Abstract—The ever increasing use of intelligent multi-agent systems poses increasing demands upon them. One of these is the ability to reason consistently under uncertainty. This, in turn, is the dominant characteristic of probabilistic learning in graphical models which, however, lack a natural decentralised formulation. The ideal would, therefore, be a unifying framework which is able to combine the strengths of both multi-agent and probabilistic inference.

In this paper we present a unified interpretation of the inference mechanisms in games and graphical models. In particular, we view fictitious play as a method of optimising the Kullback-Leibler distance between current mixed strategies and optimal mixed strategies at Nash equilibrium. In reverse, probabilistic inference in the variational mean-field framework can be viewed as fictitious game play to learn the best strategies which explain a probabilistic graphical model.

I. INTRODUCTION

Multi-agent systems which have become an attractive approach to solving complex and distributed activities when doing so by one agent is too expensive or impossible, due to physical, geographical or temporal constraints. Applications in such problems occur range from control of vehicles in wide geographical terrains, global communication networks, economics and parallel optimisation.

Each agent has its own expert knowledge and observation domain. Its action responses are based on fusing observations from its local environment and actions from some of its neighbouring agents. Although there is no central control, the aim is to achieve emergent global behaviour that solves a common goal. In real world applications, action and measurement information isn’t exact and to handle such problems, an agent needs to be able to fuse uncertain information consistently. The two ways of tackling this task is by either adapting probabilistic methods to multi-agent domains [1] or the reverse, incorporating probabilistic reasoning into current multi-agent systems [2]. The methods of the former approach, unfortunately, have lead to systems that are rigid with regards to the agent’s environment, i.e. the topology of all agent interactions. The topology needs to be known in advance and has is singly connected. In the alternative approach, one can give not attention to the agent’s environment and perform probabilistic inference regardless until a solution is reached, but its far from clear what, if at all, solutions are reached.

In general, questions one might want to ask are: How good is a solution? Will the system always converge or can it exhibit limit cycle behaviour, and under what conditions can this happen? How many equilibria are there and what is the size of the solution clusters? These are all questions about the global behaviour of locally acting agents that is hard to understand and subject to intensive research. An additional ad hoc embedding of probabilistic inference makes these questions even more difficult to answer.

The most fruitful approach to this problem, in our opinion, is that of finding a unifying cost-function view of the multi-agent learning under uncertainty. The more well known the cost function, the more can be said about the global behaviour of the system. Furthermore, extensions made in other research areas can be much more easily incorporated into multi-agent systems.

In this paper we interpret learning in games as that of finding the strategy closest to the optimal Nash strategy. Closeness is, subjectively, measured by a Kullback-Leibler distance. It is the very same distance used for probabilistic inference is performed to learn optimal models of real-world observations. In particular, a certain simplification of the Kullback-Leibler distance, known as the variational mean-field framework, is the equivalent to independent players competing to find the best explanation of the world.

In section II of this paper we review fictitious play and the cost function view of it. Section III demonstrates the use of an identical cost function in probabilistic learning and highlights the equivalences between learning games and probabilistic models. Section IV points to extensions that can be made based on the similarities observed in the previous section.

II. (LEARNING IN) GAMES: FICTITIOUS PLAY

We consider non-cooperative strategic-from games with \(I\) players, index by \(i = 1, \ldots, I\). We split the set of players into two, \(I = \{i, i\}\), costing of player \(i\) and all other players \(i = \{1, 2, \ldots, i - 1, i + 1, \ldots, I\}\). Each player has a finite set of pure strategies (or actions) \(S_i\) typically assumed to be discrete\(^1\). The space of all possible combinations of actions is given by the strategy profile \(S = \times_{i=1}^I S_i = S_1 \times S_2 \times \cdots \times S_I\). Each player’s mixed strategy is a distribution over her set of pure strategies \(Q(S_i) \in \Delta(S_i) \forall i = 1, \ldots, I\) and is an element of the space of all distributions, \(\Delta(S_i)\).

\(^1\)For reasons of notational clarity any integration over the strategy space will be denoted by an integral, irrespective of whether the space is discrete or continuous.
the pure strategy profile, the mixed strategy profile is given by $Q(S) = \times_{i=1}^I Q(S_i) = Q(S_1) \times Q(S_2) \times \cdots \times Q(S_I)$. We will use $S_i$ to mean the pure strategy profile and $Q(S_i)$ the mixed strategy profile of all other players. The payoff by each player is a function mapping all pure strategy combinations of all players on the real line, i.e. $\ell(S) : A \to \mathbb{R}$ for a given the pure strategy profile $S = \{S_1 \times S_i\}$.

