Wisdom of crowds versus groupthink: learning in groups and in isolation

Conor Mayo-Wilson, Kevin Zollman & David Danks

International Journal of Game Theory

ISSN 0020-7276

Int J Game Theory
DOI 10.1007/s00182-012-0329-7

Springer

Int J Game Theory
DOI 10.1007/s00182-012-0329-7

Wisdom of crowds versus groupthink: learning in groups and in isolation

Conor Mayo-Wilson · Kevin Zollman · David Danks

Received: 1 December 2010 / Accepted: 2 April 2012

Abstract We evaluate the asymptotic performance of boundedly-rational strategies in multi-armed bandit problems, where performance is measured in terms of the tendency (in the limit) to play optimal actions in either (i) isolation or (ii) networks of other learners. We show that, for many strategies commonly employed in economics, psychology, and machine learning, performance in isolation and performance in networks are essentially unrelated. Our results suggest that the performance of various, common boundedly-rational strategies depends crucially upon the social context (if any) in which such strategies are to be employed.

Keywords Bandit problems · Networks · Reinforcement learning · Simulating annealing · Epsilon greedy

1 Introduction

In a 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

C. Mayo-Wilson (✉)
Department of Philosophy, Baker Hall 135, Carnegie Mellon University, Pittsburgh, PA 15213, USA
e-mail: conormw@andrew.cmu.edu

K. Zollman
Department of Philosophy, Baker Hall 155D, Carnegie Mellon University, Pittsburgh, PA 15213, USA
e-mail: kzollman@andrew.cmu.edu

D. Danks
Department of Philosophy, Baker Hall 161E, Carnegie Mellon University, Pittsburgh, PA 15213, USA
e-mail: ddanks@cmu.edu

Published online: 20 April 2012

(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, seek 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), choice of drilling sites by oil companies (Keller et al. 2005), and choice of modeling techniques in the sciences (Zollman 2009).

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 study strategies which do not maximize expected utility. 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 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. An individual 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; Ellison and Fudenberg 1993; Bala and Goyal 2008; Keller et al. 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 ensure that this community will converge to the optimal action, but other social structures will not.

Although Beggs and Bala and Goyal seem to utilize essentially the same metric for the comparison of boundedly rational algorithms—convergence in the limit—their investigation are more different than they appear. Beggs considers how an *individual* does when he plays a bandit in isolation; Bala and Goyal consider how a *group* fares when each of its members confronts the same bandit. The myopic maximizer of Bala and Goyal would not converge in the limit if he was in isolation (Huttegger 2011). More surprisingly, Beggs' reinforcement learner might not converge if placed in the wrong social circumstance.

The above investigations raise a central question: what relation, if any, is there between a strategy's performance considered in isolation and its performance in a social context? To answer this question, we make precise four different ways of evaluating the asymptotic performance of strategies in multi-armed bandit problems. We then characterize the performance of a variety of boundedly rational strategies. We find that which boundedly rational strategies are judged as appropriate depends critically on (i) whether the strategy is considered in isolation or in a wider social context, and (ii) whether the strategy is evaluated in itself, or as part of a collection of complimentary strategies that can work together. Our results, we believe, make perspicuous the choices one must make before engaging in the analysis of various boundedly rational strategies.

In Sect. 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 Sect. 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

A *learning problem* is a quadruple $\langle \Omega, A, O, p \rangle$, where Ω is a set of unknown states of the world, A is a finite set of actions, O is a countable set of non-negative real numbers called outcomes, and $p = \{p(\cdot|a, \omega)\}_{a \in A, \omega \in \Omega}$ is a set if probability measures specifying the probability of obtaining a particular utility given an action and state of the world.[1]

Informally, 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 countable set O of non-negative real numbers. Importantly, the probability of obtaining an outcome given a particular action and state of the world does not change over time. For this reason, what we call a learning problem is also called a "stochastic" or "non-adversarial" multi-armed bandit problem to indicate that agents are attempting to learn a fixed state of the world, not the strategy of a competing player in a game.[2] There is a set Ω of possible *states of the world* that determines the probability distribution over O associated with each action. Agents do not know the state of the world, and their actions aim to discover it.

