example, after 100 iterations, the decentralized algorithm did a pretty good job, ending up with a little bit higher average probability of locking the targets but also with a little bit higher standard deviation. However, as the number of targets and agents increases, coordination between agents for centralized decision-making algorithm would be fully taken advantage of and a growing disparity between the two algorithms is expected to appear.

Suppose now that the target location changes in an unpredictable fashion. The learning algorithm can be suitably altered to be able to react to this change if the empirical frequencies are computed with a finite history of play. In Fig-

ure 4, we present the results where *target locations change randomly every 200 iterations* and the empirical frequency is computed with respect to the last 100 plays. The Bayesian updating is modified so that for a parameter $\theta \in (0,1)$ it takes the form,

$$\mu^{t+1} = (1-\theta) f(\mu^t, a^t) + \theta \mu^t$$

where $f(\mu^t, a^t)$ denotes the Bayesian update on μ^t by virtue of probe choices a^t (as described explicitly in (1) and (2)). In the numerical experiments reported in Figure 4, we have chosen θ to be 0.01. Note that the "pulses" representing a target present are quickly followed by exponential-like increase in the choice of that location by *one* (and only *one*) of the agents.

2.4 Scalability

Our scheme naturally addresses some critical aspects of scalability. As stated before, agents only need to know how often a given location has been probed in the past (i.e., $\lambda^t(x)$) and the updated Bayesian beliefs (i.e. $\mu^t(x)$), for $x \in X$. The specific identity of the agent(s) that executed the probes in the past is not required. This implies that memory requirements are dictated by the size of X and not by the number of agents. In Figure 4, we show the algorithm's performance for an increasing number of targets for an equal number of agents, $N = \{5, 10, 20\}$, and where the size of the grid increases squarely in the number of targets, i.e. $\{5^2, 10^2, 20^2\}$. Evidently, convergence to full target coverage slows down linearly as N increases. However, our grid configuration is in a sense a "litmus" test for scalability since it exhibits a high degree of connectivity, which in turn makes coordination more difficult to achieve. In other words, target discovery is faster in network topologies with lower degrees of connectivity (e.g. a "small worlds" topology). Another point worth emphasizing here is the fact that agents are not constrained in their choice of target locations. In a more realistic setting, the set of available locations for probes will depend on the location of the very last probe executed. This feature would speed up convergence to full target coverage.

3 Conclusions

We have presented preliminary results on a new distributed Bayesian search scheme in which all Bayesian decision makers recognize the same performance objective but do not possess the ability to communicate with each other. Our preliminary results indicate the scheme scales up gracefully and can be easily adjusted to allow for online tracking. Our future research will concentrate on the convergence and the quality of limit points. Experimental

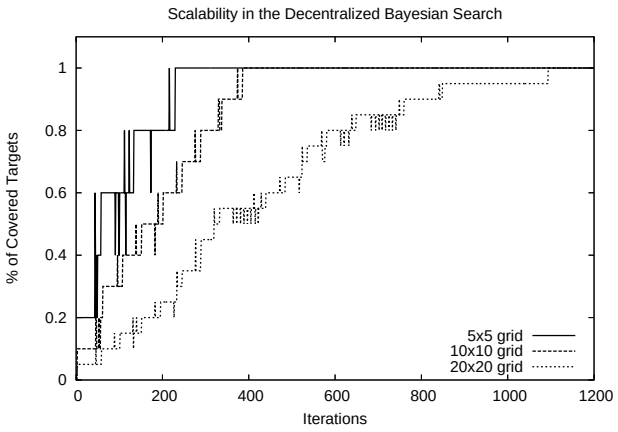


Figure 5. Test for scalability.

trials also show that the accuracy of probe, the weight for the update of the aggregated information, the structure of the loss function affect the convergence rate to full target coverage. Given the number of targets and probe accuracy, calibrating the loss function and the update weight will undoubtedly help to expedite convergence and make the new search scheme more adaptive to the diversity of real world situations.

References

[1] Fudenberg D. and Levine D., 1998. *The Theory of Learning in Games*. MIT Press.