We define the utility function then as the expected reward obtained by player $i$, given the mixed strategies $Q(S_i)$ off all of $i$'s opponents

$$R(Q(S_i), Q(S_i)) \triangleq E(\ell(S)) = \int Q(S)\ell(S) \, dS.$$ \hspace{1cm} (1)

The solution to every player’s utility maximisation goal is the known as the equilibrium. It specifies the joint mixed strategy composed of independent mixed strategies every player adopts for a particular game. In particular, the Nash equilibrium is given as the optimal mixed strategy $Q^*(S_i)$ of player $i$

$$R(Q^*(S_i), Q(S_i)) \geq R(Q(S_i), Q(S_i)) \quad \forall i = 1, \ldots, I \hspace{1cm} (2)$$
i.e. each player has no incentive to deviate from the equilibrium strategy assuming none of the other players do.

A model for learning the optimal mixed strategy attained at Nash equilibrium is fictitious play. During play, every player chooses in turn the best response to the opponents mixed strategy, which is being continually updated following their past action. Thus, starting with some prior beliefs about strategies, at discrete time-step $t$ the strategies are updated according to

$$Q(S_i) \in \left(1 - \frac{1}{t+1}\right)Q(S_i) + \frac{1}{t+1}\beta_i \hspace{1cm} (3)$$

where the best response to $\beta_i$ the other players mixed strategies is defined as the mapping

$$\beta_i : Q(S_i) \to Q(S_i) \hspace{1cm} (4)$$

and where

$$\beta_i(Q(S_i)) = \arg \max_{Q(S_i) \in \Delta(S_i)} R_i(Q(S_i), Q(S_i)). \hspace{1cm} (5)$$

At Nash equilibrium, the map reaches steady state

$$Q^*(S_i) = \beta_i(Q^*(S_i)) \quad \forall i = 1, \ldots, I \hspace{1cm} (6)$$

and every player adopts the mixed strategy that maximises the expected utility.

One objection to fictitious play has been that players almost never play mixed strategies. Instead they myopically chose a pure strategy which maximises the immediate payoff. As a result, they may constantly switch between the pure strategies with ever increasing cycle durations [3]. To overcome such problems, fictitious play has been generalised to stochastic fictitious play, in which players begin play sub-optimally and increasingly play myopic best responses as time, $t$, passes [3].

In such cases, the utility function is augmented to

$$U(Q(S_i), Q(S_i)) = \int Q(S)\ell(S) \, dS + \tau H(S_i). \hspace{1cm} (7)$$

so that the smooth best response becomes

$$\beta_i(Q(S_i)) = \arg \max_{Q(S_i) \in \Delta(S_i)} \{R_i(Q(S_i), Q(S_i)) + \tau H(S_i)\}. \hspace{1cm} (8)$$

The temperature parameter controls the degree of sub-optimality played by the players and can be regarded as determining the amount of perturbation of the expected reward $R_i(Q(S_i), Q(S_i))$. The perturbation to player $i$'s payoffs, $H(S_i)$ is required to be a smooth strictly differentiable concave function the slope of which approaches infinity as $Q(S_i)$ reaches the boundary of $\Delta(S_i)$. One of the functions which satisfies the conditions for the second term in (7) is the entropy function. Under these assumptions, the best response is a mixed strategy

$$Q(S_i) \sim \exp \left(\frac{1}{\tau}R_i(S_i, Q(S_i))\right) = \exp \left(\frac{1}{\tau} \int Q(S)\ell(S) \, dS \right) \hspace{1cm} (9)$$

Viewed from the physical perspective it can be seen that the best mixed strategy response (9) is the maximally entropic distribution subject to the constraint of maximising the value of the game [4] and subject to an annealing schedule. For, without loss of generality, we can rewrite (7) as the maximum entropy cost function to be maximised

$$U(Q(S_i), Q(S_i)) = \lambda_i \left( \int Q(S)\ell(S) \, dS - \mathcal{L}_i \right) + \tau H(S_i) \hspace{1cm} (10)$$

where $\lambda_i$ is the Lagrangian parameter controlling the constraint that the expected reward of the game, the value of the game $\mathcal{L}_i$ for player $i$, is maximal.

The maximum entropy principle, used for the estimation of probability distributions under certain types of constraints, such as ensemble averages of a quantity, can be justified as the “only consistent variational principle for the assignment of probability distributions” [5]. In its simplest form it is a modern version of Laplace’s “Principle of Insufficient Reason (or Indifference Principle)”, which states that, in the absence of reason/knowledge to prefer one over alternatives hypothesis, simply make all probabilities equal. In effect, the principle of searching for the Nash equilibrium has changed from sampling the actions until they are drawn from the equilibrium distribution to updating the mixed strategy until it converges to the intended (optimal) mixed strategy, i.e. to that of finding the distribution closest to the Nash distribution[6]. Put in terms of closeness, a myriad of possible distance measures become applicable, among the probabilistic divergence measures, such as Kullback-Leibler Divergences$^2$.