To model communication among agents, we use finite undirected graphs $G = \langle V_G, E_G \rangle$, where vertices V_G represent individual agents, and edges E_G represent pairs of agents who can 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 a *group* of learners $G' \subseteq V_G$. 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.

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 $|C|$ denote its cardinality, which if C is a sequence/function, is the length of C. Let $C^{<\mathbb{N}}$ be the set of all finite sequences with range in C, and given a seqeuence σ, let σ_n denote the nth coordinate of σ. Then define the set H of *possible histories* as follows:

[1] The bar notation $p(\cdot|a, \omega)$ ought **not** be understood as conditioining. In other words, for distinct actions a and a' and distinct states of the world ω and ω', the probability measures $p(\cdot|a, \omega)$ and $p(\cdot|a', \omega')$ are not obtained from conditioning on a third common measure; they are essentially unrelated. We use this notation because it is standard and allows us to avoid notational conflicts with other definitions that we introduce.

[2] For an exposition of the adversarial model, see Cesa-Bianchi and Lugosi (2006).

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

Here, h_n is the sequence of actions and outcomes obtained by some collection of learners at stage n of inquiry, and so 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. If G' is a group and $h \in H_{G'}$ is a history for the group, the expressions $\mathcal{A}_{g,n}(h)$ and $\mathcal{O}_{g,n}(h)$ denote the action taken and outcome obtained respectively by learner $g \in G'$ on the nth stage of inquiry.

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

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

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 better treatment. A possible network history $h \in H_G$ of length two is $\langle\langle a_1, 1\rangle, \langle a_1, 0\rangle\rangle, \langle\langle a_1, 0\rangle, \langle a_2, 0\rangle\rangle\rangle$, 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 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 following simple 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

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

Notice, however, the same history h might have a different probability if one were to srespecify 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 μ_ω^a of the action a that is constant throughout time. Hence, in any state of the world ω, there is some collection $A_\omega = \{a \in A : \mu_\omega^a \geq \mu_\omega^{a'} \text{ for all } a' \in A\}$ of *optimal* actions that maximize expected utility. It follows that the event that g plays an optimal action at stage n has a well-defined probability, which we will denote $p_\omega^S(\mathcal{A}_{g,n} \in A_\omega)$. In the next section, we study the limiting behavior of such probabilities in various strategic networks.

Some learning problems are far easierthan others; for example, if one action has higher expected utilityin every world-state, then there is relatively little for the agents to learn. We are principally interested in more difficult problems. We 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 $p_{\omega_1}^S(h) > 0$ for some $\omega_1 \in \Omega$, then there exists $\omega_2 \in \Omega$ such that $A_{\omega_1} \cap A_{\omega_2} = \emptyset$ and $p_{\omega_2}^S(h) > 0$. We 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 or fail, and no history determines an optimal action with certainty.

2.1 Six types of strategies

Although the number of differing strategies is enormous, we will focus on the behavior of six types of boundedly rational strategies: decreasing ϵ-greedy (ϵG), reinforcement learning (RL), weighted reinforcement learning (wRL), simulated annealing (SA), upper confidence bound algorithms (UCB), and what we call, $\delta\epsilon$ methods. We study these strategies for four reasons. First, the first five 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 six 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. For any sequence σ and any natural number n, the expression $\sigma \restriction n$ denotes the initial segment of σ of length n; by convention, let $\sigma \restriction n = \sigma$ if $|\sigma| < n$. 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 $\sigma \cong \sigma'$.

In the definitions of the various strategies below, let $w = \langle w_a\rangle_{a\in A}$ be a vector of non-negative real numbers.