[2] Garcia A., Reaume D. and Smith R.L., 2000. Fictitious Play for Finding System Optimal Routings in Dynamic Traffic Networks. Transportation Research B, Methods, Vol. 34 No. 2 pp 147-156

[3] Garcia A., Patek S. and Sinha S., 2005. A Decentralized Approach to Discrete Optimization via Simulation: Application to Network Flow, Department of Systems & Information Engineering Technical Report SIE-050001, University of Virginia. (available at *http://www.sys.virginia.edu/techreps/*)

[4] Garcia A., Campos E. and Reitzes J. Dynamic Pricing & Learning in Electricity Markets, Operations Research Vol. 53 No. 2 (2005) pp. 231-241

[5] Lambert T., Epelman M., Smith R. L., 2003. A Fictitious Play Approach to Large-Scale Optimization. *forthcoming* Operations Research.

[6] Monderer D. and Shapley L., 1996. Fictitious Play Property for Games with Identical Interests. Journal of Economic Theory, Vol. 68 No. 1, pp 258-265.

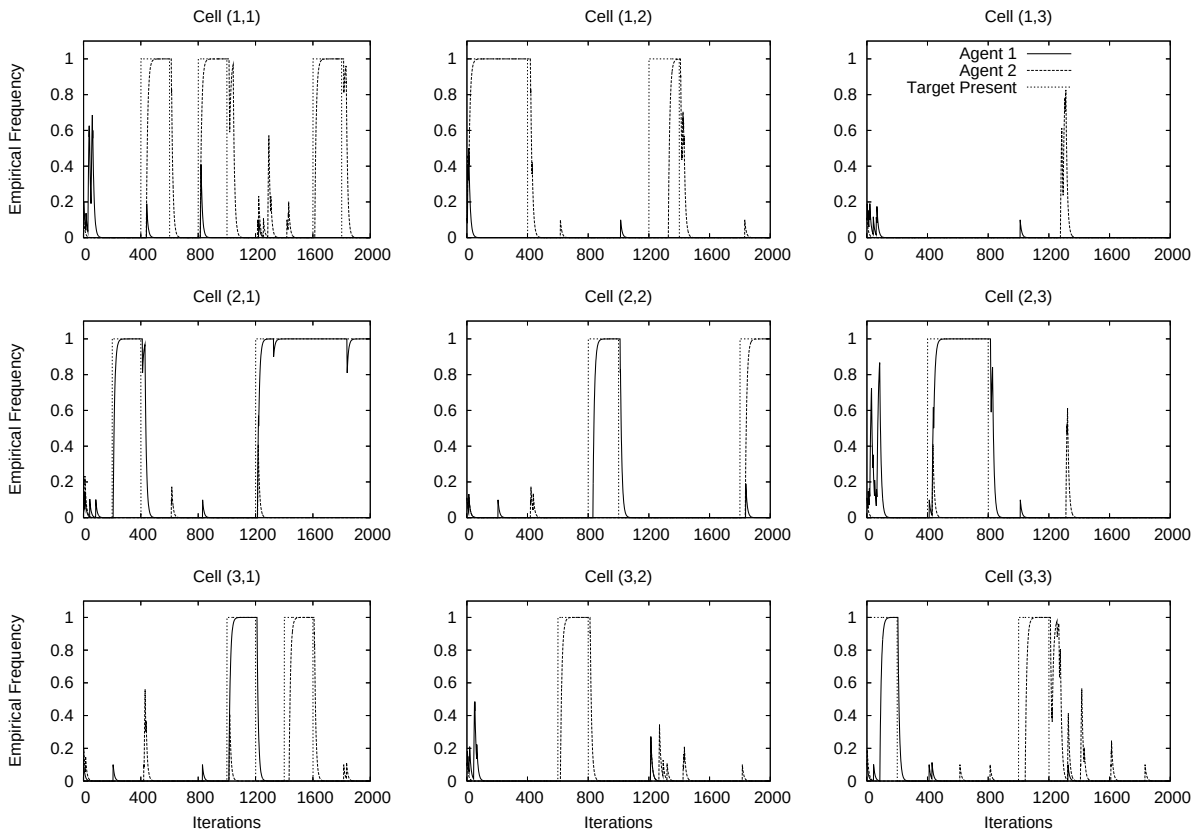


Figure 4. Performance of Algorithm with adaptive capability.

[7] Robinson J. 1951. An Iterative Method of Solving a Game. *Annals of Mathematics*, Vol. 54, No. 2, pp. 296–301.

[8] J.S. Shamma and G. Arslan. 2005. Dynamic Fictitious Play, Dynamic Gradient Play, and Distributed Convergence to Nash equilibria, IEEE Transactions on Automatic Control, Vol. 50 No 3, pp. 312-327