Wisdom of Crowds vs. Groupthink:
Learning in Groups and in Isolation

Conor Mayo-Wilson, Kevin Zollman, and David Danks

May 31, 2010

1 Introduction

In the multi-armed bandit problem, an individual is repeatedly faced with a choice between a number of potential actions, each of which yields a payoff drawn from an unknown distribution. The agent wishes to maximize her total accumulated payoff (in the finite horizon case) or converge to an optimal action in the limit (in the infinite horizon case). This very general model has been used to model a variety of economic phenomena. For example, individuals choosing between competing technologies, like different computer platforms, would like to maximize the total usefulness of the purchased technologies, but cannot know ahead of time how useful a particular technology will be. Others have suggested applying this model to the choice of treatments by doctors (Berry and Fristedt , 1985), crop choices in Africa (Bala and Goyal , 2008), and choice of drilling sites by oil companies (Keller, Rady, and Cripps , 2005).

The traditional analysis of strategies in bandit problems focuses on either a known finite number of actions or a discounted infinite sequence of actions (cf. Berry and Fristedt , 1985). In both these cases, strategies are evaluated according to their ability to maximize the (discounted) expected sum of payoffs. Recent interest in boundedly rational strategies have led some scholars to consider how strategies which do not maximize expected utility might perform. These strategies are evaluated according to their ability to converge in the infinite limit to choosing the optimal action, without considering their short or medium run behavior. For example, Beggs (2005) considers how a single individual who employs a reinforcement learning algorithm (due to Roth and Erev , 1995) would perform in a repeated multi-armed bandit problem.

Many of the above choice problems, like technology choice, are not made in isolation, but rather in a social context. A player can observe not only her own successes or failures, but those of some subset of the population of other consumers. As a result, several scholars have considered bandit problems in social settings (Bolton and Harris , 1999; Bala and Goyal , 2008; Keller, Rady, and Cripps , 2005). Bala & Goyal, for example, consider a myopic Bayesian maximizer placed in a population of other myopic Bayesian maximizers, and find that certain structures for the communication of results ensures that this community will converge to the optimal action, but other social structures will not.

Although Beggs and Bala & Goyal seem to utilize essentially the same metric for the comparison of boundedly rational algorithms – convergence in the limit – they are more different than they appear. Beggs considers how a single individual does when he plays a bandit in isolation; Bala & Goyal consider how a group fares when in a specific circumstance. It should be clear that the myopic maximizer of Bala & Goyal would not converge in the limit if he was in isolation. It is perhaps less clear that Beggs' reinforcement learner might not converge if placed in the wrong social circumstance.

This paper considers two different, but closely related, problems raised by these types of investigations. First, we wish to understand how a variety of different boundedly rational strategies perform on multi-armed bandit problems played both in isolation and within a community. We consider a very general multi-armed bandit problem and compare the performance of several different classes of boundedly rational learning rules. Second, we wish to make precise the notion of a strategy being convergent in the limit. In particular we will consider different answers to the questions (a) are we considering the strategy in total isolation or when placed in a social circumstance; and (b) are we considering only a single strategy or a set of strategies? We find that which boundedly rational strategies are judged as appropriate depends critically on how one answers these questions. This, we believe, makes perspicuous the choices one must make before engaging in the analysis of various boundedly rational strategies.

In section 2 we provide the details of our model of bandit problems and four general classes of boundedly rational strategies. These four classes were chosen to represent many of the strategies investigated in literatures in economics, psychology, computer science, and philosophy. Following this, we present the different formalizations of the notion of convergence in the limit in section 3. Here we provide the theorems which demonstrate which strategies meet the various definitions, and illustrate that the different definitions are distinct from one another. Section 4 concludes with a discussion of the applications and potential extensions of the results presented here.

2 The model of learning

We begin by modeling individual agents as embedded in a communication network. The communication network will be represented by a finite, undirected graph $G = \langle V_G, E_G \rangle$ with vertices V_G representing individual agents, and edges E_G representing pairs of agents who share information with one another. We will often write $g \in G$ when we mean that $g \in V_G$. By a similar abuse of notation, we use $G' \subseteq G$ to denote that $G' \subseteq V_G$. For any learner $g \in G$, define $N_G(g) = \{g' \in G \ : \ \{g, g'\} \in E_G\}$ to be the *neighborhood* of g in the network G. We assume $\{g\} \in N_G(g)$ for all $g \in G$, so that each individual observes the outcomes of her own choices. When the underlying network is clear from context, we write $N(g)$, dropping the subscript G.

In each time period, each agent chooses one of a finite number of *actions* A. We assume that the set of actions is constant for all times, and each action results (probabilistically) in an *outcome* (or payoff) from a set O.[1] There is a set Ω of possible *states of the world* which determine the probability distribution over O associated with each action.

A *learning problem* is a quadruple $\langle \Omega, A, O, p \rangle$, where Ω is a set of states of the world, A is a finite set of actions, O is a finite set of real numbers called outcomes, and p is a probability measure specifying the probability of obtaining a particular utility given an action and state of the world.

A history specifies (at a given time period) the actions taken and outcomes received by every individual in the graph. Formally, for any set C, let $C^{<\mathbb{N}}$ be the set of all finite sequences with range in C. Then define the set H of *possible histories* as follows:

$$H = \{h \in ((A \times O)^{<\mathbb{N}})^{<\mathbb{N}} \ : \ |h_n| = |h_k| \text{ for all } n, k \in \mathbb{N}\}$$

where h_n is the n^{th} coordinate in the history h, i.e. h_n is the sequence of actions and outcomes obtained by some collection of learners at stage n of inquiry. The requirement that $|h_n| = |h_k|$ for all $n, k \in \mathbb{N}$ captures the fact that the size of a group does not change over time. For a network G and a group $G' \subseteq G$, define:

$$H_{G',G} = \{h \in H \ : \ |h_n| = |G'| \text{ for all } n \in \mathbb{N}\}$$

When the network is clear from context, we will simply write $H_{G'}$ to simplify notation. Then H_G is the set of network histories for the entire network, and $H_{N(g)}$ is the set of neighborhood histories for the learner g.

Example 1: Let G be the undirected graph with two vertices joined by an edge. Let $\Omega = \{\omega_1, \omega_2\}$, $A = \{a_1, a_2\}$, $O = \{0, 1\}$, and

$$\begin{aligned} p(1|a_i, \omega_i) &= .7 \text{ for } i = j = 1, 2 \\ p(1|a_i, \omega_j) &= .1 \text{ for } i \neq j \end{aligned}$$

One can imagine A as possible drugs, outcomes 1 and 0 as respectively representing that a patient is cured or not, and ω_i as representing the state of the world in which a_i is the bettertreatment. A possible network history $h \in H_G$ of length two is $\langle\langle\langle a_1, 1\rangle\langle a_1, 0\rangle\rangle, \langle\langle a_1, 0\rangle\langle a_2, 0\rangle\rangle\rangle$, which denotes the history in which (i) one doctor applied treatment a_1 to two successive patients, the first of which was cured but the second of which was not, and (ii) a second doctor applied treatment a_1 to a patient who it failed to cure and then applied treatment a_2 to a second patient who was also uncured.

A *method* (also called a *strategy*) m for an agent is a function that specifies, for any particular history, a probability distribution over possible actions for the next stage. In other words, a method specifies probabilities over the agent's actions given what she knows about her own and her neighbors' past actions and

[1] For technical reasons, we assume outcomes are non-negative, and that the set of outcomes is countable.

outcomes. Of course, an agent may act deterministically simply by placing unit probability on a single action $a \in A$. A *strategic network* is a pair $S = \langle G, M \rangle$ consisting of a network G and a sequence $M = \langle m_g \rangle_{g \in G}$ specifying the strategy employed by each learner, m_g, in the network.

Together, a strategic network $S = \langle G, M \rangle$ and a learning problem $\langle \Omega, A, O, p \rangle$ determine a probability $p^S_\omega(h)$ of any finite history $h \in H_{G'}$ for any group $G' \subseteq G$. To see why, again consider Example 1. Suppose the two learners both employ the followingsimple strategy: if action a_i led to a success 1 on the previous stage, play it again with probability one; if the action failed, play the other action. Then the probability $p^S{}_{omega_1}(h)$ of the history h in Example 1 in state of the world ω_1 is

$$p^S_{\omega_1}(h) = p(1|a_1, \omega_1) \cdot p(0|a_1, \omega_1) \cdot p(0|a_1, \omega_1) \cdot p(0|a_2, \omega_1) = .7 \cdot .3 \cdot .3 \cdot .9 = .1323$$

Notice, however, the same history h might have a different probability if one were to respecify the methods employed by the agents in the network. For example, suppose the agents both employed the rule "switch actions if and only if a success is obtained." Then the history h above would have probability zero (regardless of state of the world), as the first learner continues to play action a_1 after a success.