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 by introducing some probabilistic experimentation. For instance, 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 such an 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, and more generally, let each agent begin with an initial estimate of the expected utility of each action, given by the vector $\langle w_a\rangle_{a\in A}$. Let $\hat{\mu}^a_{g,n}(h)$ be g's estimate of the expected utility of action a at stage n along history h, where n is less than or equal to the length of 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 of action a through stage n from plays in g's neighborhood. Additionally, define the set of actions which currently have the highest estimated utility:

$$\mathcal{A}^*_{g,n}(h) := \{a \in A :\ \hat{\mu}^a_{g,n}(h) \geq \hat{\mu}^{a'}_{g,n}(h) \text{ for all } a' \in A\}$$

Given (i) a vector $w = \langle w_a\rangle_{a\in A}$ 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 $\epsilon(h') \leq \epsilon(h)$), an ϵG method determined by $\langle w, \epsilon\rangle$ is any method m of the form:

$$m(h)\left(\mathcal{A}^*_{g,n}(h)\right) = 1 - \epsilon(h)$$
$$m(h)\left(A \setminus \mathcal{A}^*_{g,n}(h)\right) = \epsilon(h)$$

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 $a \in A$.

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, which represent the initial weights. Let $h \in H_{G'}$ be the history for the individuals in g's neighborhood. Let $\mathcal{W}^a_{g,n}(h)$ represent the total accumulated payoff for action a in g's neighborhood in history h through stage n, which includes the initial weight w_a. Define $\mathcal{W}_{g,n}(h) := \sum_{a \in A} \mathcal{W}^a_{g,n}(h)$, which represents the total payoff through stage n of all actions (and their initial weights) in g's neighborhood along the history h. An RL strategy m_w is defined by specifying w. For any w, the probability that an action a is played after observed history h is given by:

$$m_w(h)(a) = \frac{\mathcal{W}^a_{g,n}(h)}{\mathcal{W}_{g,n}(h)}$$

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.[3] In economics, for example, reinforcement learning has been used to model how individuals behave in repeated games in which they must learn the strategies being employed by other players.[4] 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 [see Theorem 4 below—the first half of which is a consequence of Theorem 1 in Beggs (2005)].

Weighted reinforcement learning (wRL): Although naive reinforcement learners asymptotically approach optimal behavior, the rate at which they do so is often quite slow (Beggs 2005). One way to quicken convergence is to redefine/alter the "weight" functions $\mathcal{W}^a_{g,n}$ introduced above. Instead of simply adding together the payoffs of a given action a, one can redefine $\mathcal{W}^a_{g,n}(h)$ to be a polynomial or exponential function of the payoffs obtained from playing a (Cesa-Bianchi and Lugosi 2006); we will make such a notion more precise below.

However, two facts should be observed about such wRL methods. First, the way in which the modified weight functions are used requires that utilities are bounded from above; so such methods are not well-defined in all learning problems like the remaining algorithms we consider. Second, one cannot simply use the new weight functions $\mathcal{W}^a_{g,n}$ as one did in simple reinforcement learning, namely, by choosing

[3] 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.

[4] 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).

action a with probability $\frac{\mathcal{W}^a_{g,n}(h)}{\mathcal{W}_{g,n}(h)}$. Doing so would fail to ensure that all actions are sufficiently explored, as the exponential or polynomial functions can cause the reinforcement associated with a given action to increase very quickly. Instead, one must introduce an experimentation parameter ϵ, as one does with ϵG strategies, that keeps the probability of playing a given action high.

Formally, suppose utilities are bounded above by a constant u. Define the learner g's "regret" for playing action a at stage n in h as follows:

$$\mathcal{R}^a_{g,n}(h) = \begin{cases} \frac{u - \mathcal{O}_{g,n}(h)}{m_g(h\restriction n)(a)} & \text{if } \mathcal{A}_{g,n}(h) = a \\ u - \mathcal{O}_{g,n}(h) & \text{otherwise} \end{cases}$$

Notice $\frac{1}{m_g(h\restriction n)(a)}$ is well-defined if $\mathcal{A}^a_{g,n} = a$, as in that case the probability that g's method chooses action a is strictly positive (as the set of actions is countable). Define the cumulative regret $\mathcal{R}^a_{g,\leq n}(h)$ of action a along h to be the sum of regrets $\mathcal{R}^a_{g,k}$, where $k \leq n$.

