Journal of Artificial Intelligence Research 33 (2008) Submitted 12/07; published 10/08

# On Similarities between Inference in Game Theory and Machine Learning

**Iead Rezek** I.REZEK@IMPERIAL.AC.UK
*Department of Clinical Neurosciences, Imperial College*
*London, SW7 2AZ, UK*

**David S. Leslie** DAVID.LESLIE@BRISTOL.AC.UK
*Department of Mathematics, University of Bristol*
*Bristol, BS8 1TW, UK*

**Steven Reece** REECE@ROBOTS.OX.AC.UK
**Stephen J. Roberts** SJROB@ROBOTS.OX.AC.UK
*Department of Engineering Science, University of Oxford*
*Oxford, OX1 3PJ, UK*

**Alex Rogers** ACR@ECS.SOTON.AC.UK
**Rajdeep K. Dash** RKD@ECS.SOTON.AC.UK
**Nicholas R. Jennings** NRJ@ECS.SOTON.AC.UK
*School of Electronics and Computer Science, University of Southampton*
*Southampton, SO17 1BJ, UK*

## Abstract

In this paper, we elucidate the equivalence between inference in game theory and machine learning. Our aim in so doing is to establish an equivalent vocabulary between the two domains so as to facilitate developments at the intersection of both fields, and as proof of the usefulness of this approach, we use recent developments in each field to make useful improvements to the other. More specifically, we consider the analogies between smooth best responses in fictitious play and Bayesian inference methods. Initially, we use these insights to develop and demonstrate an improved algorithm for learning in games based on probabilistic moderation. That is, by integrating over the distribution of opponent strategies (a Bayesian approach within machine learning) rather than taking a simple empirical average (the approach used in standard fictitious play) we derive a novel *moderated fictitious play* algorithm and show that it is more likely than standard fictitious play to converge to a payoff-dominant but risk-dominated Nash equilibrium in a simple coordination game. Furthermore we consider the converse case, and show how insights from game theory can be used to derive two improved mean field variational learning algorithms. We first show that the standard update rule of mean field variational learning is analogous to a Cournot adjustment within game theory. By analogy with fictitious play, we then suggest an improved update rule, and show that this results in *fictitious variational play*, an improved mean field variational learning algorithm that exhibits better convergence in highly or strongly connected graphical models. Second, we use a recent advance in fictitious play, namely dynamic fictitious play, to derive a *derivative action variational learning* algorithm, that exhibits superior convergence properties on a canonical machine learning problem (clustering a mixture distribution).

## 1. Introduction

There has recently been increasing interest in research at the intersection of game theory and machine learning (Shoham, Powers, & Grenager, 2007; Greenwald & Littman, 2007). Such work is motivated by the observation that whilst these two fields have traditionally been viewed as disparate research areas, there is actually a great deal of commonality between them that can be exploited within both fields. For example, insights from the machine learning literature on graphical models have led to the development of efficient algorithms for calculating Nash equilibria in large multi-player games (Kearns, Littman, & Singh, 2001). Similarly, the development of boosting algorithms within machine learning has been facilitated by regarding them as being engaged in a zero-sum game against a base learner (Freund & Schapire, 1997; Demiriz, Bennett, & Shawe-Taylor, 2002).

While such interdisciplinary inspiration is promising, unless there is a clearer understanding of the principal connections that exist between the two disciplines, these examples will remain isolated pieces of research. Thus, the general goal of our work is to explore in a more formal way the commonalities between game theory and machine learning. In order to do so, we consider an important problem that is central to both fields; that of making inferences based on previous observations.

We first consider game theory, where this problem occurs in the context of inferring the correct strategy to play against an opponent within a repeated game. This is generally termed *learning in games* and a common approach is to use an algorithm based on fictitious play (see Fudenberg & Levine, 1999). Here, we show that an insight from Bayesian inference (a standard machine learning technique) allow us to derive an improved fictitious play algorithm. More specifically, we show that by integrating over the distribution of opponent strategies (a standard approach within machine learning), rather than taking a simple empirical average (the approach used within the standard fictitious play algorithm), we can derive a novel *moderated fictitious play* algorithm. Moreover, we then go on to demonstrate that this algorithm is more likely than standard fictitious play to converge to a payoff-dominant but risk-dominated Nash equilibrium in a simple coordination game[1].

In the second part of the paper, we consider the mean field variational learning algorithm, which is a popular means of making inferences within machine learning. Here we show that analogies with game theory allows us to suggest two improved variational learning algorithms. We first show that the standard update rule of mean field variational learning is analogous to a Cournot adjustment process within game theory. By analogy with fictitious play, we suggest an improved update rule, which leads to an improved mean field variational learning algorithm, which we term *fictitious variational play*. By appealing to game-theoretic arguments, we prove the convergence of this procedure (in contrast standard mean-field updates can suffer from "thrashing" behaviour (Wolpert, Strauss, & Rajnarayan, 2006) similar to a Cournot process), and we show that this algorithm exhibits better convergence in highly or strongly connected graphical models. Second, we show that a recent advance in fictitious play, namely dynamic fictitious play (Shamma & Arslan, 2005), can be used to derive the novel *derivative action variational learning* algorithm. We demon-

1. We note here that a form of Bayesian learning in games is known to converge to equilibrium (Kalai & Lehrer, 1993). However, in that work the players perform Bayesian calculations in the space of all repeated game strategies, resulting in extremely complex inference problems. In contrast, we consider a Bayesian extension of fictitious play, using insights from machine learning to aid myopic decision-making.

strate the properties of this algorithm on a canonical machine learning problem (clustering a mixture distribution), and show that it again exhibits superior convergence properties compared to the standard algorithm.

When taken together, our results suggest that there is much to be gained from a closer examination of the intersection of game theory and machine learning. To this end, in this paper, we present a range of insights that allow us to derive improved learning algorithms in both fields. As such, we suggest that these initial first steps herald the possibility of more significant gains if this area is exploited in the future.

The remainder of the paper is organised as follows. In section 2 we discuss work related to the interplay of learning in games and machine learning. Then, in section 3 we discuss how techniques within machine learning can be used in relation to learning in games. We review the standard stochastic fictitious play algorithm, and then go on to derive and evaluate our *moderated fictitious play* algorithm. We then change focus, and in section 4, we show how techniques within game theory apply to machine learning algorithms. Again, we initially review the standard mean field variational learning algorithm, and then, by analogy with fictitious play and the Cournot adjustment, present in section 4.2 our *fictitious variational play* algorithm. In section 4.3 we continue this theme and incorporate insights from dynamic fictitious play to derive and evaluate our *derivative action variational learning* algorithm. Finally, we conclude and discuss future directions in section 5.

## 2. Related Work

The topics of inference and game theory have traditionally been viewed as separate research areas, and consequently little previous research has exploited their common features to achieve profitable cross-fertilisation.

One area where progress has been made is in the use of concepts from game theory to find the optimum of a multi-dimensional function. In this context, Lambert, Epelman, and Smith (2005) used fictitious play as an optimisation heuristic where players each represent a single variable and act independently to optimise a global cost function. The analysis restricts attention to the class of objective functions that are products of these independent variables, and is thus rather limited in practice.

In a similar vein, Wolpert and co-authors consider independent players who, through their actions, are attempting to maximise a global cost function (Lee & Wolpert, 2004; Wolpert, 2004). In this body of work, however, the optimisation is to be carried out with respect to the joint distributions of the variables chosen by all players. A mean-field approach is taken, resulting in independent choices for each player; the approach is very similar in flavour to that presented in Section 4 of this article. However, in this paper we explicitly use advances in the theory of learning in games to develop improved optimisation algorithms.

In the context of improving game-theoretical algorithms using techniques from machine learning and statistics, Fudenberg and Levine (1999) show that fictitious play has an interpretation as a Bayesian learning procedure. However this interpretation shows fictitious play to be a type of plug-in classifier (Ripley, 2000), and they stop short of using the full power of Bayesian techniques to improve the method. In contrast, in this article we take a fully Bayesian approach to deciding the optimal action at each play of the game.

Other articles where cross-over has been attempted, but that do not overlap greatly with the current article, include that of Demiriz et al. (2002) and Freund and Schapire (1997), who have interpreted boosting algorithms as zero sum games, and Kearns et al. (2001) who consider the use of techniques from graphical models (Jordan, Ghahramani, Jaakkola, & Saul, 1997) to help calculate equilibria in graphical games.

## 3. Fictitious Play

Fictitious play is an important model for learning in games and the source of the algorithmic developments presented later in this work. We begin by presenting the notation and terminology used in the standard game-theoretic representation of fictitious play. The reader is referred to the work of Fudenberg and Levine (1999) for a more extensive discussion.

We will consider strategic-form games with $I$ players that are indexed by $i \in \{1, \ldots, I\}$, and where we use $-i$ to index all players other than player $i$. By $\mathbb{S}^i$ we denote the finite set of pure strategies $S^i$ (also known as actions) available to player $i$, by $\mathbb{S}$ the set $\mathbb{S}^1 \times \mathbb{S}^2 \times \cdots \times \mathbb{S}^I$ of pure strategy profiles of all players, and by $\mathbb{S}^{-i}$ the set of pure strategy profiles of all players other than $i$. Each player's pay-off function is denoted by $R^i : \mathbb{S} \to \mathbb{R}$ and maps pure strategy profiles to the real line, i.e. each set of actions selected by the players is associated with a real number.