Because outcomes can be interpreted as utilities, it follows that for any state of the world ω, there is an expected value $E_\omega(a)$ of the action a that is constant throughout time. Hence, in any state of the world ω, there is some collection $A_\omega = \{a \in A \ : \ E_\omega(a) \geq E_\omega(a') \text{ for all } a' \in A\}$ of *optimal* actions that maximize expected utility. Hence, it follows that the event that g plays an optimal action at stage n has a well-defined probability, which we will denote $p^S_\omega(h^A(n, g) \in A_\omega)$. In the next section, we study the limiting behavior of such probabilities in various strategic networks.

Some learning problems are far easier than others; for example, if one action has higher expected utility in every world-state, then there is relatively little for the agents to learn. We are principally interestedin more difficult problems. Say a learning problem is *non-trivial* if no finite history reveals that a given action is optimal with certainty. In other words, a learning problem $\langle \Omega, A, O, p \rangle$ is *non-trivial* if for all strategic networks $S = \langle G, M \rangle$, and all network histories $h \in H_G$, if $p^S_{\omega_1}(h) > 0$ for some $\omega_1 \in \Omega$, then there exists $\omega_2 \in \Omega$ such that $A_{\omega_1} \cap A_{\omega_2} = \emptyset$ and $p^S_{\omega_2}(h) > 0$. Say a learning problem is *difficult* if it is non-trivial, and $1 > p(0|a, \omega) > 0$ for all $\omega \in \Omega$ and all $a \in A$. That is, no action is guaranteed to succeed, and no history determines an optimal action with certainty.

2.1 Four Types of Strategies

Although the number of differing strategies is enormous, we will focus on the behavior of four types of boundedly rational strategies: reinforcement learning (RL), simulated annealing (SA), decreasing ϵ-greedy (ϵG), and what we call, $\delta\epsilon$ methods. We study these strategies for four reasons. First, the first three types of strategies have been employed extensively in economics, computer science, statistics, and many other disciplines in which one is interested in finding the global maximum (or minimum) of a utility (respectively, cost) function. Second, all four strategies are simple and algorithmic: they can easily be simulated on computers and, given enough discipline, performed by human beings. Third, the strategies have desirable asymptotic features in the sense that, in the limit, they find the global maximum of utility functions under robust assumptions. Fourth, some of the strategies have psychological plausibility as learning rules in particular types of problems.

Before introducing the strategies, we need some notation. Denote the cardinality of S by $|S|$ which, if S is a sequence, is also its length. For any two sequences σ and σ' on any set, write $\sigma \preceq \sigma'$ if the former is an initial segment of the latter, If σ is a sequence, then $ran(\sigma)$ denotes its range when the sequence is considered as a function. For example, $ran(\langle m_1, m_2, m_3\rangle)$ is the set $\{m_1, m_2, m_3\}$ and $ran(\langle m_1, m_2, m_1\rangle)$ is the set $\{m_1, m_2\}$. When two sequences σ and σ' differ only by order of their entries (e.g. $\langle 1, 2, 3\rangle$ and $\langle 2, 1, 3\rangle$), write $\sigma \cong \sigma'$.

Reinforcement Learning (RL): Reinforcement learners begin with an initial, positive, real-valued weight for each action. On the first stage of inquiry, the agent chooses an action in proportion to the weights. For example, if there are two actions a_1 and a_2 with weights 3 and 5 respectively, then the agent chooses action a_1 with probability $\frac{3}{3+5}$ and a_2 with probability $\frac{5}{3+5}$. At subsequent stages, the agent then adds the observed outcome for all the actions taken in his neighborhood to the respective weights for the different actions.

Formally, let g be an individual, $w = \langle w_a\rangle_{a\in A}$ be a vector of positive real numbers (the initial weights), and let $h \in H_{N_G g}$ be a history for the individuals in g's neighborhood. Let $r_{a,N(g)}(h)$ represent the total accumulated payoff for action a in g's neighborhood in history h, which includes the initial weight w_a. Define $r_{N(g)}(h) := \sum_{a\in A} r_{a,N(g)}(h)$. An RL strategy is defined by specifying w. For any w, the probability that an action a is played after observed history h is given by:

$$m_r(h)(a) = \frac{r_{a,N(g)}(h)}{r_{N(g)}(h)}$$

Because w_a is positive for all $a \in A$, the chance of playing any action is always positive.

Reinforcement learning strategies are simple and appealing, and further, they have been studied extensively in psychology, economics, and computer science.[2] In economics, for example, reinforcement learning hasbeen used to model how individuals behave in repeated games in which they must learn the strate-

[2] Here, we use the phrase "reinforcement learning" as it is employed in game theory. See Beggs (2005) for a discussion of its asymptotic properties. The phrase "reinforcement learning" has related, but different, meanings in both psychology and machine learning.

gies being employed by other players.[3] Such strategies, therefore, are important, in part, because they plausibly represent how individuals actually select actions given past evidence. Moreover,RL strategies possess certain properties that make them seem rationally motivated: in isolation, an individual employing an RL method will find one or more of the optimal actions in her learning problem (Beggs , 2005).

Decreasing Epsilon Greedy (ϵG): Greedy strategies that choose, on each round, the action that currently appears best may fail to find an optimal action because they do not engage in sufficient experimentation. To address this problem, one can modify a greedy strategy as follows. Suppose $\langle \epsilon_n \rangle_{n \in \mathbb{N}}$ is a sequence of probabilities that approach zero. At stage n, an ϵG-learner plays each action which currently appears best with probability $\frac{1-\epsilon_n}{k}$, where k is the number of actions that currently appear optimal. Such a learner plays every other action with equal probability. Because the experimentation rate ϵ_n approaches zero, it follows that the ϵG learner experiments more frequently early in inquiry, and plays an estimated EU-maximizing action with greater frequency as inquiry progresses. ϵG strategies are attractive because, if ϵ_n is set to decrease at the right rate, then they will play the optimal actions with probability approaching one in all states of the world. Hence, ϵG strategies balance short-term considerations with asymptotic ones. Because they favor actions that appear to have higher EU at any given stage, such strategies approximate demands on short run rationality.

Formally, let each agent begin with an initial estimate of the expected utility of each action, given by the vector $\langle w_a \rangle_{a \in A}$. At each stage, let $\textsc{est}_g(a, h)$ be g's estimate of the expected utility of action a given history h. This is given by w_a if no one in g's neighborhood has yet played a, otherwise it is given by the current *average* payoff to action a from plays in g's neighborhood. Additionally, define the set of actions which currently have the highest estimated utility:

$$A(g,h) := \{a \in A \ : \ \textsc{est}_g(a,h) \geq \textsc{est}_g(a',h) \text{ for all } a' \in A\}$$

An ϵG method is determined by (i) a vector $\langle w_a \rangle_{a \in A}$ of non-negative real numbers representing initial estimates of the expected utility of an action a and (ii) an antitone function $\epsilon : H \to (0,1)$ (i.e $h \preceq h'$ implies $\epsilon(h') \leq \epsilon(h)$) as follows:

$$m_\epsilon(h)(a) := \begin{cases} \frac{1-\epsilon(h)}{|A(g,h)|} & \text{if } a \in A(g,h) \\ \frac{\epsilon(h)}{|A \setminus A(g,h)|} & \text{if } a \notin A(g,h) \end{cases}$$

We will often not specify the vector $\langle w_a \rangle_{a \in A}$ in the definition of an ϵG method; in such cases, assume that $w_a = 0$ for all $a \in A$.

[3] See Roth and Erev (1995) for a discussion of how well reinforcement learning fares empirically as a model of how humans behave in repeated games. The theoretical properties of reinforcement learning in games has been investigated by Argiento, et. al (2009); Beggs (2005); Hopkins (2002); Hopkins and Posch (2005); Huttegger and Skyrms (2008); Skyrms and Pemantle (2004).

Simulated Annealing (SA): In computer science, statistics, and many other fields, SA refers to a collection of techniques for minimizing some cost function.[4] In economics, the cost function might represent monetary cost; in statistical inference, a cost function might measure the degree to which an estimate (e.g., of a population mean or polynomial equation) differs from the actual value of some quantity or equation.

