The Problem: The topic of learning in multi-agent environments has received increasing attention over the past several years. Game theorists have begun to examine learning models in their study of repeated games, and reinforcement learning researchers have begun to extend their single-agent learning models to the multiple-agent case. As traditional models and methods from these two fields are adapted to tackle the problem of multi-agent learning, we need to revisit the central issue of optimality and its relation to the agents’ beliefs.

Motivation: Most interesting domains in the real world involve interaction between two or more agents acting in some environment, possibly with competing objectives. Such domains include auctions, search-and-rescue teams, market makers, and games such as soccer, just to name a few. Even seemingly single-agent domains often include a changing environment that could be better modeled as a responsive agent. In many of these environments, the maximum reward that our agent can derive often depends on the types of opponents it is facing. Thus, optimality conditions and the agent’s beliefs about its opponents are inherently dependent.

Previous Work: Most previous work focused on either Nash equilibrium or best-response criteria as optimality conditions for learning algorithms in multi-agent domains. Best-response is a fairly intuitive criteria, essentially stating that such a policy will derive the maximal expected reward for our agent given that its opponents’ policies are fixed. Nash equilibrium exists when all the agents in the system are playing best-response strategies to all other agents.

\[
R_1 = \begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix} \quad R_1 = \begin{bmatrix} 0 & -1 & 1 \\ 1 & 0 & -1 \end{bmatrix} \quad R_1 = \begin{bmatrix} 0 & 3 \\ 1 & 2 \end{bmatrix} \quad R_1 = \begin{bmatrix} 2 & 0 \\ 3 & 1 \end{bmatrix}
\]

\[
R_2 = -R_1 \quad R_2 = -R_1 \quad R_2 = \begin{bmatrix} 0 & 1 \\ 3 & 2 \end{bmatrix} \quad R_2 = \begin{bmatrix} 2 & 3 \\ 0 & 1 \end{bmatrix}
\]

(a) Matching pennies (b) Rock-Paper-Scissors (c) Hawk-Dove (d) Prisoner’s Dilemma

Figure 1: Some common examples of matrix games.

In the reinforcement learning literature, multi-agent domains are usually modeled as stochastic (Markov) games, or in the simpler case, repeated matrix games in which agents repeatedly play one-shot matrix games such as those shown in Figure 1. Most of the resulting algorithms focus their attention on choosing a single stationary policy \( \mu \) that will maximize the learner’s expected rewards in all future time periods given that we begin in time \( t \), \( \max_{\mu} \mathbb{E}_\mu \left[ \sum_{t=0}^{T} \gamma^{T-t} R^t(\mu) \right] \), where \( T \) may be finite or infinite, and \( \mu = P D(A) \), a probability distribution over possible actions. In the infinite time-horizon case, we often use a discount factor \( 0 < \gamma < 1 \). Littman [4] analyzes this framework for zero-sum games, proving convergence to the Nash equilibrium for his minimax-Q algorithm playing against another minimax-Q agent. Hu and Wellman [3] focus on general-sum games. These algorithms share the common goal of finding and playing a Nash equilibrium. Bowling and Veloso [1] and Nagayuki et al. [5] propose to relax the mutual optimality requirement of Nash equilibria by considering rational agents, which always learn to play a stationary best-response to their opponent’s strategy, even if the opponent is not playing an equilibrium strategy. The motivation is that it allows the agents to act rationally even if the opponent is not acting
rationally because of physical or computational limitations.

**Approach:** The approaches in this previous work succeed against certain types of opponents. The capabilities of these algorithms are implicitly tied to their beliefs about the opponent’s capabilities, though this link is often not made explicit. We thus propose a general classification that categorizes algorithms by the cross-product of their possible strategies and their possible beliefs about the opponent’s strategy, \( \mathcal{H} \times \mathcal{B} \). An agent’s possible strategies can be classified based upon the amount of history it has in memory, from \( \mathcal{H}_0 \) to \( \mathcal{H}_\infty \). Given more memory, the agent can formulate more complex policies, since policies are maps from histories to action distributions. An agent’s belief classification mirrors the strategy classification in the obvious way. Agents that believe their opponent is memoryless are classified as \( B_0 \) players, \( B_t \) players believe that the opponent bases its strategy on the previous \( t \)-periods of play, and so forth.

Within each league of players \( \mathcal{H}_i \times \mathcal{B}_j \), we assume that the players are fully rational in the sense that they can fully use their available histories to construct their future policy. However, an important observation is that the resulting “rational” behavior depends on their beliefs about the opponent. If we believe that our opponent is a memoryless player, then even if we are an \( \mathcal{H}_\infty \) player, our fully rational strategy is to simply model the opponent’s stationary strategy and play our stationary best response. Thus, our belief capacity and our history capacity are inter-related. Without a rich set of possible beliefs about our opponent, we cannot make good use of our available history.

**Impact:** Currently most multi-agent learning algorithms do not explicitly account for the agent’s beliefs about its opponents. Our approach makes this belief a central part of the design of future algorithms and greatly affects how we evaluate the effectiveness of new algorithms.

We have demonstrated this by creating a new algorithm called PHC-Confuser that possesses specific beliefs about its opponents. Using its belief that the opponent is a best-response learner, the PHC-Confuser algorithm can model the opponent’s changing policies over time and exploit the opponent’s behavior to achieve unbounded undiscounted rewards over time in certain games. Against other types of fair opponents, it also plays reasonably well, achieving near-Nash outcomes in most cases.

**Future Work:** Future work includes the exploration of other algorithms in the league \( \mathcal{H}_\infty \times \mathcal{B}_t \) that can win against larger sets of fair opponents. This may include extensions of the PHC-Confuser algorithm, as well as enhancements to algorithms such as the Multiplicative Weight approach by Freund and Schapire [2]. Opponent identification may also play an important role in this work. Our goal is to create algorithms that can achieve reasonable goals against a wide range of opponent players.

**Research Support:** This research is supported in part by a Graduate Research Fellowship from the National Science Foundation.

**References:**