This simple model is usually extended to allow players to use mixed strategies $\pi^i \in \Delta(\mathbb{S}^i)$, where $\Delta(\mathbb{S}^i)$ denotes the set of probability distributions over the pure strategy set $\mathbb{S}^i$. Hence each $\pi^i$ is a probability distribution on the discrete space $\mathbb{S}^i$. Writing $\pi = (\pi^1, \pi^2, \ldots, \pi^I)$ for the probability distribution on $\mathbb{S}$ which is the product of the individual mixed strategies, $\pi^i$, we extend the reward functions to the space of mixed strategies by setting

$$R^i(\pi) = \mathbb{E}_\pi R^i(S) \tag{1}$$

where $\mathbb{E}_\pi$ denotes expectation with respect to the pure strategy profile $S \in \mathbb{S}$ selected according to the distribution $\pi$. Similarly

$$R^i(S^i, \pi^{-i}) = \mathbb{E}_{\pi^{-i}} R^i(S^i, S^{-i}) \tag{2}$$

where $\mathbb{E}_{\pi^{-i}}$ denotes expectation with respect to $S^{-i} \in \mathbb{S}^{-i}$.

The standard solution concept in game theory is the Nash equilibrium. This is a mixed strategy profile $\pi$ such that for each $i$

$$R^i(\pi) \geq R^i(\tilde{\pi}^i, \pi^{-i}) \quad \text{for all } \tilde{\pi}^i \in \Delta(\mathbb{S}^i). \tag{3}$$

In other words, a Nash equilibrium is a set of mixed strategies such that no player can increase their expected reward by unilaterally changing their strategy.

If all players receive an identical reward then we have what is known as a partnership game. In this case, players are acting independently while trying to optimise a global objective function. This special case is important since it corresponds to a distributed optimisation problem, where the objective function represents the reward function of the game. At a Nash equilibrium it is impossible to improve the expected value of this objective function by changing the probability distribution of a single player. Thus Nash equilibria correspond to local optima of the objective function.

Fictitious play proceeds by assuming that during repeated play of a game, every player monitors the action of their opponent. The players continually update estimates $\sigma$ of their opponents' mixed strategies by taking the empirical average of past action choices of the other players. Given an estimate of play, a player selects a best response (i.e. an action that maximizes their expected reward given their beliefs). Thus, at time $t$, the estimates are updated according to

$$\sigma^i_{t+1} = \left(1 - \frac{1}{t+1}\right) \sigma^i_t + \frac{1}{t+1}\, b^i(\sigma^{-i}_t) \tag{4}$$

where $b^i(\sigma^{-i}_t)$, the best response to the other players' empirical mixed strategies, satisfies

$$b^i(\sigma^{-i}_t) \in \operatorname*{argmax}_{S^i \in \mathbb{S}^i} R^i(S^i, \sigma^{-i}_t). \tag{5}$$

In certain classes of games, including the partnership games mentioned previously, beliefs that evolve according to equation 4 are known to converge to Nash equilibrium. On the other hand, there also exist games for which non-convergence of equation 4 has been shown Fudenberg and Levine.

### 3.1 Stochastic Fictitious Play

Now, one objection to fictitious play has been the discontinuity of the best response function, which means that players almost always play pure strategies, even when beliefs have converged to a Nash equilibrium in mixed strategies. To overcome such problems, fictitious play has been generalized to stochastic fictitious play (see Fudenberg & Levine, 1999) which employs a smooth best response function, defined by

$$\beta^i(\pi^{-i}) = \operatorname*{argmax}_{\pi^i \in \Delta(\mathbb{S}^i)} R^i(\pi^i, \pi^{-i}) + \tau v^i(\pi^i) \tag{6}$$

where $\tau$ is a temperature parameter and $v^i$ is a smooth, strictly differentiable concave function such that as $\pi^i$ approaches the boundary of $\Delta(\mathbb{S}^i)$ the slope of $v^i$ becomes infinite. One popular choice of smoothing function $v^i$ is the entropy function, which results in the logistic choice function with the noise or temperature parameter $\tau$

$$\beta^i(\pi^{-i})(S^i) = \frac{1}{Z} \exp\left\{ \frac{1}{\tau} R^i(S^i, \pi^{-i}) \right\} \tag{7}$$

where the partition function $Z$ ensures that the best response adds to unity. Thus, in stochastic fictitious play, players choose at every round an action randomly selected using the smooth best response to their current estimate of the opponents' probability of play.

The estimates under this process are also known to converge in several classes of games, including partnership games (Hofbauer & Hopkins, 2005) which will be discussed again in section 4. Several further extensions of fictitious play have been introduced in attempts to extend the classes of games in which convergence can be achieved, including weakened fictitious play (Leslie & Collins, 2005; van der Genugten, 2000) and dynamic fictitious play (Shamma & Arslan, 2005). We will use these extensions in the second part of the paper to improve convergence of modifications of variational learning.

### 3.2 Moderated Fictitious Play

In fictitious play, each player "learns" by using an empirical average of the past action choices of the other players to estimate their current mixed strategy. This estimate can be thought of as the maximum likelihood estimate (MLE) of the opponent's mixed strategy at time $t$ under the assumption that all the actions of each player $i$ have been selected using a multinomial distribution with parameter $\pi^{-i}$ such that

$$\sigma_t^{-i} = \hat{\pi}_t^{-i} = \underset{\pi^{-i}}{\operatorname{argmax}} \prod_{u=0}^{t} P(S_u^{-i}; \pi^{-i}) \tag{8}$$

where $P(S_u^{-i}; \pi^{-i})$ can be modelled by a product of multinomial distributions $P(S_u^{-i}; \pi^{-i}) = \prod_{j=1, j \neq i}^{I} \pi^j(S_u^j)$. Fudenberg and Levine note that this choice of $\pi_t^{-i}$ corresponds to the maximum *a posteriori* estimate of the opponent mixed strategies in a Bayesian model.

However, from a machine learning perspective, the logistic choice best response function, given in equation 7, may be viewed as a single layer neural network with a sigmoid activation function (Bishop, 2006). Substituting the unknown parameter, $\pi_t^{-i}$, in equation 7 by its maximum likelihood estimate (or by a Bayesian point estimate) fails to take into account anything that is known about the parameter's distribution; classifiers using this approach have been called "plug-in" classifiers (Ripley, 2000). From a Bayesian perspective, better predictions can be obtained by integrating out the parameter's distribution and thus computing the *posterior predictive* best response function (Gelman, Carlin, Stern, & Rubin, 2000). This process is known in the neural network literature as "moderation" (MacKay, 1992).

This suggests a modification of fictitious play that we term *moderated fictitious play*. In this, every player uses the posterior predictive best response, obtained by integrating out over all opponent mixed strategies, weighted by their posterior distribution given the previously observed actions. As it is conventional to choose to use the posterior mean as the point estimate, the strategy now chosen by player $i$ is

$$\tilde{\beta}_t^i = \int \beta^i(\pi_t^{-i}) P(\pi_t^{-i} \mid S_{1:t}^{-i}) \, d\,\pi_t^{-i} \tag{9}$$

where $P(\pi_t^{-i} \mid S_{1:t}^{-i})$ is the posterior probability of the opponents' mixed strategies $\pi_t^{-i}$ given the observed history $S_{1:t}^{-i}$ of play from time 1 to $t$.

Since we model the observed pure strategies of player $j$ as observations of a multinomial random variable with parameters $\pi^j$, we place a uniform Dirichlet prior, $Dir(\pi^j; \alpha_0^j)$, on each $\pi^j$, with all parameters $\alpha_{0_k}^j = 1$. The posterior distribution of $\pi_t^{-i}$ is therefore again a product of independent Dirichlet distributions,

$$P(\pi_t^{-i} \mid s_{1:t}^{-i}; \alpha_t^{-i}) = \prod_{j \neq i} Dir(\pi_t^j; \alpha_t^j) \tag{10}$$

with $\alpha_t^j(s^j) = 1 + \sum_{u=1}^{t} \mathbf{I}\{s_u^j = s^j\}$, where $\mathbf{I}$ is an indicator function.