Formally, let $\sigma = \langle \langle w_a \rangle_{a \in A}, \langle q_{a,a'} \rangle_{a,a' \in A}, T \rangle$ be a triple in which (i) $\langle w_a \rangle_{a \in A}$ is a vector of non-negative real numbers representing initial estimates of the expected utility of an action a, (ii) $\langle q_{a,a'} \rangle_{a,a' \in A}$ is a vector of numbers from the open unit interval $(0,1)$ representing initial *transition probabilities*, that is, the probability the method will switch from action a to a' on successive stages of inquiry, and (iii) $T : H \to \mathbb{R}^{\geq 0}$ is a monotone map (i.e. if $h \preceq h'$, then $T(h) \leq T(h')$) from the set of histories to non-negative real numbers which is called a *cooling schedule*. For all $h \in H_{N(g),n+1}$ and $a \in A$, define:

$$s(h,a) = T(h) \cdot \max\{0, \text{EST}_g(h^A(n,g), h \upharpoonright n) - \text{EST}_g(a, h \upharpoonright n)\}$$

Here, s stands for "switch." Then the SA method determined by $\sigma = \langle \langle w_a \rangle_{a \in A}, \langle q_{a,a'} \rangle_{a,a' \in A}, T \rangle$ is defined as follows:

$$\begin{aligned} m_\sigma(\langle - \rangle)(a) &= \frac{1}{|A|} \\ m_\sigma(h)(a) &= \begin{cases} q_{a',a} \cdot e^{-s(a,h)} & \text{if } a \neq a' = h^A(n,g) \\ 1 - \sum_{a'' \in A \setminus \{a'\}} q_{a',a''} \cdot e^{-s(a'',h)} & \text{if } a = a' = h^A(n,g) \end{cases} \end{aligned}$$

Like ϵG methods, we will often not explicitly specify the vector $\langle w_a \rangle_{a \in A}$ in the definition of an SA method; in such cases, assume that $w_a = 0$ for all $a \in A$.

In our model of learning, SA strategies are similar to ϵG strategies. SA strategies may experiment frequently with differing actions at the outset of inquiry, but they have a "cooling schedule" that ensures that the rate of experimentation drops as inquiry progresses. SA strategies and ϵG strategies, however, differ in an important sense. SA strategies specify the probability of *switching* from one action to another; the probability of switching is higher if the switch involves moving to an action with higher EU, and lower if the switch appears to be costly. Importantly, however, SA strategies do not "default" to playing the action with the highest EU, but rather, the chance of playing any action depends crucially on the previous action taken.

[4] For an overview of SA methods and applications see Bertsimas and Tsitsiklis (1993), which considers SA methods in **non** "noisy" learning problems in which the action space is finite. Bertsimas and Tsitsiklis (1993) provides references for those interested in SA methods in infinite action spaces. For an overview of SA methods in the presence of "noise", see Branke et al. (2008). Many of the SA algorithms for learning in noisy environments assume that one can draw finite samples of any size at successive stages of inquiry. As this is not permitted in our model (because agents can choose exactly one action), what we call SA strategies are closer to the original SA methods for learning in non-noisy environments.

The fourth class of methods that we consider consists of intuitively plausible algorithms, though they have not been studied prior to this paper.

Delta-Epsilon ($\delta\epsilon$): $\delta\epsilon$ strategies are generalizations of ϵG strategies. Like ϵG strategies, $\delta\epsilon$ methods play the action which has performed best most frequently, and experiment with some probability ϵ_n on the n^{th} round, where ϵ_n decreases over time. The difference between the two types of strategies is that each $\delta\epsilon$ method has some "favorite" action a^* that it favors in early rounds. Hence, there is some sequence of (non-increasing) probabilities δ_n with which $\delta\epsilon$ methods play the favorite action a^* on the n^{th} round. The currently best actions are, therefore, played with probability $1 - \delta_n - \epsilon_n$ on the n^{th} stage of inquiry.

Formally, let $a^* \in A$, and $\delta, \epsilon : H \to [0, 1)$ be antitone maps such that $\delta(h) + \epsilon(h) \leq 1$. Then the $\delta\epsilon$ method determined by a quadruple $\langle \langle w_a \rangle_{a \in A}, \delta, \epsilon, a^* \rangle$, is defined as follows:

$$m_{\delta,\epsilon,a^*}(h)(a) := \begin{cases} \frac{1-(\epsilon(h)+\delta(h))}{|A(g,h)|} & \text{if } a \neq a^* \text{ and } a \in A(g,h) \\ \frac{\epsilon(h)}{|A \setminus A(g,h)|} & \text{if } a \neq a^* \text{ and } a \notin A(g,h) \\ \delta(h) + \frac{1-(\epsilon(h)+\delta(h))}{|A(g,h)|} & \text{if } a = a^* \text{ and } a \in A(g,h) \\ \delta(h) + \frac{\epsilon(h)}{|A(g,h)|} & \text{if } a = a^* \text{ and } a \notin A(g,h) \end{cases}$$

Every ϵG methods is a $\delta\epsilon$ method if one sets δ to be the constant function 0. Like SA methods, we will often not specify the vector $\langle w_a \rangle_{a \in A}$ in the definition of a $\delta\epsilon$ method; in such cases, assume that $w_a = 0$ for all $a \in A$.

$\delta\epsilon$ methods capture a plausible feature of human learning: individuals may have a bias, perhaps unconscious, toward a particular option (e.g., a type of technology) for whatever reason. The δ parameter specifies the degree to which they have this bias. Individuals will occasionally forgo the apparently better option in order to experiment with their particular favorite technology. The ϵ parameters, in contrast, specify a learner's tendency to "experiment" with entirely unfamiliar actions.

3 Individual versus Group Rationality

One of the predominant ways of evaluating these various boundedly rational strategies is by comparing their asymptotic properties. Which of these rules will, in the limit, converge to playing one of the optimal actions? One of the central claims of this section is that there are at least four different ways one might make this precise, and that whether a learning rule converges depends on how exactly one defines convergence.

Our four ways of characterizing long run convergence differ on two dimensions. First, one can consider the performance of either only a single strategy or a set of strategies. Second, one can consider the performance of a strategy (or strategies) when they are isolated from other individuals or when they are in groups with other strategies. These two dimensions yield four distinct notions of convergence, each satisfied by different (sets of) strategies.

We first consider the most basic case: a single agent playing in the absence of any others. Let $S_m = \langle G = \{g\}, \langle m \rangle \rangle$ be the *isolated network* with exactly one learner employing the strategy m.

Definition 1. A strategy m is *isolation consistent* (IC) if for all $\omega \in \Omega$:

$$\lim_{n \to \infty} p_{\omega}^{S_m}(h^A(n, g) \in A_{\omega}) = 1$$

IC requires that a single learner employing strategy m in isolation converges, with probability one, to an optimal action. IC is the weakest criterion for individual epistemic rationality that we consider. It is well-known that, regardless of the difficulty of the learning problem, some ϵG and SA-strategies are IC. Similarly, some $\delta\epsilon$ strategies are IC. Under mild assumptions, all RL methods can also be shown to be IC

Theorem 1. Some SA, ϵG, and $\delta\epsilon$ strategies are always (i.e. in every learning problem) IC. If $\langle \Omega, A, O, p \rangle$ is a learning problem in which there are constants $k_2 > k_1 > 0$ such that $p(o|a, \omega) = 0$ if $o \notin [k_1, k_2]$, then all RL methods are IC.

The second case is convergence of an individual learner in a network of other, not necessarily similar, learners. This notion requires thatthe learner converge to playing an optimal action in any arbitrary network. Let $S = \langle G, M \rangle$ be a strategic network, $g \in G$, and m be a method. Write $S_{g,m}$ for the strategic network obtained from S by replacing g's method m_g with the alternative method m.

Definition 2. A strategy m is *universally consistent* (UC) if for any strategic network $S = \langle G, M \rangle$ and any $g \in G$:

$$\lim_{n \to \infty} p_{\omega}^{S_{g,m}}(h^A(n, g) \in A_{\omega}) = 1$$

UC strategies always exist, regardless of the difficulty of the learning problem, since one can simply employ an IC strategy and ignore one's neighbors. Furthermore, by definition, any UC strategy is IC, since the isolated network is a strategic network. The converse, however, is false in general:

Theorem 2. In all difficult learning problems, there are RL, SA, ϵG, and $\delta\epsilon$ strategies that are IC but not UC. In addition, if $\langle \Omega, A, O, p \rangle$ is a non-trivial learning problem in which there are constants $k_2 > k_1 > 0$ such that $p(o|a, \omega) = 0$ if $o \notin [k_1, k_2]$, then all RL methods are IC but not UC.