From the analogy of fictitious play and maximum entropy, we can draw a simple (and for this case rather uninteresting) graphical model of the game, shown in figure 1. The actions

$^2$Note that the entropy function $\int Q(S)\log Q(S) \, dS$ is relative to an appropriately defined measure, $\mu(S)$, $\int Q(S)\log Q(S) - \log(\mu(S)) \, dS$, which typically is assumed to be flat.
are random variables with unknown distributions and referred to as latent random variables (shown as clear nodes in 1). The rewards, akin to measurements are given, i.e. observed, and considered instantiated random variables (shaded nodes).

The maximum entropy criterion is well known and used in machine learning, for example in the estimation of probabilistic models, known as graphical models. Thus, from the machine learning point of view, fictitious play resembles the well known Expectation Maximisation (EM) algorithm [7], which has a cost function that is, in principle, identical to (10).

Consider again \( I \) players, index by \( \forall i = 1, \cdots, I \), and split into two sets, \( I = \{ i, i' \} \). The pure strategy space of each player is denoted by \( S_i \), that of the opponent players by \( S_{i'} \), and the strategy profile by \( S \). The mixed strategy profile is given by \( Q(S) \), where make the specific assumption that all players, and subsequently their mixed strategies, are independent

\[
Q(S) = \prod_{i=1}^{I} Q(S_i). \tag{11}
\]

This is known as the “mean field” assumption in the variational learning framework.

The payoffs are now generalised from discrete matrix payoffs to payoff functions, \( \ell(S, \theta) \) with parameters \( \theta \). Typically in machine learning, the payoffs form take of a model postulated to underly the experimental data generating processes (a.k.a. generative model). Furthermore, omnipresent observation noise is captured by (more often than not) additive random perturbations which follow certain probability distributions. The logarithm of the distributions is used for mathematical convenience (the exponential family of distributions plays a very dominant role in machine learning). Thus, the payoffs for modelling observations \( D \) are written as

\[
\ell(S, D|\theta) \triangleq \log(p(S, D|\theta)) \tag{12}
\]

The analogy between fictitious play and the variational methods can then be established through the cost function. In particular, the variational learning approach finds the maximum entropy equilibrium distribution, \( Q(S) \) subject to maximising the expected reward (log-probability of the model)

\[
D(Q(S)||P(S|\theta)) = \int Q(S) \ell(S, D|\theta) \ dS + \tau H(S)
= \int Q(S) \log \left( \frac{\tau Q(S)}{p(S, D|\theta)} \right) \ dS \tag{13}
\]

The second equality highlights the fact that the variational cost functions (13) is a Kullback-Leibler divergence, as would by the fictitious play cost function (7) with exponentiated payoffs \( R \).

The optimal distribution \( Q(S_i) \) of player \( i \) is the one which minimises the Kullback-Leibler divergence (13), assuming all other players \( i \) adopt the mixture strategy \( Q(S_{i'}) \). It is obtained by partial differentiation of (13). In fact, the definition of Nash equilibrium (2) is a partial differential in disguise. Thus, the variational best response is a mixed strategy distribution which takes the general form [8]

\[
Q(S_i) \propto \exp \left\{ \frac{1}{\tau} \int Q(S_{i'}) \ell(S, D|\theta) \ dS \right\} \tag{14}
\]

This is identical to smooth fictitious play best response (16), which as shown to converge to equilibrium for fictitious play [9] and for the E.M. algorithm [10].

Due to payoffs begin functions rather than simple lookup tables, as in the case of matrix games, the action maximising the reward needs to be computed first. For one particular variational learning method, the Expectation Maximisation (EM), this step is performed by finding the parameter \( \theta \) at which the payoff function reaches its maximum. Variational “play” then proceeds in two steps, commonly referred to as the E- and M-step:

- The E-step computes player \( i \)‘s best response, \( Q(S_i) \), based on the opponents’ mixed strategies \( Q(S_{i'}) \).
- The M-step finds the total expected reward maximising action \( \theta \).

There are some important differences machine learning and learning in games is the reward.

a) Reward: The major difference is the reward. Rewards are very frequently concave (log) functions. This represents the fact that machine learning tasks deal with modelling real observations. Our lack of complete information about nature, however, makes minimum risk strategy the optimal choice and hence a concave utility [11]. Thus, from a game theoretic view, machine learning is a multi-player game against nature. The players are the probability distributions imposed on parameters of the experimental data model. They may be independent players, characterised by independent distributions. Thus, they compete against each other to find the best sparse mixed strategy against nature.