There are multiple approaches to estimating the integral in equation 9. A generally applicable approach is to sample $N$ opponent mixed strategies, $\Pi_n^{-i}$, from the posterior distribution and use a Monte Carlo approximation of the integral in equation 9, given by

$$\tilde{\beta}_t^i \approx \frac{1}{N} \sum_{n=1}^{N} \beta^i(\Pi_n^{-i}). \tag{11}$$

To investigate the effect of moderation we also consider an analytic expression for $\tilde{\beta}_t^i$ that makes use of two approximations. The first approximates the distribution in equation 10 by a normal distribution

$$\mathcal{N}\left(\mu; \Sigma\right) \tag{12}$$

with mean vector

$$\mu = \alpha_t^{-i} / \bar{\alpha}_t^{-i} \tag{13}$$

and covariance matrix

$$\Sigma = \frac{1}{\bar{\alpha}_t^{-i}} \begin{pmatrix} \mu_1(1-\mu_1) & \dots & -\mu_1\mu_K \\ & \ddots & \\ -\mu_K\mu_1 & \dots & \mu_K(1-\mu_K) \end{pmatrix} \tag{14}$$

where $K = |S^{-i}|$ and $\bar{\alpha}_t^{-i} = \sum_k \alpha_{t_k}$ (Bernardo & Smith, 1994). The second, given in MacKay (1992) for the case of two action choices, approximates the integral of a sigmoid

$$g\left(\frac{a}{\tau}\right) = \frac{1}{1 + \exp\left(\frac{a}{\tau}\right)}$$

with respect to a normal distribution, $P(a) = \mathcal{N}(a; m, \sigma^2)$ with mean $m$ and variance $\sigma^2$, by the modified sigmoid

$$\int g\left(\frac{a}{\tau}\right) P(a)\, d\, a \approx g\left(\frac{1}{\tau}\kappa(\sigma) m\right) \tag{15}$$

where

$$\kappa(\sigma) = \left(1 + \frac{\pi\sigma^2}{8}\right)^{-\frac{1}{2}}. \tag{16}$$

We see from equations 15 and 16 that the effect of moderation is to scale high rewards down in proportion to the variance (and thus uncertainty of estimated opponent mixed strategy) and shift the probability value of any action closer to 0.5 (i.e. down from unity or up from zero to 0.5). At the onset of play when little is known about the opponent, playing both actions with equal probability is an intuitively reasonable course of action.

To test the general moderated fictitious play of equation 9 with Dirichlet posterior distributions, we investigate games with varying degrees of risk dominance, since in these cases the equilibrium selection of strategies is strongly dependent upon the players' beliefs about the other players' probabilities for each action. We compared the probability of moderated and stochastic fictitious play converging to the payoff dominated solution for games in which the payoffs are described by the payoff matrix

$$R = \begin{pmatrix} (1,1) & (0,r) \\ (r,0) & (10,10) \end{pmatrix} \tag{17}$$

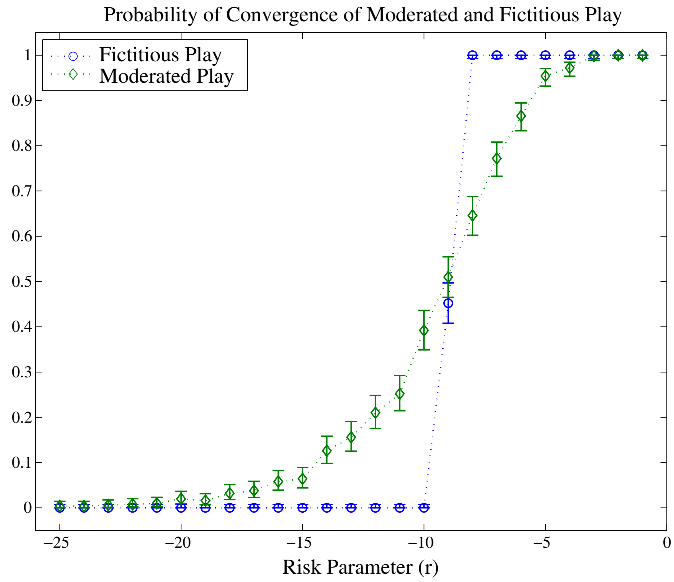



Figure 1: Probability of convergence (with 95% confidence intervals) of fictitious and moderated play to the payoff-dominant but risk-dominated equilibrium of the game represented by the payoff matrix shown in equation 17.

where the factor $r$ determined the degree of risk dominance of the action pair 1/1 and was set to range from $r = -1, -2 \cdots - 25$.

In this game, it is clearly best if both players choose action 2 (since in doing so they both receive a reward of 10). However for many learning processes (see Fudenberg & Levine, 1999 and Young, 1998, for example) the players see a few initial actions and become convinced that the opponent is playing a strategy that means choosing action 2 is bad (since the penalty for playing 2 against 1 is high). We find that by taking uncertainty over strategies into account at the start, through the use of our moderated fictitious play algorithm, we are less likely to get stuck playing action 1, and so convergence to the strategy at $(10, 10)$ is more likely.

For each value of $r$ we ran 500 plays of the game and measured the convergence rates for moderated and stochastic fictitious play using matching initial conditions. Both algorithms used the same smooth best response function. Specifically, Boltzmann smooth best responses with temperature parameter $\tau = 0.1$. We present the results in Figure 1. Stochastic fictitious play converges to the (1,1) equilibrium for all games in which action (2,2) is risk dominated (i.e. for $r < -10$). As soon as action (2/2) is no longer risk dominated then stochastic fictitious play does converge. In contrast, moderated play exhibits a much smoother overall convergence characteristic. Moderated play achieves convergence to action (2/2) over a much greater range of values of $r$, though with varying degrees of probability. Thus for most risk dominated games examined ($r = -25 \cdots - 10$), moderated play is more likely to converge to the payoff-dominant equilibrium than stochastic fictitious play.

Thus, by using an insight from machine learning, and specifically, the standard procedure in Bayesian inference of integrating over the distribution of the opponent's strategies, rather than taking an empirical average, we have been able to derive an algorithm based on stochastic fictitious play with a smoother and thus more predictable convergence behaviour.

## 4. Variational Learning

Having shown in the first part of the paper how insights from machine learning can be used to derive improved algorithms for fictitious play, we now consider the converse case. More specifically, we consider a popular machine learning algorithm, the mean field variational learning algorithm, and show how this can be viewed as learning in a game where all players receive identical rewards. We proceed to show how insights from game theory (specifically, relating variational leaning to a Cournot process and fictitious play) can be used to derive improved algorithms.

We start by reviewing the mean field variational learning algorithm, and first note that it and other methods are typically used to infer the probability distribution of some latent (or hidden) variables, based on the evidence provided by another set of observable variables. Computing these probability distributions, however, requires an integration step which is frequently intractable (MacKay, 2003). To tackle this problem one can use Markov chain Monte Carlo methods (Robert & Casella, 1999) to obtain asymptotically optimal results, or alternatively use approximate analytical methods, such as variational approaches, if faster algorithms are preferred. Among the variational methods, the mean field variational approach to distribution estimation (Jordan et al., 1997) has been applied to real world problems ranging from bioinformatics (Husmeier, Dybowski, & Roberts, 2004) to finite element analysis (Liu, Besterfield, & Belytschko, 1988), and now here to games.

### 4.1 Mean Field Variational Method

The typical starting point in variational learning is the distributional form of a model, postulated to underlie the experimental data generating processes (i.e. the generative model). The distribution will usually be instantiated with some observations, $\mathcal{D}$, and defined over a set $I$ of latent variables which are indexed by $i = 1, \cdots I$. We denote the domain of the latent variable $S^i$ by $\mathbb{S}^i$, and an element by $s^i \in \mathbb{S}^i$. Note that, for ease of exposition later in the text, we re-use and newly define $S$ as we intend to make the connection to the earlier definition of $S$ as the strategy profile. We often desire the marginal distribution $p^i \in \Delta(\mathbb{S}^i)$ of the latent variable $i$, taken from the set of all marginal distributions $\Delta(\mathbb{S}^i)$ over $S^i$.

In the absence of any detailed knowledge about dependence or independence of the variables, we define the joint distribution $p \in \Delta(\mathbb{S})$ for the set of all distributions over $\mathbb{S}$, where $\mathbb{S} = \mathbb{S}^1 \times \mathbb{S}^2 \times \cdots \mathbb{S}^I$ is the profile domain of the latent variables. For mathematical convenience we resort to the logarithm of the density

$$\ell(S \mid \mathcal{D}, \theta) \triangleq \log\left(p(S \mid \mathcal{D}, \theta)\right) \tag{18}$$

and parameterise the distribution $p$ with $\theta \in \Theta$.

Due to the intractability of integrating equation 18 with respect to $S \in \mathbb{S}$, variational learning methods (Jordan et al., 1997) approach the problem by finding the distribution,

$q \in \Delta(\mathbb{S})$, which minimises the criterion

$$\mathcal{F} = \int \cdots \int q(s)\ell(s \mid \mathcal{D}, \theta)\, d\, s + \tau H\left(q\right) \tag{19}$$

where $H(\cdot)$ is the entropy function

$$H\left(q\right) = -\int \cdots \int q(s) \log q(s)\, d\, s. \tag{20}$$

This is equivalent to minimising

$$D(q||p) = \int \cdots \int q(s) \log \left( \frac{q(s)^{\tau}}{p(s \mid \mathcal{D}, \theta)} \right)\, d\, s \tag{21}$$

and highlights the fact that the variational cost function described in equation 19 is a Kullback-Leibler (KL) divergence between the marginal log-likelihood $\log\left(p(\mathcal{D})\right)$ and the negative free energy (Jordan et al., 1997).