The general result that not all IC strategies are UC is unsurprising given the generality of the definitions of strategies, actions, and worlds. One can simply define a pathological strategy that behaves well in isolation, but chooses suboptimal actions when in networks. The important feature of the above theorem is that *plausible* strategies, like some RL and SA strategies, are IC but fail to be UC. The reason for such failure is rather easy to explain. Consider SA strategies first. Recall the "cooling schedule" of a SA strategy specifies the probability with which a learner will choose some seemingly inferior action. In SA strategies, the cooling schedule must be finely tuned so as to ensure that learners experiment (i) sufficiently often so as to ensure they find an optimal action, and (ii) sufficiently infrequently so as to ensure they play an optimal action with probability approaching one in the long-run. Such fine-tuning is very fragile: in large networks, learners might acquire information too quickly and fail to experiment enough to find an optimal action. Similar remarks apply to ϵG and $\delta\epsilon$ methods.

RL strategies fail to be UC for a different reason. At each stage of inquiry, RL learners calculate the *total* utility that has been obtained by playing some action in the past, where the totals include the utilities obtained by all of one's neighbors. If a reinforcement learner is surrounded by enough neighbors who are choosing inferior actions, then the cumulative utility obtained by plays of suboptimal actions might be higher than that of optimal actions. Thus, a RL method might converge to playing a suboptimal action with probability one in the limit.

This argument that RL-strategies fail to be UC, however, trades on the existence of learners with no interest in finding optimal actions. It seems unfair to require a learner to find optimal actions when his or her neighbors are intent on deceiving him or her. When only RL methods are present in a finite network, then Theorem 5 shows that, under most assumptions, every learner is guaranteed to find optimal actions. That is, RL methods work well together as a *group*.

The third and fourth notions of convergence focus on the behavior of a group of strategies, either in "isolation" (i.e., with no other methods in the network) or in a larger network. One natural idea is to impose no constraints on the network in which the group is situated. Such an idea is, in our view, misguided. Say a network is *connected* if there is a finite sequence of edges between any two learners. Consider now individuals in unconnected networks: these learners never communicate at all, and so it makes little sense to think of such networks as social groups. Moreover, there are few interesting theoretical connections that can be drawn when one requires convergence of a "group" even in unconnected networks. We thus restrict our attention to connected networks, where far more interesting relationships between group and individual rationality emerge. To see why, we first introduce some definitions.

Definition 3 (N-Network)**.** Let $S = \langle G, M \rangle$ be a strategic network, and let N be a sequence of methods of the same length as M. Then S is called a *N-network* if $N \cong M$.

Definition 4 (Group Isolation Consistency)**.** Let N be a sequence of methods. Then N is *group isolation consistent* (GIC) if for all connected N-networks $S = \langle G, M \rangle$, all $g \in G$, and all $\omega \in \Omega$:

$$\lim_{n \to \infty} p_\omega^S(h^A(n, g) \in A_\omega) = 1$$

Definition 5 (Group Universal Consistency)**.** Let N be a sequence of methods. Then N is *group universally consistent* (GUC) if for all networks $S = \langle G, M \rangle$, if $S' = \langle G', M' \rangle$ is a connected N-subnetwork of S, then for all $g \in G'$ and all $\omega \in \Omega$:

$$\lim_{n \to \infty} p^S_\omega(h^A(n,g) \in A_\omega) = 1$$

Characterizing group rationality in terms of sequences of methods is important because doing so allows one to characterize exactly how many of a given strategy are employed in a network. However, in many circumstances, one is only interested in the underlying set of methods used in a network. To this end, define:

Definition 6 (Group Universal/Isolation Consistency (Sets))**.** Let $\mathcal{M}$ be a set of methods. Then $\mathcal{M}$ is GIC (respectively, GUC) if for for every sequence of methods M such that $ran(M) = \mathcal{M}$, the sequence M is GIC (respectively, GUC).

So a set $\mathcal{M}$ is GIC if, for all connected networks that have only methods in $\mathcal{M}$ and each method in $\mathcal{M}$ is occurs at least once in the network, each learner in the network converges to playing optimal actions. A set $\mathcal{M}$ is GUC if, for all networks in which each method in $\mathcal{M}$ is represented at least once and those employing $\mathcal{M}$ are connected by paths of learners using $\mathcal{M}$, each agent in the subnetwork employing $\mathcal{M}$ converges.

The names encode a deliberate analogy: GIC stands to GUC as IC stands to UC. Just as an IC method is only required to converge when no other methods are present, so a GIC sequence of methods is only required to find optimal actions when no other methods are present in the network. And just a UC method must converge regardless of the other methods around it, a GUC sequence of methods must converge to optimal actions regardless of other methods in the network.

Clearly, any sequence (respectively set) of UC strategies M is both GUC and GIC, since the UC methods are just those that converge regardless of those around them. It thus follows immediately that GUC and GIC groups exist. Interestingly, however, GUC sequences of methods need not contain *any* strategy that is even IC (let alone UC).

Theorem 3. *In difficult learning problems, there are sequences and sets of $\delta\epsilon$ methods M such that M is GUC, but no m in M is IC.*

Still more surprising is the fact that there are IC methods that form groups that fail to be GIC:

Theorem 4. *In difficult learning problems, there are sequences M (respectively sets) of $\delta\epsilon$ methods that are not GIC, but such that every coordinate (respectively element) m of M is IC. In fact, M can even be a constant sequence consisting of one method repeated some finite number of times. Similarly for SA and ϵG methods.*

Finally, because all ϵG strategies are $\delta\epsilon$ strategies, we obtain the following corollary that shows that, depending on the balance between dogmatism and tendency to experiment, a method may behave in any number of ways when employed in isolation and when in networks.

Corollary 1. In difficult learning problems, there exist different sequences (respectively sets) M of $\delta\epsilon$ methods such that

1. Each member (respectively, coordinate) of M is IC but not UC; or

2. Each member (respectively, coordinate) of M is IC, but M is not GIC; or

3. M is GUC, but no member (respectively, coordinate) of M is IC.

The only conceptual relationship not discussed in the above corollary is the relationship between GUC and GIC. It is clear that if M is GUC, then it is also GIC. The converse is false in general, and RL methods provide an especially strong counterexample:

Theorem 5. Suppose $\langle \Omega, A, O, p \rangle$ is a non-trivial learning problem in which there are constants $k_2 > k_1 > 0$ such that $p(o|a, \omega) = 0$ if $o \notin [k_1, k_2]$. Then every finite sequence of RL methods is GIC, but no such sequence is GUC.

4 Discussion

We believe that the most important part of our results is the demonstration that judgments of individual rationality and group rationality need not coincide. Rational (by one standard) individuals can form an irrational group, and rational groups can be composed of irrational individuals. Recent interest in the "wisdom of crowds" has already suggested that groups might outperform individual members, and our analyses demonstrate a different way in which the group can be wiser than the individual. Conversely, the popular notion of "groupthink," in which a group of intelligent individuals converge prematurely on an incorrect conclusion, is one instance of our more general finding that certain types of strategies succeed in isolation but fail when collected into a group. These formal results thus highlight the importance of clarity when one argues that a particular method is "rational" or "intelligent": much can depend on how that term is specified, regardless of whether one is focused on individuals or groups.

These analyses are, however, only a first step in understanding the connections between individual and group rationality in learning. There are a variety of methods which satisfy none of the conditions specified above, but are nonetheless convergent in a particular setting. Bala and Goyal (2008) provide one such illustration. We also have focused on reinforcement learning as it is understood in the game theory literature; the related-but-different RL methods in psychology and machine learning presumably exhibit different convergence properties. Additional investigation into more limited notions of group rationality than the ones offered here are likely to illustrate the virtues of other boundedly rational learning rules, and may potentially reveal further conceptual distinctions.

In addition to considering other methods, these analyses should be extended to different formal frameworks for representing inquiry. We have focused on the case of multi-armed bandit problems, but these are clearly only one way to model learning and inquiry. It is unknown how our formal results translate to different settings. One natural connection is to consider learning in competitive game-theoretic contexts. Theorems about the performance in multi-armed bandits are often used to help understand how these rules perform in games, and so our convergence results should be extended to these domains.

There are also a range of natural applications for this framework. As already suggested, understanding how various boundedly rational strategies perform in a multi-armed bandit problem can have important implications to a variety of different economic phenomena, and in particular, on models of the influence that social factors can have on various strategies for learning in multi-armed bandits. This framework also provides a natural representation of many cases of inquiry by a scientific community.

More generally, this investigation provides crucial groundwork for understanding the difference between judgments of convergence of various types by boundedly rational strategies. It thus provides a means by which one can better understand the behavior of such methods in isolation and in groups.

References

[1] Argiento, R., Pemantle, R., Skyrms, B., and Volkov, S. (2009) "Learning to Signal: Analysis of a Micro-Level Reinforcement Model," *Stochastic Processes and Their Applications* 119(2), 373390.

[2] Bala, V. and Goyal, S. (2008) "Learning in networks." To appear in, *Handbook of Mathematical Economics.* Eds. J. Benhabib, A. Bisin and M.O. Jackson.