By comparison with economic applications of game theory [12], the currency of machine learning games is the “information”, in the sense of log-probability. Just like normal
currency, it has no additional attributes to it. Conversion between different monetary currencies is via exchange rates, whilst the information currency is converted through use of conditional probabilities.

b) Temperature: Another difference is the fact, that the greater part of machine learning algorithms keep the temperature parameter constant, bar for example simulated annealing [13], annealed Markov Chain Monte Carlo [14] and deterministic annealing [15].

c) Perturbation function: Most crucially and unlike stochastic fictitious play in (7), the parameters of the perturbation processes are not fixed. They are unknown and, in the Bayesian framework, have associated prior and posterior distributions. These distributions can then be considered as additional players in the game. The value of the game in the Bayesian framework is usually known as the evidence, the marginal probability of the observations. In the EM learning the value becomes the maximum likelihood of the data.

Rewards, in the EM method, were exactly maximised. That is in the sense that a point estimate for the parameters of the payoff function $\theta$ was obtained. However, uncertainty about the rewards may also exists. To deal with this uncertainty in the Bayesian framework, priors are imposed on the payoff function parameters. This leads to an extension of the variational learning scheme, in which the Maximum entropy is not taken but with respect to a particular hypothesis - that of the prior.

The adjusted cost function (15) takes the form of

$$D(Q(S,\theta)||P(S,\theta)) = \int\int Q(S)Q(\theta)\log\left(\frac{\tau Q(S)Q(\theta)}{P(S,\theta)p(\theta)}\right) dS d\theta$$

This is the Kullback-Leibler divergence between the exact posterior distribution, $P(S,\theta,D)$, and an approximation to it, $Q(S,\theta)$. This KL divergence is typically used when integration to obtain of exact marginal distributions, such as $P(S)$ or $P(\theta)$, is intractable [16]. The KL divergence between the approximate and true posterior is minimised leading to coupled update equations for the distributions. In the mean field assumptions they take the general form

$$Q(S_{i}) \propto \exp\left\{\frac{1}{\tau} \int\int Q(S)Q(\theta)\log(\ell(S,\theta,D)) dS d\theta\right\}$$

These are then iterated until convergence. In essence, all variables are now random variables with associated distributions. Thus, the set of players in now augmented to include players whose strategy space are the payoff function parameters.

IV. IMPLICATIONS AND EXTENSIONS TO GAME PLAYING

A. Global vs Local Rewards

So far, no specific assumption has been made about the form of the payoffs. In their simplest form, the payoffs may be private to each player

$$\ell(Q) = \sum_{i} \ell(S_{i})$$

One the other hand, we may presume that some that some player’s payoffs affect only few other players. For example, consider a 3 player extension of the 2-player matching pennies game. In the extension the third player’s rewards depend on how the other 2 players play against each other, i.e. they do not consider player 3 in their play. The game designer thus imposes a a structural form on the payoff relationships, for example as

$$\ell(Q) = \ell(S_{1}|S_{2}) + \ell(S_{2}|S_{1}) + \ell(S_{3}|S_{1},S_{2}).$$

Viewed from a probabilistic derivation “payoffs” (13), the structural relationships between their individual components may stem from apriori knowledge of causal relationships between parameters of the experimental data model. Thus, they specifically state knowledge relating local to global reward functions. In general, for computing best response, each player must consider players affecting her, players she influences and other players affecting the players she influences (in graph theory terms the Markov blanket, that is parents, children and children’s parents [17].

B. Independent Mixed Strategies

In addition to assumptions affecting the payoffs, these is the assumption about the independence of players (11). Both, the machine learning as well as game theoretic learning exposition above assumed independent players. This assumption naturally arose from the problem statement, such as distributed control. However, the machine learning literature has developed methods for finding marginal equilibrium distributions when there exist some probabilistic dependence between the players’ mixed strategies. For certain dependence topologies, e.g. trees, exact methods can be derived [18]. Structure of mixed strategy distributions encode a more direct dependence between the players, one mediated through probabilistic inference rather than payoffs. It is conceivable that such dependence structures find their equivalences in coalition formation in games.

C. Quality of Equilibrium

Quantifying the tightness of the bound (15), and hence the quality of the equilibrium , is subject to current research. Generally speaking it will depend on degree of discrepancy between the true and approximate distribution. One approach uses sampling methods have been used to improve on the equilibrium distribution[19]. Alternatively, the empirical integration in (16) can be performed and has been applied to large scale optimisation problems[20].

Acknowledgements: This research was undertaken as part of the ARGUS II DARP (Defence and Aerospace Research Partnership). This is a collaborative project involving BAE SYSTEMS, QinetiQ, Rolls-Royce, Oxford University and Southampton University, and is funded by the industrial partners together with the EPSRC, MoD and DTI.
REFERENCES