Within variational learning, the "mean field" approach makes the assumption that all latent variables are independent. Thus, the distribution profile, $q$, for the latent variables simplifies to

$$q(s) \triangleq \prod_{i=1}^{I} q^i(s^i) \tag{22}$$

where $q^i \in \Delta(\mathbb{S}^i)$. On the basis of the mean field assumption, the optimal distribution of variable $i$ is the one that minimises the KL divergence in equation 21, assuming that all other variables $-i$ adopt the distribution $q^{-i} \in \Delta(\mathbb{S}^{-i})$, and can be obtained by partial differentiation of equation 21. The model-free update equation for $q^i$, under these assumptions (Haft, Hofmann, & Tresp, 1999), takes the general form

$$q^i(s^i) \propto \exp\left\{ \frac{1}{\tau} \int \cdots \int q^{@i}(s^{@i}) \ell(s^i, s^{@i} \mid \mathcal{D}; \theta)\, d\, s^{@i} \right\} \tag{23}$$

where the index set $@i$ denotes the Markov blanket of variable $i$ (i.e. the set of nodes neighbouring node $i$). In a 2 player game this set consist of just the opponent, while in a graphical game the Markov blanket consists of all players affecting player $i$'s actions (Kearns et al., 2001; Pearl, 1988).

The variational algorithm thus proceeds by iterating the update equation 23 until the KL divergence expressed in equation 21 has converged to a stationary value — possibly a local extremum. In the case of the EM algorithm, one round of updates by equation 23 will be interlaced with one round of updates of parameter $\theta$. The update equations for $\theta$ are obtained by differentiation of equation 19 with respect to $\theta$.

From a game-theoretic perspective, the log-probability shown in equation 18, with instantiated observations $\mathcal{D}$, can be interpreted as a global reward function

$$r(S \mid \theta) \equiv \log\left(p(S \mid \mathcal{D}, \theta)\right) \tag{24}$$

parameterised by $\theta$. The $I$ latent variables then become synonymous with the players, each player having a strategy space consisting of all possible values of $S^i$ and mixed strategy

$q^i$. Interpreted in terms of players, the task of probabilistic inference is now a partnership game, played by $I$ players jointly against nature (Grünwald & Dawid, 2004). The total expected reward

$$\mathcal{L}(\theta) \equiv \int \cdots \int q(s) \log \left( p(s \mid \mathcal{D}, \theta) \right) \, d\, s \tag{25}$$

given some set of mixed strategies, is simply the expected log-likelihood. In the Bayesian setting, with priors on $\theta$, the global reward is the full log-posterior distribution and the equivalent total expected reward is the marginal log-likelihood, or evidence. If $p$ factorises then it can often be represented as a graphical model. Within a game-theoretic interpretation, graphical models can be seen as players having their own, local, reward functions, and such graphical games (Kearns et al., 2001) are an active research area which implicitly makes use of the analogy between game theory and graphical models.

### 4.2 Fictitious Variational Play

The variational mean field algorithm, described in the previous section, suggests that the mixed strategies that maximise equation 24 can be determined by iterating over equation 23 and gradually taking the limit $\tau \to 0$. This is analogous to a Cournot adjustment process in the game theory literature, with the modification that *smooth* best responses are used in place of best responses[2]. However, a well known shortcoming of the Cournot process' iteration of best response is that it can often fail to converge and, instead, exhibit cyclic behaviour. Consider, for instance, the partnership game with reward matrix

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

If the players commence play by choosing different actions, the Cournot process fails to converge and the iterated best responses will exhibit cyclic solutions.

Such cyclic behaviour can indeed be observed in the variational mean field algorithm. Whilst not commonly reported, cycles can also occur in highly connected graphical models (e.g. Markov Random Fields or Coupled Hidden Markov Models), and when the probability distribution being approximated does not really support the imposed independence assumption in the mean field approach (i.e. when random variables are strongly instead of weakly connected). Clearly, especially in the latter case, even random initialisation cannot avoid the mean field algorithm's problem of convergence. This phenomenon is described as "thrashing" by Wolpert et al.'s (2006).

***Example****: This simple example illustrates our point. Suppose we have two observations,* $\mathcal{D} = \{y_1, y_2\}$ *with* $y_1 = 10$ *and* $y_2 = -10$. *We know that each observation is drawn from a two-component mixture of one dimensional normal distributions, with mixing probability* $0.5$ *and both variances set to* $1$. *The means of the normal distributions are chosen from the*

2. In the Cournot adjustment process, players use strategies that are a best response to the action the opponents used in the previous period. In many cases, this process does not converge (Fudenberg & Levine, 1999).

*set $\{-10, 10\}$. Therefore, we have for*

$$\begin{aligned}\ell(\mu^1, \mu^2 \mid \mathcal{D}) = & \log(0.5(\phi(d_1 - \mu^1) + \phi(d_1 - \mu^2))) \\ & + \log(0.5(\phi(d_2 - \mu^1) + \phi(d_2 - \mu^2)))\end{aligned}$$

*where $\phi(y_i - \mu^j)$ for $i, j = 1, 2$ denotes the density of $y_i - \mu^j$ under a standard normal distribution. Now by symmetry, it should be clear that $l_{same} \equiv l(10, 10) = l(-10, -10) = \log\phi(0) + \log\phi(20) \approx -201.8379$ and $l_{different} \equiv l(10, -10) = l(-10, 10) = 2\log(0.5(\phi(0) + \phi(20))) \approx -3.2242$. From a game theory perspective, this means that the partnership game we are playing is simply given by the $2 \times 2$ matrix*

$$\begin{pmatrix} -201.8379 & -3.2242 \\ -3.2242 & -201.8379 \end{pmatrix}.$$

*If we choose two q-distributions of equation 23 for each component mean, initialise them with $q^i(\mu^i = 10) = 0.9 = 1 - q^i(\mu^i = -10)$ for component index $i = \{1, 2\}$ (i.e. each marginal distribution places weight 0.9 on having the mean at 10), and update them simultaneously, both distributions switch virtually all of their mass to the $-10$ point. This shouldn't be a surprise, given that we clearly need to have $\mu^1 \neq \mu^2$. The problem is that both components of the mean field approximation jump at the same time. At the next step, the reverse move will happen. Each time things get more extreme, and continuous cycling occurs.*

In the work by Wolpert et al. (2006) it was suggested that thrashing can be avoided by adjusting the distributions toward the best responses, instead of setting them to *be* the best responses. Here, we use the analogy with game theory to rigorously develop this approach, and prove its convergence. We start by noting that fictitious play does exactly this; it modifies the Cournot process by adjusting players' mixed strategies towards the smooth best response, instead of setting it to be the smooth best response. This suggests a fictitious play-like modification to the standard mean field algorithm, in which best responses computed at time $t$ are mixed with the current distributions:

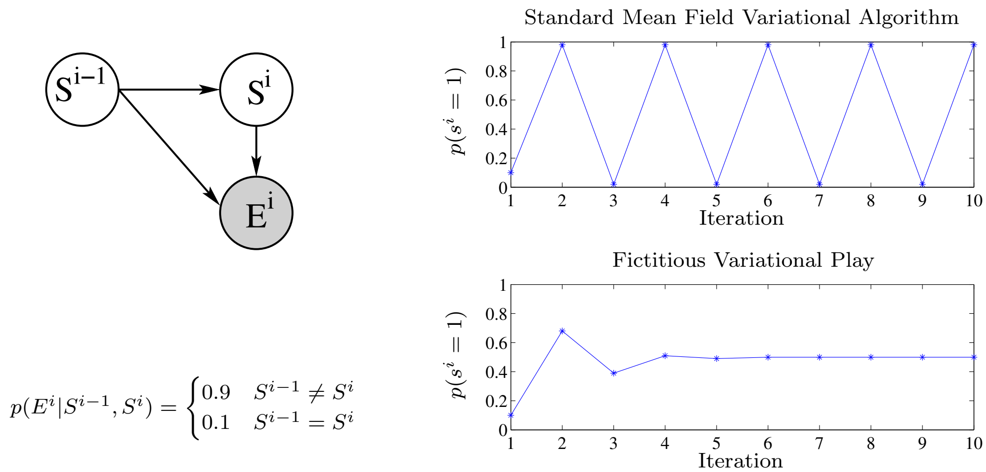
$$q^i_{t+1} = (1 - \lambda_t) q^i_t + \lambda_t \beta^i_{MF}\left(q^{-i}_t\right) \tag{26}$$

where $\lambda_t$ is a sequence satisfying the usual Robins-Monro conditions (Bishop, 2006) such that $\sum_{t=1}^{\infty} \lambda_t = \infty$, $\lim_{t\to\infty} \lambda_t = 0$, $\sum_{t=1}^{\infty} \lambda_t^2 < \infty$, and $\beta^i_{MF}\left(q^{-i}_t\right)$ denotes the best response function for distribution $i$ and is given by equation 23. We call this process variational fictitious play.

The game theory literature proves that stochastic fictitious play in partnership games is guaranteed to converge when $\mathbb{S}^i$ is finite (Leslie & Collins, 2005). This allows us to prove the following result:

**Theorem:** If each random variable is discrete, the variational fictitious play converges.

**Proof:** In the case of discrete random variables, equation 26 defines a generalised weakened fictitious play process for the game where each player receives reward $\log(p(s \mid \mathcal{D}, \theta))$. This




(a) (b)

Figure 2: Comparison of the performance of the standard mean field variational algorithm and our improved algorithm when applied to an exemplar binary hidden Markov model.

process is known to converge (Leslie & Collins, 2005).