[3] Beggs, A. (2005) "On the Convergence of Reinforcement Learning." *Journal of Economic Theory.* 122: 1-36.

[4] Berry, D. A., and Fristedt, B. (1985) *Bandit Problems: Sequential Allocation of Experiments*, Chapman and Hall.

[5] Bertsimas, D. and Tsitsiklis, J. (1993) "Simulated Annealing." 8 (1): 10-15.

[6] Bolton, P., and Harris, C. (1999) "Strategic Experimentation," *Econometrica* 67(2), 349374.

[7] Branke, J., Meisel S., and Schmidt C. (2008) "Simulated annealing in the Presence of Noise." *Journal of Heuristics.* 14 (6) : 627-654.

[8] Hong, L. and Page, S. (2001) "Problem Solving by Heterogeneous Agents." *Journal of Economic Theory.* 97 (1): 123-163.

[9] Hong, L. and Page, S. (2004) "Groups of Diverse Problem solvers Can Outperform Groups of High-Ability Problem Solvers." *Proceedings of the National Academy of Sciences.* 101 (46): 16385 – 16389.

[10] Hopkins, E. (2002 "Two Competing Models of How People Learn in Games," *Econometrica* 70(6), 21412166.

[11] Hopkins, E. and Posch, M. (2005) “Attainability of Boundary Points under Reinforcement Learning,” *Games and Economic Behavior* 53(1), 110125.

[12] Huttegger, S. and Skyrms, B. (2008) “Emergence of Information Transfer by Inductive Learning,” *Studia Logica* 89, 237256.

[13] Keller, G., Rady, S., and Cripps, M (2005) “Strategic Experimentation with Exponential Bandits,” *Econometrica* 73(1), 3968.

[14] Kitcher, P. (1990) “The Division of Cognitive Labor.” *Journal of Philosophy.* 87 (1): 5-22.

[15] Mayo-Wilson, C., Zollman, K., and Danks, D. “Wisdom of the Crowds vs. Groupthink: Connections between Individual and Group Epistemology.” Carnegie Mellon University, Department of Philosophy. Technical Report No. 187.

[16] Roth, A. and Erev, I. (1995) “Learning in Extensive-Form Games: Experimental Data and Simple Dynamic Models in the Intermediate Term.” *Games and Economic Behavior.* 8: 164 – 212.

[17] Shephard, R. N. (1957) “Stimulus and Response Generalization: A stochastic Model Relating Generalization to Distance in Psychological Space.” *Psychometrika,* 22: 325-345.

[18] Skyrms, B. and Pemantle, R. (2004) “Network Formation by Reinforcement Learning: The Long and Medium Run,” *Mathematical Social Sciences* 48, 315327.

[19] Strevens, M. (2003) “The Role of the Priority Rule in Science,” *Journal of Philosophy.* Vol. 100.

[20] Weisberg, M. and Muldoon, R. (2009a) “Epistemic Landscapes and the Division of Cognitive Labor.” Forthcoming in *Philosophy of Science.*

[21] Zollman, K. (2007) *Network Epistemology.* Ph.D. Dissertation. University of California, Irvine. Department of Logic and Philosophy of Science.

[22] Zollman, K. (2009) “The Epistemic Benefit of Transient Diversity.” *Erkenntnis.* 72(1): 17-35.

[23] Zollman, K. (2010) “The Communication Structure of Epistemic Communities.” *Philosophy of Science.* 74(5): 574-587.

[24] Zollman, K. (2010) “Social Structure and the Effects of Conformity.” *Synthese.* 172(3): 317-340.

5 Formal Definitions

5.1 Notational Conventions

In the following appendix, 2^S will denote the power set of S. Let $S^{<\mathbb{N}}$ denote all finite sequences over S, and let $S^{\mathbb{N}}$ be the set of all infinite sequences over S. We will use $\langle - \rangle$ to denote the empty sequence. Write σ_n to denote the n^{th} coordinate of σ, and let $\sigma \upharpoonright n$ to denote the initial segment of the sequence σ of length n; we stipulate that $\sigma \upharpoonright n = \sigma$ if n is greater than the length of σ.

If σ is a subsequence of σ', then write $\sigma \sqsubseteq \sigma'$, and write $\sigma \sqsubset \sigma'$ if the subsequence is strict. If $\sigma \preceq \sigma'$, then $\sigma \sqsubseteq \sigma'$, but not vice versa. For example $\langle 1, 2 \rangle \sqsubseteq \langle 3, 1, 5, 2 \rangle$, but the former is not an initial segment of the latter.

Given a network G and a group $G' \subseteq G$, for any $n \in \mathbb{N}$ we let $H_{G',n}$ denote sequences of $H_{G'}$ of length n. Because (i) the set of outcomes, actions, and individuals G are all at most countable, and (ii) the set of finite sequences over countable sets is countable, we obtain:

Lemma 1. *H, $H_{G'}$, H_G, H_n, $H_{G',n}$, and $H_{G,n}$ are countable.*

Write $h^A(n, g)$ to denote the action taken by g on the n^{th} stage of inquiry, and $h^O(n, g)$ to denote the outcome obtained. If $h \in H_{G'}$ has length 1 (i.e. h represents the actions/outcomes of group G' at the first stage of inquiry), write $h^A(g)$ and $h^O(g)$ to denote the initial action taken and outcome obtained by the learner $g \in G'$. Similarly, if $h \in H_{G'}$ is such that $|h_n| = 1$ for all $n \leq |h|$ (i.e. h represents the history of exactly one learner), write $h^A(n)$ and $h^O(n)$ to denote the action and outcome respectively taken/obtained at stage n.

For a network G and a group $G' \subseteq G$, a *complete group history* for G' is an infinite sequence $\langle h_n \rangle_{n \in \mathbb{N}}$ of (finite) group histories such that $h_n \in H_{G',n}$ and $h_n \prec h_k$ for all $n < k$. Denote the set of complete group histories for G' by $\overline{H}_{G'}$. Define complete individual histories $\overline{H}_g$, and complete network histories $\overline{H}_G$ similarly.

5.2 Measurable Spaces of Histories

Let G be a network, $G' \subseteq G$, and define $\mathcal{H}_G = \cup_{G' \subseteq G} \overline{H}_{G'}$ to be the set of all complete histories for all groups G' in the network G. For any group history $h \in H_{G',n}$ of length n, define:

$$[h] = \{\overline{h} \in \overline{H}_{G'} \ : \ \overline{h}_n = h\}$$

In other words, $[h]$ is the set of complete group histories extending the finite group history h. It is easy to see that the sets $[h]$ form a basis for a topology. Let τ_G be the topology in which open sets are unions of sets of the form $[h]$, where $G' \subseteq G$ and $h \in H_{G'}$. Let $\mathcal{F}_G = \sigma(\tau_G)$ be the σ-closure of τ_G, i.e. $\mathcal{F}_G$ is the Borel algebra generated by τ_G. Then $\langle \mathcal{H}_G, \mathcal{F}_G \rangle$ is a measurable space.

Lemma 2. The following sets are measurable (i.e. events) in $\langle \mathcal{H}_G, \mathcal{F}_G \rangle$:

1. $[h^A(n,g) = a] := \{\overline{h} \in \mathcal{H}_G \ : \ \overline{h}_n^A(n,g) = a\}$ for fixed $a \in A$ and $g \in G$

2. $[G'$ plays A' infinitely often$] := \{\overline{h} \in \mathcal{H}_G \ : \forall n \in \mathbb{N} \exists k \geq n \exists g \in G'(\overline{h}_k^A(k,g) \in A')\}$ for fixed $A' \subseteq A$ and $G' \subseteq G$

3. $[\lim_{n\to\infty} \text{EST}_g(a, h_n) = r \ ] := \{\overline{h} \in \mathcal{H}_G \ : \ \lim_{n\to\infty} \text{EST}_g(a, \overline{h}_n) = r\}$ for fixed $a \in A$, $g \in G$, and $r \in \mathbb{R}$.

4. $[\lim_{n\to\infty} m(h_n)(A_\omega) = 1] = \{\overline{h} \in \mathcal{H}_G \ : \ \lim_{n\to\infty} m(\overline{h}_n)(A_\omega) = 1\}$, where ω is a fixed state of the world, and m is a fixed method.