A wRL learner uses regrets as follows. Let $\phi : \mathbb{R} \to \mathbb{R}^+ \cup \{0\}$ be an increasing and twice-differentiable weight function, and let $\epsilon : H \to (0, 1)$ be an antitone function. Then the wRL method $m_{\phi,\epsilon}$ associated with ϕ and ϵ is given by:

$$m_{\phi,\epsilon}(h)(a) = (1 - \epsilon(h)) \cdot \frac{\phi'(\mathcal{R}^a_{g,\leq n}(h))}{\sum_{a \in A} \phi'(\mathcal{R}^a_{g,\leq n}(h))} + \frac{\epsilon(h)}{|A|}$$

Common choices for ϕ include $\phi(x) = \frac{1}{\lambda} e^{\lambda x}$ where $\lambda > 0$, and $\phi(x) = x^r_+$ where $r \geq 2$.

Simulated annealing (SA): In computer science, statistics, and many other fields, SA refers to a collection of techniques for minimizing some cost function.[5] 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.

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

[5] 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.

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.

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 positive 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 matrix 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) $T : H \to \mathbb{R}^+ \cup \{0\}$ 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 $a \in A$ and for all histories h of length $n + 1$ for a learner g, define:

$$s(h, a) = T(h) \cdot \max \left\{ 0, \hat{\mu}_{g,n}^{\mathcal{A}_{g,n}(h)}(h) - \hat{\mu}_{g,n}^{a}(h) \right\}$$

Here, s stands for "switch." The switch function $s(h, a)$ is large if the action a appears to have greater value than the action last played along h, and it is 0 otherwise. Moreover, the switch function decreases with the length of h because it incorporates the cooling schedule $T(h)$.

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. The value of $m_\sigma(\langle - \rangle)(a)$, we assume, is either zero or one so that some initial action is fixed, and

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

Since the switch function increases in the estimated value of the action a, the method m assigns greater probability to those actions with higher estimated value. Further, by the first case of the definition, because the switch function decreases with the length of h, the probability that m switches actions decreases with time. 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 $a \in A$.

Delta-epsilon ($\delta\epsilon$): The fourth class of methods that we consider consists of intuitively plausible algorithms, though they have not been studied prior to this paper. $\delta\epsilon$ 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 nth round, where ϵ_n decreases over time. The difference between the two types of strategies is that each $\delta\epsilon$ method has some set of "favorite" actions $A_f \subseteq A$ that it favors in early rounds. Hence, there is some sequence of (non-increasing) probabilities δ_n with which $\delta\epsilon$ methods plays its favorite action A_f on the nth round. The currently best actions are, therefore, played with probability $1 - \delta_n - \epsilon_n$ on the nth stage of inquiry.

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

$$m(h)\left(\mathcal{A}^*_{g,n}(h) \setminus A_f\right) = 1 - (\epsilon(h) + \delta(h))$$
$$m(h)\left(A \setminus \left(\mathcal{A}^*_{g,n}(h) \cup A_f\right)\right) = \epsilon(h)$$
$$m(h)(A_f) = \delta(h)$$

Every ϵG method is a $\delta\epsilon$ method, where $A' = \emptyset$ and δ is the constant function 0.

$\delta\epsilon$ 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. There is substantial psychological evidence that people exhibit initial action/option preferences (i.e., prior to observing any outcomes) that are grounded in aspects of the action or option that are unrelated to its utility, but which nonetheless bias people's choices in early stages of learning (Kuhlman and Marshello 1975; Stanovich and West 1998; Baron and Ritov 2004). 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.

Upper confidence bound (UCB): At each stage of inquiry, a UCB learner calculates an "upper confidence bound" u_a for eachaction a; intuitively, the bound u_a might represent learner's beliefs about the highest possible value that action a may have. The learner then chooses the action with the greatest upper confidence bound.