**Remark:** While an equivalent theorem is not yet available for continuous random variables, a similar approach may yield suitable results when combined with recent work of Hofbauer and Sorin (2006).

***Example:*** *As an example of the change the fictitious update scheme has on the standard variational algorithm, consider the binary hidden Markov model shown in Figure 2a. The model contains two latent nodes, indexed by* $i$ *and* $i-1$*, which are jointly parent to the observed variable* $E^i$*. Both latent variables take values* $S^{i-1}, S^i \in \{0,1\}$*. The model is specified such that observation is best explained (with probability* 0.9*; see Figure 2b), if the two neighbouring states take different values. Further, the joint prior for states* $i-1$ *and* $i$ *is the uniform distribution. For simplicity we omit the parameter* $\theta$*, which encodes observation state probabilities and the prior distribution.*

- *Consider initialising so that both the distributions* $q^1(s^1 = 1)$ *and* $q^2(s^2 = 1)$ *are close to* 1*. The result is a periodic flipping of the state probability distributions,* $q^1$ *and* $q^2$*, at every update iteration of the mean field algorithm (top graph in Figure 2b). In contrast, the fictitious variational play scheme gradually forces the algorithm to converge (bottom graph in Figure 2b).*

- *Initialisation at random reduces the likelihood of the mean field algorithm's failure to converge. However, empirically, we could still observe such cyclic solutions in* 50% *of all the random starts of the hidden Markov model training procedure. In contrast, the modified mean field variational algorithm converges reliably in all cases.*

### 4.3 Derivative Action Variational Algorithm

In the previous sections we have shown that, when viewed from a game theoretic perspective, the smooth best response function is equivalent to the mean field update rule and, consequently, we were able to apply results from game theory to derive an improved variational learning algorithm. In this section we go further and use this equivalence to incorporate a recent development in fictitious play, specifically dynamic fictitious play, into the variational mean field method.

In dynamic fictitious play (Shamma & Arslan, 2005) the standard best response expression, shown in equation 7, is extended to include an additional term

$$\beta\left(q_t^{-i} + \underbrace{\gamma \frac{d}{dt} q^{-i}}_{\text{new}}\right). \tag{27}$$

This additional derivative term acts as an "anticipatory" component of the opponent's play and is expected to increase the speed of convergence in the partnership game considered here. Note that, on convergence, the derivative terms become zero and the fixed points are exactly the same as for the standard variational algorithm.

Based on the correspondence between the best response function in fictitious play and the model-free update expression of variational learning in section 4.1, we incorporate this additional differential term into the update equation of our variational learning algorithm to derive a *derivative action variational algorithm* (DAVA) that displays improved convergence properties compared to the standard variational algorithm.

To illustrate this procedure, we consider the case of applying this derivative action variational algorithm to the problem of performing clustering (i.e. learning and labelling a mixture distribution in order to best fit an experimental data set). We chose clustering as it represents a canonical problem, widely reported in the literature of machine learning. We consider a standard mixture distribution consisting of $K$ one-dimensional Gaussian distributions, $\mathcal{N}(d|\mu_k, \beta_k)$ for $k = 1, \ldots, K$, given by

$$p(d|\theta) = \sum_{k=1}^{K} \pi_k \mathcal{N}(d|\mu_k, \beta_k) \tag{28}$$

where $\theta$ is the set of distribution parameters $\{\mu_1, \beta_1, \pi_1, \cdots, \mu_K, , \beta_K, \pi_K\}$. Here, $\mu_k$ and $\beta_k$ are the mean and precision of each Gaussian, and $\pi_k$ is their weighting within the mixture. The usual approach to learning this distribution (Redner & Walker, 1984) is to assume the existence of an indicator (or labelling) variable, indexed by $i$ and which takes values $S^i \in \{1, \cdots K\}$, for each sample, $d^i$, and formulate the complete likelihood

$$p(S^i, \mu_1, \ldots, \mu_K, \beta_1, \ldots, \beta_K \mid d^i) \propto \prod_{k=1}^{K} \pi_k^{\delta(S^i=k)} \mathcal{N}(d^i|\mu_k, \beta_k)^{\delta(S^i=k)} p(\theta) \tag{29}$$

where $p(\theta)$ denote the parameter prior distributions. To obtain analytic coupled update equations for each member of the parameter set, $\{\pi_1, \mu_1, \beta_1, \cdots, \pi_K, \mu_K, \beta_K\}$, we use the model discussed in Uedaa and Ghahramani (2002), which describes the appropriate choice of the approximating marginal posterior distribution for each parameter. By evaluating the integral in equation 23, after replacing the generic latent variables $S$ by the model parameters $\theta$ and the indicator variables $S$, Uedaa and Ghahramani show that a closed form expression can be derived for the approximating marginal posterior distributions for all parameters. To compute the approximate marginal posterior mean distribution, for example, the set of variables $S^i$ and $S^{@i}$ in equation 23 become place holders for $\mu_k$ and $\{\beta_k, \pi, S^1, \cdots, S^N\}$, respectively. In other words, to compute the posterior distribution of $\mu_k$, the logarithm of equation 23 must be averaged with respect to the distributions $q(\beta_k)$, $q(\pi)$, and $q(S^i), i = 1, \cdots N$.

The computations result in a closed form solution for the marginal posterior for each element of the parameter set. Thus, the posterior of the means, $\mu_k$, is a normal distribution

$$q^{\mu_k} \triangleq \mathcal{N}(\mu_k | m_k, \tau_k) \tag{30}$$

with mean $m_k$ and precision $\beta_k$ and where

$$\begin{aligned} \mu_k &= \left(c_k b_k \bar{d}_k + \tau_0 \mu_0\right) \tau_k \\ \bar{\lambda} &= \textstyle\sum_i \lambda_k^i \\ \lambda_k^i &= q^i(S^i = k) \\ \bar{d}_k &= \textstyle\sum_i \lambda_k^i d^i \\ \tau_k &= c_k b_k \bar{\lambda}_k + \tau_0 \end{aligned}$$

and $\mu_0$ and $\tau_0$ are, respectively, the mean and precision of the Gaussian prior for $\mu_k$. The posterior for the precisions is a Gamma distribution

$$q^{\beta_k} \triangleq \Gamma(\beta_k | b_k, c_k) \tag{31}$$

where

$$\begin{aligned} c_k &= \frac{1}{2}\bar{\lambda}_k + \alpha_0 \\ b_k &= \frac{1}{2} \textstyle\sum_i \lambda_k^i (d^i - \mu_k)^2) + \frac{1}{2}\bar{\lambda}_k \tau_k^{-1} + \beta_0 \end{aligned}$$

where $\alpha_0$ and $\beta_0$ are, respectively, are the shape and precision parameters of the Gamma prior for the precisions $\beta_k$. Finally, the posterior of the mixture weights is a $K$-dimensional Dirichlet distribution

$$q^{\pi} \triangleq Dir(\pi | \kappa) \tag{32}$$

where

$$\kappa = \textstyle\sum_i \lambda_k^i + \kappa_0$$

and $\kappa_0$ is the parameter of the Dirichlet prior for $\pi$. Finally, the distribution of the component labels $s^i$ has the form

$$q^i(S^i) = \frac{1}{Z_{S^i}} \prod_{k=1}^{K} \tilde{\pi}_k^{\delta(S^i = k)} \tilde{\beta}_k^{\frac{1}{2}} \exp\left\{ -\frac{b_k c_k}{2} \left[ (d^i - m_k)^2 + \frac{1}{\tau_k} \right] \right\} \tag{33}$$

where the normalising constant $Z_{S^i}$ is computed over the finite set of states of $S^i$. The values for $\tilde{\pi}_k$ and $\tilde{\beta}_k$ are computed using the relations

$$\begin{aligned}\tilde{\pi}_k =& \Psi(\kappa_k) - \Psi\left(\textstyle\sum_{l=1}^{K} \kappa_l\right)\\ \tilde{\beta}_k =& \Psi(c_k) + \log(b_k).\end{aligned}$$

$\Psi$ is the digamma function.

As noted earlier, the update equations can be interpreted as the best response of a particular parameter given the values of the others. Thus, the additional derivative term seen in dynamic fictitious play can, in principle, be included into the variational update equations 30, 31, 32, and 33. If it is desired, however, to obtain closed form solutions for the modified best response functions, or update equations, the derivative term can only be incorporated in the discrete distribution, given by equation 33, as follows.