There is another measurable space that will be employed in several lemmas and theorems below. For a fixed $a \in A$, let H_a be the set of individual histories such that only the action a is played by the individual, i.e.

$$H_a := \{h \in H \ : \ |h_n| = 1 \text{ and } h^A(n) = a \text{ for all } n \leq |h|\}$$

Similarly, define $\mathcal{H}_a = \overline{H}_a$, τ_a, and $\mathcal{F}_a$ to be respectively the sets (i) of complete individual histories in which only action a is played, (ii) the topology generated by the basic open sets $[h_a]$, where $h_a \in H_a$, and (iii) the σ-algebra generated by τ_a. Then just as in Lemma 2, one obtains that the following sets are measurable in $\langle \mathcal{H}_a, \mathcal{F}_a \rangle$:

Lemma 3. The following sets are measurable (i.e. events) in $\langle \mathcal{H}_a, \mathcal{F}_a \rangle$:

1. $[h^O(n) \in O'] := \{\overline{h} \in \mathcal{H}_a \ : \ \overline{h}_n^O(n,g) \in O'\}$ for fixed $O' \subseteq O$.

2. $[\lim_{n\to\infty} \text{EST}(a, h_n) = r] := \{\overline{h} \in \mathcal{H}_a \ : \ \lim_{n\to\infty} \text{EST}(a, \overline{h}_n) = r\}$ for fixed $a \in A$, and $r \in \mathbb{R}$.

Notice the parameters G' and g are dropped from the above events because there is, by definition, only one learner in each of the histories in H_a.

5.3 Probabilities of Histories and Complete Histories

Given a strategic network $S = \langle G, M \rangle$, a collection of learners $G' \subseteq G$, and a state of the world ω, one can define, by recursion on the length of a history $h \in H_{G'}$, the probability $p^S_{G',\omega,n}(h)$ that each learner $g \in G'$ performs the action and obtains the outcomes specified by the history h of length n.

$$\begin{aligned} p^S_{G',\omega,0}(\langle - \rangle) &= 1 \\ p^S_{G',\omega,n+1}(h) &:= p^S_{G',\omega,n}(h \restriction n) \cdot \ \Pi_{g \in G'} \ m_g(h \restriction n)(h^A(n+1,g)) \\ &\quad \cdot p(h^O(n+1,g)|h^A(n+1,g),\omega) \end{aligned}$$

Given a strategic network $S = \langle G, M \rangle$ and a state of the world $\omega \in \Omega$, one can define p^S_ω to be the unique, countably additive probability measure on $\langle \mathcal{H}_G, \mathcal{F}_G \rangle$ such that $p^S_\omega([h]) = p^S_{G',\omega,n}(h)$ for all $h \in H_{G',n}$ and all $n \in \mathbb{N}$. The measure p^S_ω exists and is unique by standard measure theoretic constructions. Details are available in Mayo-Wilson, Zollman, and Danks (2010). By abuse of notation, we do not distinguish between $p^S_{G',\omega,n}(h)$ and its extension $p^S_\omega([h])$ in the ensuing proofs, as the expressions denote the same quantities.

For technical reasons, it will also be helpful to specify a probability measure on the space $\langle \mathcal{H}_a, \mathcal{F}_a \rangle$. Let m_a be the method that always plays action a, and $S_a = \langle \{g\}, \langle m_a \rangle \rangle$ be a network with one agent who employs m_a. For each $\omega \in \Omega$, define $p^a_\omega = p^{S_a}_\omega$. It immediately follows that $p^a_\omega([h]) = \Pi_{n \leq |h|}\ p(h^O(n,g)|a, \omega)$ for all $h \in \mathcal{H}_a$.

5.4 Basic Lemmas

Lemma 4. $p^a_\omega(\lim_{n\to\infty} \textsc{est}_g(a, h_n) = E_\omega[a]) = 1$

Proof: Let $X_n : \mathcal{H}_a \to \mathbb{R}$ be the random variable $\overline{h} \mapsto \overline{h}^O_n(n,g)$, and apply the strong law of large numbers to $\langle X_n \rangle_{n \in \mathbb{N}}$.

Lemma 5. Let $S = \langle G, M \rangle$ be any strategic network, $G' \subseteq G$, $a \in A$, $h_a \in H_a$, and $h \in H_{G'}$. Suppose $h_a \sqsubseteq h$. Then $p^a_\omega([h_a]) \geq p^S_\omega([h])$ for all $\omega \in \Omega$.

Proof: Recall that both $p^a_\omega([h_a])$ and $p^S_\omega([h])$ are defined to be products of numbers less than or equal to one. Because $h_a \sqsubseteq h$, every term in the product $p^a_\omega([h_a])$ appears in the product $p^S_\omega([h])$. Hence, $p^a_\omega([h_a]) \geq p^S_\omega([h])$.

Lemma 6. Let $S = \langle G, M \rangle$ be any strategic network, $G' \subseteq G$, and $a \in A$, $h_a \in H_a$, and $h \in H_{G'}$. Suppose $E \in \mathcal{F}_G$ and $E_a \in \mathcal{F}_a$ are such that

1. For every $\overline{h}_a \in E_a$, there is $\overline{h} \in E$ such that $\overline{h}_a \sqsubseteq \overline{h}$, and
2. For every $\overline{h} \in E$, there is $\overline{h}_a \in E_a$ such that $\overline{h}_a \sqsubseteq \overline{h}$.

Then $p^a_\omega(E_a) \geq p^S_\omega(E)$.

Proof: Follows from the previous lemma and the constructions of p^S_ω and p^a_ω. See Mayo-Wilson, Zollman, and Danks (2010) for details.

Lemma 7. Let $S = \langle G, M \rangle$ be any strategic network, $g \in G$, and $a \in A$. Then for all $\omega \in \Omega$:

$$p^S_\omega(\lim_{n\to\infty} \textsc{est}_g(a, h_n) = E_\omega[a] \mid N_G(g) \text{ plays } a \text{ infinitely often}) = 1$$

so long as $p^S_\omega(N_G(g) \text{ plays } a \text{ infinitely often}) > 0$.

Proof: Fix $g \in G$ and let

$$E_g := [\lim_{n\to\infty} \textsc{est}_g(a, h_n) \neq E_\omega[a]] \cap [N_G(g) \text{ plays } a \text{ infinitely often}].$$

For all $\overline{h} \in E_g$, let $\overline{h}_a$ be the sequence consisting of all of the coordinates of $\overline{h}$ in which the action a is played; because a is played infinitely often in $\overline{h}$ (by definition of E_g), the sequence $\overline{h}_a$ is infinitely long. Define:

$$E_{g,a} := \{\overline{h}_a \in \overline{H}_a \ : \ \overline{h} \in E_g\}.$$

Because the limit of estimates of the EU of a is wrong in every $\overline{h} \in E_g$, it is likewise wrong in every $\overline{h}_a \in E_{g,a}$. By Lemma 4, it follows that $p^a_\omega(E_{g,a}) = 0$. By Lemma 6, it follows that $p^S_\omega(E_g) \leq p^a_\omega(E_{g,a}) = 0$.

Lemma 8. Let $S = \langle G, M \rangle$ be a strategic network, $g \in G$, $\omega \in \Omega$. Suppose that $p^S_\omega(\lim_{n\to\infty} m_g(h_n)(A_\omega) = 1) = 1$. Then $\lim_{n\to\infty} p^S_\omega(h^A(n,g) \in A_\omega) = 1$.

Proof: Let $\epsilon \in \mathbb{Q} \cap (0,1)$, and let $n \in \mathbb{N}$. Define:

$$\begin{aligned}
F_{n,\epsilon} &:= \{h \in H_{N(g),n} \ : \ m_g(h)(A_\omega) > 1-\epsilon\} \\
\overline{F}_{n,\epsilon} &:= \{\overline{h} \in \overline{H}_{N(g)} \ : \ m_g(\overline{h}_n)(A_\omega) > 1-\epsilon\} \\
\overline{E}_{n,\epsilon} &:= \{\overline{h} \in \overline{H}_{N(g)} \ : \ m_g(\overline{h}_k)(A_\omega) > 1-\epsilon \text{ for all } k \geq n\}
\end{aligned}$$

Clearly, $\overline{E}_{n,\epsilon} \subseteq \overline{F}_{n,\epsilon}$. It follows that:

$$\begin{aligned}
p^S_\omega(h^A(n+1,g) \in A_\omega) &= \sum_{h \in H_{N(g),n}} p^S_\omega(h) \cdot m_g(h)(A_\omega) \\
&= \sum_{h \in F_{n,\epsilon}} p^S_\omega(h) \cdot m_g(h)(A_\omega) + \sum_{h \in H_{N(g),n} \setminus F_{n,\epsilon}} p^S_\omega(h) \cdot m_g(h)(A_\omega) \\
&\geq \sum_{h \in F_{n,\epsilon}} p^S_\omega(h) \cdot m_g(h)(A_\omega) \\
&\geq \sum_{h \in F_{n,\epsilon}} p^S_\omega(h) \cdot (1-\epsilon) \\
&= p^S_\omega(\overline{F}_{n,\epsilon}) \cdot (1-\epsilon) \\
&\geq p^S_\omega(\overline{E}_{n,\epsilon}) \cdot (1-\epsilon)
\end{aligned}$$

