

Convergence of Probability Collectives with adaptive choice of temperature parameters

Michalis Smyrnakis and David S. Leslie

Department of Mathematics University of Bristol, UK

1 Introduction

There are numerous applications of multi-agent systems like disaster management [1], sensor networks [2], traffic control [3] and scheduling problems [4] where agents should coordinate to achieve a common goal. In most of these cases a centralized solution is inefficient because of the scale and the complexity of the problems and thus distributed solutions are required.

When the objective is optimization this is naturally formulated as an n -player game [5] [6]. The agents take an action concerning their environment and from their action they receive a global reward that is the same for all the agents. Then this reward acts as a potential. The same stands for the players of a partnership strategic form game, where players have to take an action and their common global reward depends on the actions of the other players.

Many different learning techniques have been used to solve multi-agent optimization problems such as Q-learning [7], minimax-Q learning [8], opponent modeling [9], WOLF [10] and others. However very few of these have theoretical convergence results. On the other hand game-theoretic algorithms such as adaptive play and fictitious play have been proved to converge although in practice this convergence can be very slow [11, 12]. In this paper we will prove that in discrete action space a variation of Probability Collectives (PCs) [13], converges to an optimum. We do this by relating it to generalised weakened fictitious play [14], which is known to converge in the potential games that are of special interest in multiagent systems.

2 Probability Collectives

Most optimization methods search for an $s \in \mathbf{S}$ that optimizes a reward function $G(s)$. In contrast Probability Collectives search for a product distribution $q(s) = \prod_i q_i(s_i)$, where $q_i(s_i)$ is the probability that agent i chooses an action $s_i \in S_i$, that optimizes the expected value of $G(s)$, $E_q(G(s))$ [13].

We have to encode the constraints that arise from the fact that each q_i must be a probability distribution. These constraints are the following:

$$\begin{aligned} \sum_{s_i \in S_i} q_i(s_i) &= 1 \\ q_i(s_i) &\geq 0 \quad \text{for each } i \quad \text{for each } s_i \in S_i \end{aligned}$$

The first constraint can be encoded using Lagrange multipliers. We can enforce the second constraint using a barrier function ϕ , satisfying $\phi(q_i(s_i)) > 0$ when $q_i(s_i) > 0$ and $\phi(q_i(s_i)) = \infty$ when $q_i(s_i) \leq 0$. We can choose $\phi(q) = \kappa + q \ln q$, where κ is a constant to ensure that $\phi(q) > 0 \forall q > 0$. Since the minimum of $q \ln q$ is $-\frac{1}{e}$ we can set $\kappa = \frac{1}{e} + \epsilon$, where ϵ is an arbitrary small number close to zero. Then the objective function that we want to maximize is the following:

$$L(q, b_i) = E_q(G) - \sum_i \frac{1}{b_i} (r_i - S(q_i)) - \sum_i \lambda_i (\sum q_i(s_i) - 1) \quad (1)$$

where $r_i = \sum_{S_i} \kappa$, $b_i > 0$ are inverse temperature parameters, λ_i are Lagrange multipliers and $S(q_i) = -\sum_{S_i} q_i(s_i) \ln q_i(s_i)$ is Shannon's entropy.

A solution can be found if we search for the critical points of (1), setting its derivative to zero. Brouwer's fixed point theorem ensures that solutions always exist. Their form up to a multiplicative constant is the following:

$$q_i(s_i = m) \propto e^{E_{q_{-i}}(G|s_i=m)b_i} \quad (2)$$

where $E_{q_{-i}}(G|x_i)$ is the expected value of the reward function under the probability distribution $q_1 \times \dots \times q_{i-1} \times q_{i+1} \times \dots \times q_N$ conditional on the value $s_i = m$.