For notational clarity we will only add the anticipatory component to the estimate of the means, $\mu_k$, and use this only for the update of the $s^k$. That is, we will consider only the inclusion of the $\frac{d}{dt}q^{\mu_k}$ term to equation 33. Furthermore, we will approximate the derivative $\frac{d}{dt}q^{\mu_k}$ by the discrete difference in the distributions between iterations $t$ and $t-1$

$$\frac{d}{dt}q^{\mu_k} \approx q_t^{\mu_k} - q_{t-1}^{\mu_k}. \tag{34}$$

It is also possible to implement a smoothed version of the derivative to provide added robustness against random fluctuations, as was done in the equivalent fictitious play algorithm described by Shamma and Arslan (2005). When we introduce the derivative term into equation 33 the update equation for the component labels becomes

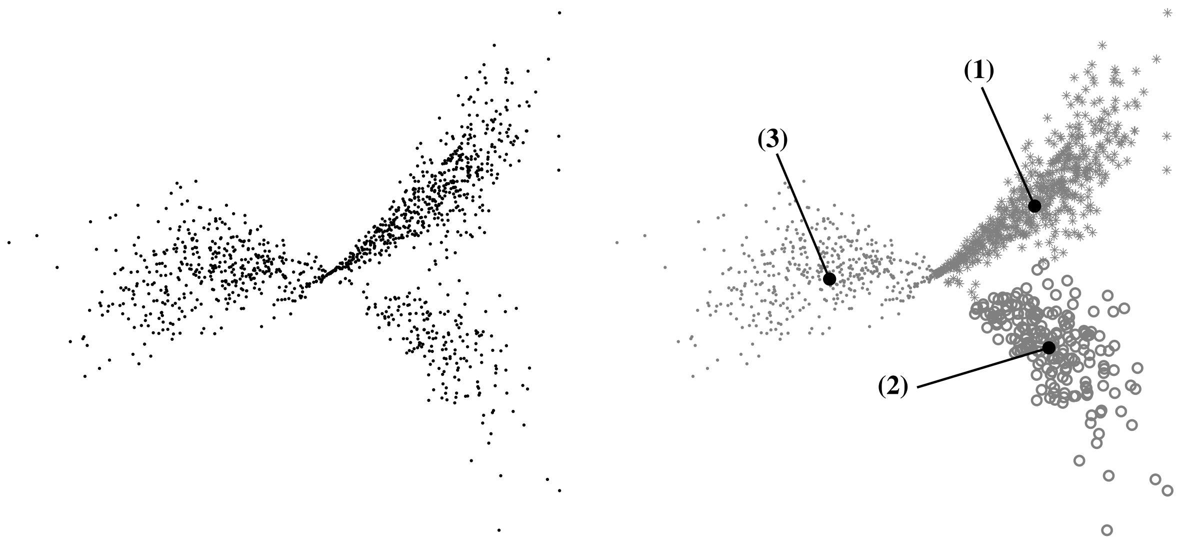
$$\begin{aligned}q^i(S^i) \propto \prod_{k=1}^{K} \tilde{\pi}_k^{\delta(S^i=k)} \tilde{\beta}_k^{\frac{1}{2}} \exp\Bigg\{ -\frac{b_k c_k}{2}(1+\gamma)\left[\left(d^i - m_{k_t}\right)^2 + \frac{1}{\tau_{k_t}}\right] \\ + \frac{b_k c_k}{2}\gamma\left[\left(d^i - m_{k_{t-1}}\right)^2 + \frac{1}{\tau_{k_{t-1}}}\right]\Bigg\}\end{aligned} \tag{35}$$

where the parameters of $q_t^{\mu_k}$ are denoted by $m_{k_t}$ and $\tau_{k_t}$. The differential coefficient, $\gamma$, allows us to control the degree by which the differential term contributes toward the overall update equation.

In order to demonstrate empirically the convergence properties of the derivative action variational inference algorithm, we compared the algorithm using either update equation 33 or update equation 35 to test data. To control the problem setting, we generated synthetic test data by drawing data points from a mixture of 3 Gaussian distributions and then applying a non-linear mapping function[3]. This data is shown in Figure 3 along with the optimal clustering solution.

We then performed 200 runs comparing the standard variational algorithm (implemented as described in Penny & Roberts, 2002) with the derivative action variational algorithm. During each run, both algorithms were started from an identically and randomly initialised

3. Equation 35 generalises to two dimensions or higher by replacing the quadratic term $(d^i - m_{k_t})^2$ with the inner product $(\vec{d^i} - \vec{m}_{k_t})^\top(\vec{d^i} - \vec{m}_{k_t})$. This assumes vector valued data samples, $\vec{d^i}\ \forall i$, are Gaussian distributed with means $\vec{m}_k$ and homoscedastic precision $\beta_k$.




(a) (b)

Figure 3: Synthetic data (a) used for empirical study of the convergence properties of the derivative action variational inference algorithm and the optimal clustering solution (b) showing the centres of each of the three Gaussians in the mixture.

set of model parameters and at each iteration we measured the value of the Kullback-Leibler divergence of equation 21. The algorithms were terminated after 60 iterations (chosen to be well above the number of iterations required to achieve a relative change of Kullback-Leibler divergence of less than $10^{-5}$). The differential coefficient, $\gamma$, in equation 35 was set to a constant value throughout. For illustrative purposes we chose the following values for this coefficient: $\gamma = 0.5$, 1.0, 1.5, and 2.0.

We first consider the difference in the convergence rate between the standard variational algorithm and the derivative action variational algorithm. Due to the nature of both algorithms, different initial conditions can lead to different final clustering solutions. In addition, the difference in the update rule of the two algorithms means that, although we start both algorithms with identical initial conditions, they do not necessarily result in identical final clustering solutions. Thus, we analyse the difference in the Kullback-Leibler divergence at each iteration, for just those runs that did in fact produce identical clustering solutions at termination. In Figure 4 we compare these algorithms for the case when $\gamma = 0.5$. As can be clearly seen, the DAVA converges everywhere more quickly than the standard algorithm.

Our experiments also indicate that the choice of the differential coefficient, $\gamma$, has a significant effect on the convergence behaviour of DAVA. This is seen in Figure 5 for the case when $\gamma = 1.0$. Compared to the results in Figure 4, the magnitude of the difference in Kullback-Leibler divergence is much larger, indicating a substantial increase in the convergence rate.

By comparing the times at which each algorithm reached equilibrium (indicated by a relative change of Kullback-Leibler divergence of less than $10^{-5}$), for all four values of $\gamma$,

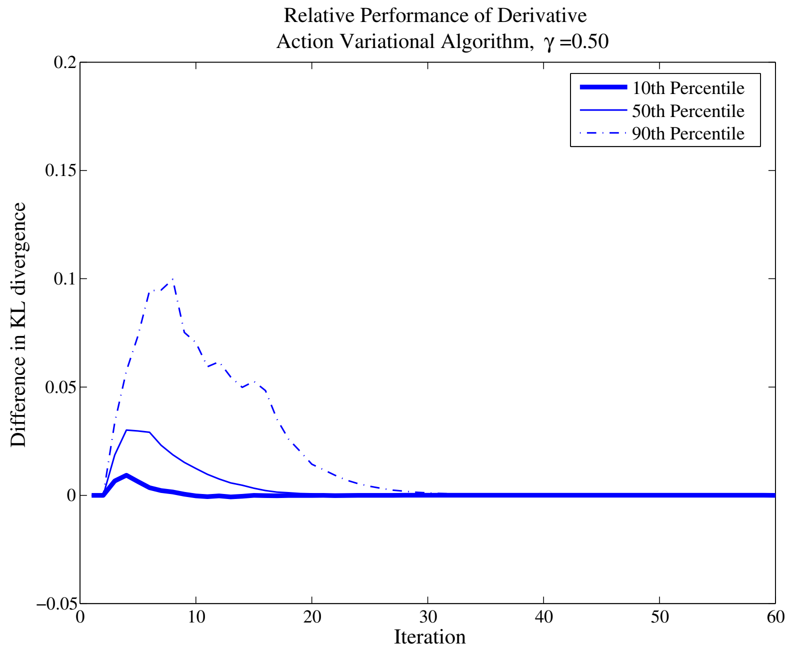



Figure 4: For comparison of the standard and the derivative action variational algorithm ($\gamma = 0.5$), the Kullback-Leibler divergence values, obtained from both algorithms and at every iteration, were compared. Shown here are the estimates of the $10^{th}$, $50^{th}$ (the median) and $90^{th}$ percentiles of the differences between the standard KL and the derivative action KL, i.e. $KL_{standard}(t) - KL_{DAVA}(t)$ at iteration $t$. A positive value suggests that the current solution of the standard algorithm is worse than than of the derivative action algorithm. A zero value implies that the solutions found by both algorithms are identical. At initialisation the KL differences are zero as both algorithms have identical initialisation conditions.

the role of the differential coefficient in improving convergence rates becomes apparent. In particular, Table 1 shows the relative convergence rate improvement shown by the derivative action variational algorithm, compared to the standard variational algorithm. The results indicate that the median improvement can be as much as 28% when $\gamma = 2.0$. In addition, as the value of the differential coefficient increases, the variance in the convergence rate increases, and thus, there is a widening gap between the best and worst performance improvements.