Notice that $\overline{E}_{1,\epsilon} \subseteq \overline{E}_{2,\epsilon} \subseteq \ldots$, and so it follows that $\lim_{n\to\infty} p^S_\omega(\overline{E}_{n,\epsilon}) = p^S_\omega(\cup_{n\in\mathbb{N}} \overline{E}_{n,\epsilon})$. Now by assumption $p^S_\omega(\lim_{n\to\infty} m_g(h_n)(A_\omega) = 1) = 1$. Furthermore, $[\lim_{n\to\infty} m_g(h_n)(A_\omega) = 1] = \cap_{\delta \in \mathbb{Q}\cap(0,1)} \cup_{n\in\mathbb{N}} \overline{E}_{n,\delta}$. So it follows that

$$\begin{aligned}
1 &= p^S_\omega(\lim_{n\to\infty} m_g(h_n)(A_\omega) = 1) \\
&= p^S_\omega(\cap_{\delta \in \mathbb{Q}\cap(0,1)} \cup_{n\in\mathbb{N}} \overline{E}_{n,\delta}) \\
&\leq p^S_\omega(\cup_{n\in\mathbb{N}} \overline{E}_{n,\epsilon}) \\
&= \lim_{n\to\infty} p^S_\omega(\overline{E}_{n,\epsilon}) \\
&\leq \frac{1}{1-\epsilon} \cdot \lim_{n\to\infty} p^S_\omega(h^A(n+1,g) \in A_\omega) \text{ by the argument above}
\end{aligned}$$

As ϵ was chosen arbitrarily from the $\mathbb{Q} \cap (0,1)$, the result follows.

Lemma 9. Let $S = \langle G, M \rangle$ be a strategic network, $g \in G$, $A' \subseteq A$, and $\omega \in \Omega$. If $\lim_{n\to\infty} p^S_\omega(h^A(n,g) \in A') = 1$, then $p^S_\omega(g \text{ plays } A' \text{ infinitely often}) = 1$.

Proof: By contraposition. Suppose $p^S_\omega(g$ does **not** play A' infinitely often $)$ is positive. By definition, $[g$ does **not** play A' infinitely often $] = \cup_{n \in \mathbb{N}} \cap_{k \geq n} [h^A(k, g) \notin A']$, and so (by countable additivity), there is some $j \in \mathbb{N}$ such that $p^S_\omega(\cap_{k \geq j}[h^A(k, g) \notin A']) = r > 0$. It follows that $p^S_\omega(h^A(k, g) \in A') \leq 1 - r$ for all $k \geq j$. Hence, $\lim_{n \to \infty} p^S_\omega(h^A(n, g) \in A') \leq 1 - r < 1$.

Corollary 2. Let $S = \langle G, M \rangle$ be a strategic network, $g \in G$, and $\omega \in \Omega$. Suppose that there is some $n \in \mathbb{N}$ such that $p^S_\omega(\bigcap_{k>n}[h^A(k, g) \notin A_\omega]) > 0$. Then $\lim_{n \to \infty} p^S_\omega(h^A(n, g) \in A_\omega) < 1$.

5.5 Proofs of Major Propositions

In the following two propositions, let $\epsilon : H \to \mathbb{R}^{\geq 0}$ be the function $\epsilon(h) = \frac{1}{|h|^{|h_1|}}$, and let m_ϵ be the ϵG method determined by ϵ.

Proposition 1. In all learning problems, m_ϵ is IC.

Proof: Consider the isolated network $S_{m_\epsilon} = \langle \{g\}, \langle m_\epsilon \rangle \rangle$. Let $a \in A$ and $n \in \mathbb{N}$. Define $E_n = [h^A(n) = a]$. Then by definition of the method m_ϵ, every action on stage n is always played with probability at least $\frac{1}{|A| \cdot n}$ (recall that A is finite, and so this is a real number). It follows that: $p^{S_{m_\epsilon}}_\omega(E_n \mid \cap_{k<n} E^c_k) \geq \frac{1}{|A| \cdot n}$, and hence

$$\sum_{n \in \mathbb{N}} p^{S_{m_\epsilon}}_\omega(E_n \mid \cap_{k<n} E^c_k) = \infty$$

By the Borel-Cantelli Lemma, it follows that $p^{S_{m_\epsilon}}_\omega(E_n$ infinitely often$) = 1$. In other words, the only learner in S_{m_ϵ} plays a infinitely often. As a was chosen arbitrarily, every action in A is played infinitely often. By Lemma 7, g's estimates of the expected utility of each action approach the true expected utility in every possible state of the world almost surely. Because m_ϵ plays the (estimated) EU maximizing actions with probability approaching one in every state of the world, it follows that $p^{S_{m_\epsilon}}_\omega(\lim_{n \to \infty} m_\epsilon(h_n)(A_\omega) = 1) = 1$. By Lemma 8, the result follows.

Proposition 2. Let $\langle \Omega, A, O, p \rangle$ be a difficult learning problem. Then $\langle m_\epsilon, m_\epsilon \rangle$ is not GIC.

Proof: Let $S = \langle G = \{g_1, g_2\}, \langle m_\epsilon, m_\epsilon \rangle \rangle$ be the strategic network consisting of exactly two researchers, both of whom employ the method m_ϵ. Let $\omega_1 \in \Omega$. As the learning problem is non-trivial, there is some $\omega_2 \in \Omega$ such that $A_{\omega_1} \cap A_{\omega_2} = \emptyset$. As the learning problem is difficult, there is some history $h \in H_G$ such that (i) every action in A_{ω_1} has garnered zero payoff along h, (ii) some action in A_{ω_2} has garnered positive payoff along h, and (iii) $p^S_{\omega_1}(h) > 0$. Suppose h has length

n. Define:

$$\begin{aligned} E &= [h] \cap \bigcap_{g \in G} \bigcap_{j > n} [h^A(j,g) \notin A_{\omega_1}] \\ F &= [h] \cap \bigcap_{g \in G} \bigcap_{j > n} [h^A(j,g) \in A_{\omega_2}] \\ F_k &= [h] \cap \bigcap_{g \in G} \bigcap_{n < j < n+k} [h^A(j,g) \in A_{\omega_2}] \end{aligned}$$

Notice first that $F \subseteq E$, and so $p^S_{\omega_1}(F) \leq p^S_{\omega_1}(E)$. Thus, it suffices to show that $p^S_{\omega_1}(F) > 0$. Next notice that $F_1 \supseteq F_2 \supseteq \ldots F$, and so $\lim_{k \to \infty} p^S_{\omega_1}(F_k) = p^S_{\omega_1}(F)$. Because m_ϵ chooses action in $A \setminus A(g,h)$ with probability at most $\frac{1}{|h|^2}$, it is easy to check, by induction on k, that

$$p^S_{\omega_1}(F_k) \geq p^S_{\omega_1}([h]) \cdot \Pi_{n<j<k} \ (1 - \frac{1}{j^2})^2.$$

The term under the product sign is squared because g_1 and g_2 choose their actions independently of one another. It follows that:

$$p^S_{\omega_1}(F) = \lim_{k \to \infty} p^S_{\omega_1}(F_k) \geq \lim_{k \to \infty} p^S_{\omega_1}([h]) \cdot \Pi_{n<j<k} \ (1 - \frac{1}{j^2})^2 > 0$$

where the last inequality follows from the fact that $p^S_{\omega_1}(h) > 0$. By Corollary 2, the result follows. Notice the same proof works for any finite sequence that has range m_ϵ and length greater than or equal to two.

Proposition 3. Let $\langle \Omega, A, O, p \rangle$ be any learning problem. Let m be the SA method determined by the following. The transition probabilities $q_{a,a'}$ equal $\frac{1}{|A|}$ for all $a, a' \in A$, and the cooling schedule $T : H \to \mathbb{R}$ is the function $T(h) = \log(|h|^{|h_1|})$ (here, log is the natural logarithm). Then m is IC. If $\langle \Omega, A, O, p \rangle$ is difficult, then $\langle m, m \rangle$ is not GIC.

Proof: The proofs of the two claims are exactly analogous to those of Propositions 1 and 2. See Mayo-Wilson, Zollman, and Danks (2010) for details.

In the following two propositions, let $m^{\delta\epsilon}_a$ be the $\delta\epsilon$ method determined by the triple $\langle a, \epsilon, \delta \rangle$, where $\epsilon(h) = 0$ and

$$\delta(h) = \begin{cases} 1 \text{ if } a \in A(g,h) \\ \frac{1}{|h|} \text{ otherwise} \end{cases}$$

Proposition 4. Let $\langle \Omega, A, O, p \rangle$ be a non-trivial learning problem. Then $m^{\delta\epsilon}_a$ is not IC.