Because there is no easy analytical solution of equation (2) we could search for solutions by iteratively setting:

$$q_i^{t+1} = k_i(q_{-i}^t, b_i^t). \quad (3)$$

$$\text{where } k_i(q_{-i}^t, b_i^t)(s_i = m) = \frac{e^{E_{q_{-i}^t}(G(s)|s_i=m)b_i^t}}{\sum_{y_i} e^{E_{q_{-i}^t}(G(s)|y_i)b_i^t}}$$

However the simultaneous update of all the distributions q_i has as an effect that it is possible to observe thrashing [13], it is worth noting that a similar phenomenon is well known in game theory. This describes the situation where each agent updates his distribution according to the previous values of the other agents' distributions, but since all the agents are changing their distributions there is no guarantee that the new product distribution q^{t+1} will increase the Maxent Lagrangian. To avoid thrashing we instead consider updates of the form:

$$q_i^{t+1} = (1 - \alpha^{t+1})q_i^t + \alpha^{t+1}k_i(q_{-i}^t, b_i^t). \quad (4)$$

where $\alpha^t \rightarrow 0$ as $t \rightarrow \infty$. We will adaptively choose the temperature parameter b_i^t using the gradient ascent of (1) with respect to b_i which results in the following update rule:

$$b_i^t = b_i^{t-1} + \frac{1}{(b_i^{t-1})^2} \gamma (r_i - S(q_i^{t-1})). \quad (5)$$

where $0 < \gamma < 1$ is a stepsize parameter.

3 Convergence of Probability Collectives

The convergence of a distributed optimization algorithm to a local or a global optimum is equivalent to the convergence to Nash equilibrium of a learning algorithm in games. It is important to know that there will be convergence at least to a local maximum, which is stable, rather than producing arbitrary solutions.

Theorem 1. *Our Probability Collectives algorithm, using the Brouwer updating rule (4) and gradient ascent to update the temperature parameter (5) will converge to the set of local maxima of G .*

Proof. Leslie and Collins (2006) proved that stochastic fictitious play with vanishing smoothing results in a generalised weakened fictitious play process and thus converges to the set of Nash equilibria in partnership games. The updates of stochastic fictitious play are of the following form:

$$q_i^{t+1} = (1 - a^{t+1})q_i^t + a^{t+1}(\overline{BR}_i^t(q_{-i}^t) + M_i^{t+1}) \quad (6)$$

where \overline{BR}_i^t is a smooth best response function and the M_i^{t+1} are martingale differences. A common choice of \overline{BR}_i^t is the Boltzman function [11]. The probability collectives updating rule as it is described in (4) results in updates that can also be expressed as (6) if the temperature parameters $b_i^t \rightarrow \infty$. Hence if we can show that $b_i^t \rightarrow \infty$ for all i , the q_i^t 's follow a generalised weakened fictitious play process, and therefore converge to a Nash equilibrium of the partnership game which corresponds to the optimum of G .

From this point we will consider only the updates of a single agent i and for simplicity of notation we will write $b_i^t = b_t$. Then we can rewrite (5) as $b_t = b_{t-1} + \frac{f(t)}{b_{t-1}^2}$, where $f(t) \geq \gamma e = \delta$. Depending on the choice of γ the value of δ can be arbitrarily small. We are going to use a γ such that $0 < \delta < \frac{b_0^3}{2}$

Initially we will show that b_t is greater than the series defined by $c_0 = b_0$ and $c_t = c_{t-1} + \frac{\delta}{c_{t-1}^2}$. So if c_t diverges b_t will also diverge. Afterwards we will prove that c_t diverges to finish our proof.