However, this increasing variance does not imply that DAVA is converging to inferior solutions. This can be seen in Figure 6 which shows an analysis of the quality of the solutions reached by each algorithm, for all 200 runs (not just those where both algorithms converged to the same clustering solution). At moderate values of $\gamma$ ($\gamma \leq 1.5$) the derivative component can assist in finding a better solution. This is indicated by the proportion of positive KL divergence differences (in Figure 6). Positive values imply that the standard algorithm often finds worse solutions compared to DAVA at the particular setting of $\gamma$. Increasing the value

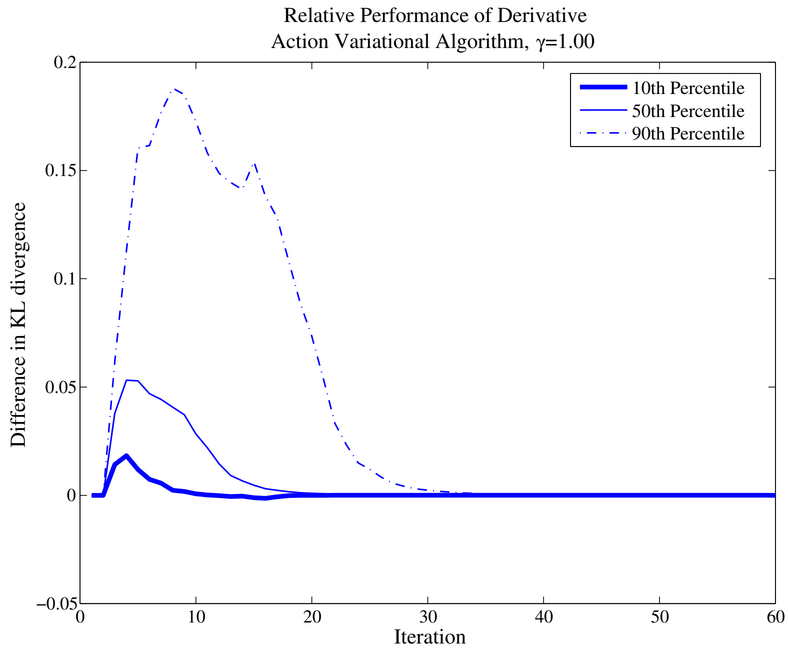



Figure 5: Estimates of the $10^{th}$, $50^{th}$ (the median) and $90^{th}$ percentiles of the differences in the Kullback-Leibler divergence values, at each iteration, after 200 runs of the standard and the derivative action variational algorithm ($\gamma = 1.0$).

of the differential coefficient further increases the variance in the clustering solutions that are generated by the algorithm.

Having evaluated the performance of the DAVA algorithm on a synthetic data set, and investigated its performance over a range of values for the derivative coefficient, $\gamma$, we now consider a more realistic experimental setting and apply the algorithm to medical magnetic resonance data (see Figure 7). This data consists of a $100 \times 100$ pixel region of a slice through a tumour patient's brain. Data was collected using both T2 and proton density (PD) spin sequences (shown in Figure 7a), which are used directly to form a two-dimensional feature space.

A 10 component Gaussian mixture model is fitted to this data space, and as before, we use the DAVA algorithm derived earlier to learn and label the most appropriate mixture distribution. Following the synthetic data experiments, the DAVA derivative coefficient $\gamma$ was set to 1.0. This being the best compromise between speed and robustness of the algorithm. We let the algorithms run for 100 iterations and measured the KL divergence at each iteration to monitor convergence. Figure 7b shows the resulting segmentation from the DAVA algorithm.

On this dataset, the algorithm converged to $K = 5$ classes; identical to the Markov Chain Monte Carlo clustering reported in the work of Roberts, Holmes, and Denison (2001). The segmentation clearly shows spatial structure which was not incorporated into the segmen-

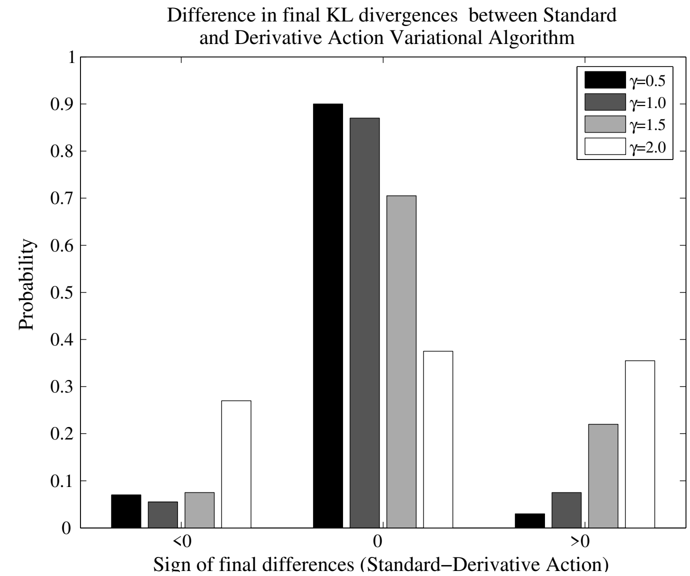



Figure 6: Relative difference in Kullback-Leibler divergence values at equilibrium for the standard and the derivative action variational algorithms with derivative coefficients: $\gamma = 0.5$, 1.0, 1.5, and 2.0.

| $\gamma$ | $10^{th}$ percentile | $50^{th}$ percentile | $90^{th}$ percentile |
|---|---|---|---|
| 0.5 | 0% | 9 % | 21 % |
| 1.0 | 0% | 19 % | 39 % |
| 1.5 | -2% | 22% | 46 % |
| 2.0 | 8 % | 28% | 53 % |

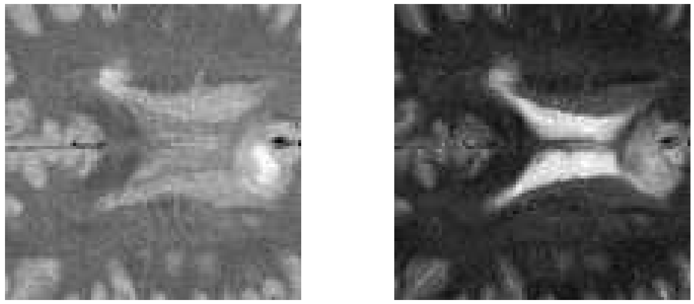
Table 1: Relative convergence rate improvement of the derivation action variational algorithm over the standard variational algorithms.

tation process *a priori*, and the algorithm was approximately 1.5 times faster than the standard segmentation algorithm[4].

In summary, by adding a derivative term to the variational learning algorithm, we have produced an algorithm that, in empirical studies, shows improved convergence rates compared to the standard variational algorithm. Indeed, this convergence rate improvement has been achieved by applying the additional derivative response term to the mean components of the mixture model parameters only, and, thus, we believe that further improvement is possible if other parameters are treated similarly.

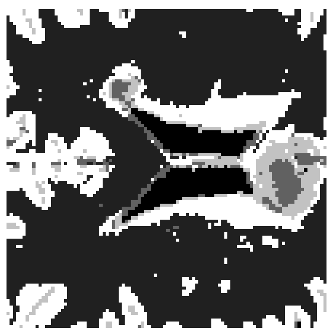
4. To determine the factor of speedup between the two algorithms we averaged the KL divergence values of every iteration and for each segmentation algorithm. Based on the mean KL divergence curve the iteration number at which the algorithm converged could be calculated. This is the iteration at which the relative change in KL divergence between 2 successive iterations was less than $10^{-4}$. The ratio of the convergence points of the DAVA and the standard algorithm then produced our indicator of speed.

Slice of a T2 magnetic resonance image through a tumour in a patient's brain.



(a)

The proton density image corresponding to the T2 MR image.



(b)

Figure 7: The results of the segmentation using the DAVA algorithm with $\gamma = 1.0$. For this data the segmentation was obtained at approximately half the time it took the standard segmentation algorithm.

## 5. Conclusions and Future Work

In this work we have shown an equivalence between game-theoretical learning in partnership games and the variational learning algorithm common in machine learning. In this comparison, probabilistic inference using the mean field variational learning is seen to be a Cournot adjustment process in a partnership game. Likewise, the smooth best response function used by players is a plug-in single layer classifier with a sigmoid activation function.

By exploiting this analysis and the insights derived from it, we have been able to show that insights from one area may be applied to the other to derive improved fictitious play and variational learning algorithms. In empirical comparisons, these algorithms showed improved convergence properties compared to the standard counterparts.

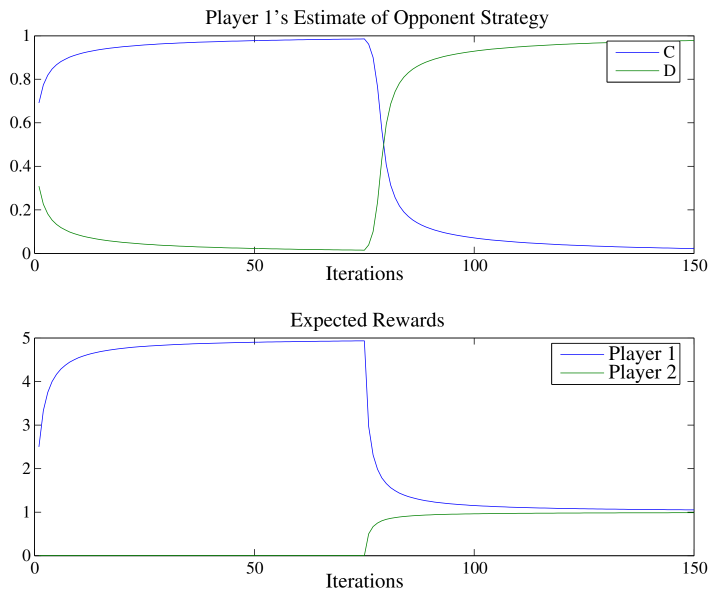



Figure 8: Using dynamic logistic regression (DLR) for adaptive estimation of opponent mixed strategies. In a game of repeated prisoner's dilemma one player changes strategies from "always co-operate" to "always defect" midway through the game. Using DLR as model of learning, the other player adapts to the changes based on the uncertainty between the predicted and observed opponent strategies.