Proof: Let S be the isolated network consisting of one learner g employing the method $m^{\delta\epsilon}_a$. As the learning problem is non-trivial, there is some $\omega \in \Omega$ such that $a \notin A_\omega$. This implies that $[h^A(n) \notin A_\omega] \subseteq [h^A(n) = a]$. Define

E to be the set of histories along which only the action a is played, i.e., $E = \bigcap_{n \in \mathbb{N}} \ [h^A(n) = a]$. By Corollary 2, it suffices to show that $p^S_\omega(E) > 0$. In fact, we show E has probability one. To do so, note that, by convention, the initial weights assigned to each action in A are zero, so that a appears to be an optimal action on the first stage, i.e. $a \in A(g, \langle - \rangle)$. So g plays a with probability one on the first stage. Because outcomes are non-negative, it follows that regardless of the outcome of the first play, a remains seemingly optimal at stage 2, and so on. Hence, regardless of the state of the world, in *every* history h for the isolated network S with positive probability, the only action played along h is a. It follows that $p^S_\omega(E) = 1$.

Proposition 5. Let $\langle \Omega, A, O, p \rangle$ be any learning problem, and $M = \langle m^{\delta\epsilon}_a \rangle_{a \in A}$. Then M is GUC.

Proof: Let $S = \langle G, N \rangle$ be any strategic network containing a connected M-subnetwork $S' = \langle G', M \rangle$. Let $\omega \in \Omega$. Pick some $a \in A_\omega$, and some $g \in G'$ such that $m_g = m^{\delta\epsilon}_a$. Let $E_n = [h^A(n, g) = a]$, so that $p^S_\omega(E_n | \cap_{k<n} E^c_k) \geq \frac{1}{n}$ by definition of $m^{\delta\epsilon}_a$. By the Second Borel-Cantelli Lemma, it follows that $p^S_\omega(g \text{ plays } a \ \text{ infinitely often }) = 1$.

By Lemma 7, it follows that, almost surely, every learner in $N_G(g)$ has an estimate of the EU of a that approaches the actual EU of a in ω. Because $a \in A_\omega$, the definition of the strategies $\{m^{\delta\epsilon}_{a'}\}_{a' \in A}$ and Lemma 8 imply that, almost surely, every learner in $N_G(g) \cap G'$ plays actions in A_ω with probability approaching one.

By Lemma 9, it follows that, almost surely, every learner in $N_G(g) \cap G'$ plays plays actions in A_ω infinitely often. Because A_ω is finite, by the pigeonhole principle, it follows that if an individual plays actions from A_ω infinitely often, then there is some $a' \in A_\omega$ that he plays infinitely often. It follows that, almost surely, for every learner in $g' \in N_G(g) \cap G'$, there is some action $a_{g'} \in A_\omega$ that g' plays infinitely often.

Now let $g'' \in G'$ be an agent such that g'' is a neighbor of some neighbor g' of g. We can repeat the argument above. Since g' plays some optimal action $a_{g'} \in A_\omega$ infinitely often almost surely, then by Lemma 7, it follows that g'' has an estimate of the EU of $a_{g'}$ that approaches the actual EU of $a_{g'}$ almost surely. By the definition of the strategies $\{m^{\delta\epsilon}_{a'}\}_{a' \in A}$ and Lemma 8, it then follows that g'' plays actions in A_ω with probability approaching one. So neighbors of neighbors of g play EU maximizing actions with probability approaching one if they are in G'.

In general, let $\pi(g, g')$ be the length of the shortest path between g and g' in G. By induction on $n \in \mathbb{N}$, we see that for any agent $g' \in G'$ such that $\pi(g, g') = n$, g' plays EU maximizing actions with probability approaching one. Because the subnetwork $S = \langle G', M \rangle$ is connected, there is always a finite path between g and g' for all $g' \in G$.

5.6 Proofs of Theorems

Proof of Theorem 1: That all RL strategies are IC under the assumptions of the theorem follows from Theorem 5, which is a trivial generalization of the proof of Theorem 1 in Beggs (2005). That some ϵG methods are isolation consistent follows from Proposition 1. Because every ϵG method is a $\delta\epsilon$ method, it follows that some $\delta\epsilon$ methods are IC. Finally, that some SA methods are isolation consistent follows from Proposition 3. For conditions characterizing when a wide class of SA methods are IC, see Bertsimas and Tsitsiklis (1993).

Proof of Theorem 2: Follows from Theorems 4 and 5.
Proof of Theorem 3: Follows immediately from Propositions 4 and 5.
Proof of Theorem 4: Follows immediately from Propositions 1, 2, 3.
Proof of Theorem 5: First, we show that every finite sequence of RL methods is GIC. Let M be any finite sequence of RL methods, and let $S = \langle G, N \rangle$ be any M-network (in fact, one need not assume G is connected). Pick $g \in G$ and $\omega \in \Omega$. We must show that $\lim_{n\to\infty} p^S_\omega(h^A(n, g) \in A_\omega) = 1$.

To do so, we adapt the proof of Theorem 1 in Beggs (2005) in the following way. It suffices to consider the case in which A contains exactly two actions a_1 and a_2. Beggs defines two random variables $A_i(n)$ and $\pi_i(n)$ (where $i = 1, 2$), which respectively represent the total utility acquired by playing action a_i through stage n and the payoff acquired by playing action a_i on stage n. In our model, these two random variables are the mappings $A_i(n) : \overline{H}_G \to \mathbb{R}^+$ and $\pi_i(n) : \overline{H}_G \to \mathbb{R}^+$ defined respectively by the equations $A_i(n)(\overline{h}) = r_{a_i,N(g)}(\overline{h}_n)$ and $\pi_i(n)(\overline{h}) = \sum_{g' \in N(g)} \overline{h}^O_n(n, g)$. Because the strategic network S contains only finitely many agents by assumption, the assumptions of the theorem imply that the variables $A_i(n)$ and $\pi_i(n)$ are bounded and can be plugged directly into the proof of Theorem 1 in Beggs (2005) to yield the result.

Next we show that no finite sequence of RL methods is GUC in any non-trivial learning problem in which there are constants $k_2 > k_1 > 0$ such that $p(o|a, \omega) = 0$ if $o \notin [k_1, k_2]$. Let M be a finite sequence of RL methods. It suffices to find (i) a strategic network $S = \langle G, N \rangle$ with a connected M-subnetwork $S' = \langle G', M \rangle$; (ii) a learner $g \in G'$; and (iii) a state of the world $\omega \in \Omega$ such that $\lim_{n\to\infty} p^S_\omega(h^A(n, g) \in A_\omega) \neq 1$.

To construct S, first take a sequence of learners of the same cardinality as M and place them in a singly-connected row, so that the first is the neighbor to the second, the second is a neighbor to the first and third, the third is a neighbor to the second and fourth, and so on. Assign the first learner on the line to play the first strategy in M, the second to play the second, and so on. Denote the resulting strategic network by $S' = \langle G', M \rangle$; notice S' is a connected M-network.

Next, we augment S' to form a larger network S as follows. Find the least natural number $n \in \mathbb{N}$ such that $n \cdot k_1 > 3 \cdot k_2$. Add n agents to G' and add an edge from each of the n new agents to each old agent $g \in G'$. Call the resulting network G. Pick some action $a \in A$, and assign each new agent the strategy m_a, which plays the action a deterministically. Call the resulting strategic network

S; notice that S contains S' as a connected M-subnetwork.

Let ω be a state of the world in which $a \notin A_\omega$ (such an a exists because the learning problem is non-trivial by assumption). We claim that

$$(*) \lim_{k \to \infty} p^S_\omega(h^A(k, g) \in A_\omega) < 1$$

for all $g \in G'$, and so M is not GUC. Let $g \in G'$. By construction, regardless of history, g has at least n neighbors each playing the action a at any stage. By assumption, $p(o|a, \omega) > 0$ only if $o \geq k_1 > 0$, and so it follows that the sum of the payoffs to the agents playing a in g's neighborhood is at least $n \cdot k_1$ at every stage. In contrast, g has at most 3 neighbors playing any other action $a' \in A$. Since payoffs are bounded above by k_2, the sum of payoffs to agents playing actions other than a in g's neighborhood is at most $3 \cdot k_2 < n \cdot k_1$. It follows that, in the limit, one half is strictly less than ratio of (i) the total utility accumulated by agents playing a in g' neighborhood to (ii) the total utility accumulated by playing all actions. As g is a reinforcement learner, g, therefore, plays action $a^* \notin A_\omega$ with probability greater than one half in the limit, and $(*)$ follows.