We are going to prove that $b_t \geq c_t \forall t$ using induction. We know that $b_0 \geq c_0$ (since they are equal) and we will assume that $b_{t-1} \geq c_{t-1}$. Then we will show that $b_t > c_t$.

$$\begin{aligned} b_t - c_t &= b_{t-1} - c_{t-1} + \frac{f_t}{b_{t-1}^2} - \frac{\delta}{c_{t-1}^2} \\ &\geq b_{t-1} - c_{t-1} + \delta \left(\frac{1}{b_{t-1}^2} - \frac{1}{c_{t-1}^2} \right) \end{aligned} \quad (7)$$

The derivative of (7) with respect to b_{t-1} is:

$$\frac{\partial b_t - c_t}{\partial b_{t-1}} = 1 - \frac{2e}{b_{t-1}^3} \quad (8)$$

which is always positive, by our choice of γ . Thus $b_t - c_t$ is minimized at 0 when $b_{t-1} = c_{t-1}$ and $f_t = \delta$. Thus by induction $b_t \geq c_t \forall t$.

Since we have proved that $b_t \geq c_t \forall t$, it will suffice to prove that c_t diverges. Note that $c_t = c_0 + \frac{\delta}{c_{t-1}^2} + \dots + \frac{\delta}{c_0^2} = c_0 + \sum_{i=1}^t \frac{\delta}{c_{i-1}^2}$. Suppose $c_t \not\rightarrow \infty$, so $\exists C > 0$ such that $c_t \leq C \forall t$. Hence $\frac{1}{c_t^2} \geq \frac{1}{C^2} \forall t$, and $C \geq c_t = c_0 + \sum_{i=1}^t \frac{\delta}{c_{i-1}^2} > c_0 + \frac{t\delta}{C^2} \rightarrow \infty$. This is a contradiction, so we must have $c_t \rightarrow \infty$

□

References

1. Kitano, H., Tadokoro, S., Noda, I., Matsubara, H., Takahashi, T., Shinjou, A., Shimada, S.: Robocup rescue: Search and rescue in large-scale disasters as a domain for autonomous agents research. In: Proc. of IEEE Conf. on System, Man and Cybernetics. (1999) (5 pages)
2. Kho, J., Rogers, A., Jennings, N.R.: Decentralized control of adaptive sampling in wireless sensor networks. *ACM Trans. Sen. Netw.* **5**(3) (2009) 1–35
3. van Leeuwen, P., Hesselink, H., Rohling, J.: Scheduling aircraft using constraint satisfaction. *Electr. Notes Theor. Comput. Sci.* **76** (2002)
4. Stranjak, A., Dutta, P.S., Ebden, M., Rogers, A., Vytelingum, P.: A multi-agent simulation system for prediction and scheduling of aero engine overhaul. In: *AA-MAS '08: Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems.* (2008) 81–88
5. Wolpert, D.H., Tumer, K.: An introduction to collective intelligence. Technical report, NASA (1999)
6. Arslan, G., Marden, J.R., Shamma, J.S.: Autonomous vehicle-target assignment: A game-theoretical formulation. *Journal of Dynamic Systems, Measurement, and Control* **129**(5) (2007) 584–596
7. Crites, R.H., Barto, A.: Improving elevator performance using reinforcement learning. In: *Advances in Neural Information Processing Systems 8.* (1996)
8. Littman, M.: Markov games as a framework for multiagent reinforcement learning. In: *Proceedings of the Eleventh International Conference of Machine learning.* (1994)
9. Uther, W., Veloso, M.: Adversarial reinforcement learning. Technical report, Carnegie Mellon University (1997)
10. Bowling, M., Veloso, M.: Multiagent learning using a variable learning rate. *Artificial Intelligence* **136** (2002) 215–250
11. Fudenberg, D., Levine, D.: *The theory of Learning in Games.* The MIT Press (1998)
12. Monderer, D., Shapley, L.: Potential games. *Games and Economic Behavior* **14** (1996) 124–143
13. Wolpert, D.H., Strauss, C.E.M., Rajnarayan, D.: Advances in distributed optimization using probability collectives. *Advances in Complex Systems (ACS)* **9**(04) (2006) 383–436
14. Leslie, D.S., Collins, E.: Generalised weakened fictitious play. *Games and Economic Behavior* **56**(2) (2006) 285–298