We believe these initial results are particularly exciting. Thus, whilst there still remains further analysis to be performed (specifically, we would like to prove convergence of moderated fictitious play and compare the performance of our variational algorithms on large real world data sets), the work clearly shows the value of studying the intersection of machine learning and game theory.

One conclusion from this work is that almost any machine learning algorithm can be put to use as a model of players learning from each other. Consider, for instance adaptive classification algorithms which are capable of adapting to changes in the learner's environment. Most game-theoretic methods are tuned toward achieving equilibrium and estimating, for example, Nash mixed strategy profiles. While this is desirable in stationary environments, such algorithms fail to adapt to non-stationary environments, and are consequently of little practical use in dynamic real-world applications. The research presented in this paper strongly suggests that adaptive classifiers might prove useful.

As a proof of concept, we have implemented the dynamic logistic regression (DLR), presented in Penny and Roberts (1999), as a model of play for one player (player 1) in a game of repeated prisoner's dilemma. The other player (player 2) was set to play an "always co-operate" strategy for the first 75 rounds of play. For the second set of 75 rounds, player 2 was set to play a "always defect" strategy. The task of player 1 is to detect this change in the opponent's behaviour and compute the best responses according to the updated estimate of player 2's strategy. The estimates of player 1 of player 2's strategy for the entire game are shown in Figure 8, together with both players' expected rewards. The DLR adaptively changes the one-step ahead predicted opponent strategy on the basis of the uncertainty that results from incorporation of the most recently observed opponent action (see Penny & Roberts, 1999 for details; the observations considered in this work map directly to the observed actions made by the opponent). The decision about which action to play follows then as usual (i.e. compute the best response according to the updated estimate using equation 9).

There are two things we would like to point out. First, as implemented here, the input required for DLR is simply the recently observed opponent action, and the decision made by the DLR is the action drawn from the best response function. However, DLR also allows for a vector of inputs, and as a consequence, players can be made to respond not just to the opponent's actions, but also to other context or application specific variables. Second, the DLR estimation described in Penny and Roberts (1999) is fully Bayesian. Missing data can be naturally embedded in the Bayesian estimation and this is demonstrated in Lowne, Haw, and Roberts (2006). Mapping this fact back to the use of DLR as model of play implies that players can keep track of the opponent's behaviour *without* the need to follow or observe their every move. Missing observations, for instance, could result in an increased uncertainty in the opponent's predicted play and, within, a reversal toward the appropriate best response (as it might have existed at the onset of play).

Similar ideas of using dynamic, instead of static, estimators of opponent strategy have recently been presented in the work of Smyrnakis and Leslie (2008). Extending further the use of machine learning techniques to allow dynamic estimation of both strategies and environmental parameters will allow game theoretical learning to become more generally applicable in real-world scenarios.

## Acknowledgments

The authors would like the thank the reviewers, whose suggestions led to significant improvements in both the content and clarity of the final paper. This research was undertaken as part of the ARGUS II DARP and ALADDIN projects. ARGUS II DARP (Defence and Aerospace Research Partnership) 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. ALADDIN (Autonomous Learning Agents for Decentralised Data and Information Systems) is jointly funded by a BAE Systems and EPSRC (Engineering and Physical Science Research Council) strategic partnership (EP/C548051/1).

## References

Bernardo, J. M., & Smith, A. F. M. (1994). *Bayesian Theory.* John Wiley and Sons.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning.* Oxford University Press, Oxford.

Demiriz, A., Bennett, K. P., & Shawe-Taylor, J. (2002). Linear programming boosting via column generation. *Machine Learning*, *46*(1–3), 225–254.

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, *55*(1), 119–139.

Fudenberg, D., & Levine, D. K. (1999). *The Theory of Learning in Games*. MIT Press.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2000). *Bayesian Data Analysis*. Chapman & Hall/CRC.

Greenwald, A., & Littman, M. L. (2007). Introduction to the special issue on learning and computational game theory. *Machine Learning*, *67*(1-2), 3–6.

Grünwald, P. D., & Dawid, A. P. (2004). Game Theory, Maximum Entropy, Minimum Discrepancy and Robust Bayesian Decision Theory. *Annals of Statistics*, *32*, 1367–1433.

Haft, M., Hofmann, R., & Tresp, V. (1999). Model-Independent Mean Field Theory as a Local Method for Approximate Propagation of Information. *Computation in Neural Systems*, *10*, 93–105.

Hofbauer, J., & Hopkins, E. (2005). Learning in perturbed asymmetric games. *Games and Economic Behavior*, *52*, 133–157.

Hofbauer, J., & Sorin, S. (2006). Best response dynamics for continuous zero-sum games. *Discrete and Continuous Dynamical Systems*, *B6*, 215–224.

Husmeier, D., Dybowski, R., & Roberts, S. J. (Eds.). (2004). *Probabilistic Modeling in Bioinformatics and Medical Informatics*. Advanced Information and Knowledge Processing. Springer Verlag.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., & Saul, L. K. (1997). An Introduction to Variational Methods for Graphical Models. In Jordan, M. I. (Ed.), *Learning in Graphical Models*. Kluwer Academic Press.

Kalai, E., & Lehrer, E. (1993). Rational Learning Leads to Nash Equilibrium. *Econometrica*, *61*(5), 1019–1045.

Kearns, M., Littman, M. L., & Singh, S. (2001). Graphical Models for Game Theory. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, pp. 253–260.

Lambert, T., Epelman, M. A., & Smith, R. L. (2005). A Fictitious Play Approach to Large-Scale Optimization. *Operations Research*, *53*(3), 477–489.

Lee, C. F., & Wolpert, D. H. (2004). Product Distribution Theory for Control of Multi-Agent Systems. In *AAMAS '04: Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems*, pp. 522–529, New York, USA.

Leslie, D. S., & Collins, E. J. (2005). Generalised weakened fictitious play. *Games and Economic Behavior*, *56*(2), 285–298.

Liu, W., Besterfield, G., & Belytschko, T. (1988). Variational approach to probabilistic finite elements. *Journal of Engineering Mechanics*, *114*(12), 2115–2133.

Lowne, D., Haw, C., & Roberts, S. (2006). An adaptive, sparse-feedback EEG classifier for self-paced BCI. In *Proceedings of the Third International Workshop on Brain-Computer Interfaces, Graz, Austria*.

MacKay, D. J. C. (1992). The Evidence Framework Applied to Classification Networks. *Neural Computation*, *4*(5), 720–736.

MacKay, D. J. C. (2003). *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers.

Penny, W., & Roberts, S. (1999). Dynamic logistic regression. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN'99)*, Vol. 3, pp. 1562 – 1567.

Penny, W., & Roberts, S. (2002). Bayesian Multivariate Autoregressive Models with Structured Priors. *IEE Proceedings on Vision, Signal and Image Processing*, *149*(1), 33–41.

Redner, R. A., & Walker, H. F. (1984). Mixture Densities, Maximum Likelihood and the EM Algorithm. *SIAM Review*, *26*(2), 195–239.

Ripley, B. (2000). *Pattern Recognition and Neural Networks*. Cambridge University Press.

Robert, C. P., & Casella, G. (1999). *Monte Carlo Statistical Methods*. Springer-Verlag: New York.

Roberts, S., Holmes, C., & Denison, D. (2001). Minimum Entropy data partitioning using Reversible Jump Markov Chain Monte Carlo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *23*(8), 909–915.

Shamma, J. S., & Arslan, G. (2005). Dynamic Fictitious Play, Dynamic Gradient Play and Distributed Convergence to Nash Equilibria. *IEEE Transactions on Automatic Control*, *50*(3), 312–327.

Shoham, Y., Powers, R., & Grenager, T. (2007). If multi-agent learning is the answer, what is the question. *Artificial Intelligence*, *171*(7), 365–377.

Smyrnakis, M., & Leslie, D. (2008). Stochastic Fictitious Play using particle Filters to update the beliefs of opponents strategies. In *Proceedings of the First International Workshop on Optimisation in Multi-Agent Systems (OPTMAS) at the Seventh International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2008)*.

Uedaa, N., & Ghahramani, Z. (2002). Bayesian model search for mixture models based on optimizing variational bounds. *Neural Networks*, *15*, 1223–1241.

van der Genugten, B. (2000). A Weakened Form of Fictitious Play in Two-Person Zero-Sum Games. *International Game Theory Review*, *2*(4), 307–328.

Wolpert, D. H. (2004). Information Theory - The Bridge Connecting Bounded Rational Game Theory and Statistical Physics. arXiv.org:cond-mat/0402508.

Wolpert, D., Strauss, C., & Rajnarayan, D. (2006). Advances in Distributed Optimization Using Probability Collectives. *Advances in Complex Systems*, *9*(4), 383–436.

Young, H. P. (1998). *Individual Strategy and Social Structure: An Evolutionary Theory of Institutions*. Princeton University Press.