The motivation for UCB algorithms is fairly simple. If an action is played infrequently, then a learner may inaccurately estimate the action's value due to insufficient sample size. As a result, if the learner were to use a confidence interval (rather than a point-estimate) to represent his or her beliefs about the action's value, the resulting interval would be rather wide, containing values much higher than the sample mean. UCB algorithms force a learner to play actions sufficiently often so that the "confidence bounds" (which are not actually the bounds of confidence intervals) narrow in on the true value of each action. At early stages of inquiry, there will be a large difference between the confidence bounds and the estimated value of actions (heuristically, the confidence intervals with different actions are wide and overlapping), and so UCB learners explore many actions. At later stages of inquiry, when actions have been played sufficiently often, the confidence bounds and sample mean differ only slightly (heuristically, the confidence intervals are narrow), and so UCB learners will typically play those actions with highest estimated value, which, in the limit, are also the actions with the highest actual value. In this way, UCB algorithms achieve a balance between "exploration" and "exploitation."

The algorithms are defined formally as follows. For every group $G' \subseteq G$ and every group history $h \in H_{G'}$, let $\mathcal{O}^a_{G',n}(h)$ denote the outcome obtained when action a is played for the nth time by some member in G'. So if $\mathcal{T}^a_{G',n}(h)$ is the number of times a is played along h through stage n, then the outcomes obtained from playing a can be represented by a vector $\mathcal{O}^a_{G'}(h) := \langle \mathcal{O}^a_{G',1}(h), \mathcal{O}^a_{G',2}(h), \ldots, \mathcal{O}^a_{G',\mathcal{T}^a_{G',|h|}(h)}(h)\rangle$. UCB algorithms compute a confidence bound for the action a using this vector, and the total number of plays of *all* actions along h (which is just $|h| \cdot |h_1|$).

Let $\text{UCB} = \{\text{UCB}_{ni} : \mathbb{R}^i \to \mathbb{R} \cup \{\infty\} \mid n \in \mathbb{N}, i \leq n\}$ be a collection of Borel measurable functions satisfying the following two properties:

1. For fixed i, the functions UCB_{ni} are non-decreasing in n. In other words, for all fixed $i \in \mathbb{N}$ and all $r \in \mathbb{R}^i$, one has $\text{UCB}_{mi}(r) \geq \text{UCB}_{ni}(r)$ whenever $m \geq n \geq i$.
2. For any sequence of i.i.d. random variables $\mathcal{X}_1, \mathcal{X}_2, \ldots$ with finite mean $\mu < \infty$, and for any $r < \mu$

$$p\left(\text{UCB}_{ni}\left(\mathcal{X}_1, \ldots, \mathcal{X}_i\right) < r \text{ for some } i \leq n\right) = o\left(n^{-1}\right)$$

where the o in the second condition is to be understood as indicating "little o-notation" rather than an outcome. For any learner g and any neighborhood history $h \in H_{N_G(g)}$, let $\text{UCB}_a(h) = \text{UCB}_{|h|\cdot|h_1|, \mathcal{T}^a_{N_G(g),|h|}(h)}(\mathcal{O}^a_{N_G(g)}(h))$. Then the UCB algorithm m_{UCB} associated with UCB maps any history h to the action a that maximizes $\text{UCB}_a(h)$; if there is more than one such action, the method chooses the action with lowest index. Property (1) asserts that one's current upper confidence bound for a should not decrease unless a is played again; intuitively, it captures the fact that one's confidence in the value of a should depend only on the outcomes obtained when a has been played, and not on how frequently other actions in the network have been employed. Property (2) asserts that the upper confidence bound for all actions quickly approaches the true value of the action, and moreover, it does not remain below the true value for very long.

UCB strategies are desirable for a number of reasons. First, many of the common upper confidence bound functions are easily computable, and hence, UCB algorithms are easy to implement. Among other reasons, computation is eased by the fact that UCB algorithms are deterministic and hence, do not require use of a randomizing device like the other strategies above. Second, because the upper confidence bounds u_a approach the actions' actual values μ^a_ω at a quick enough rate (by the second property above), UCB learners will, in every learning problem, converge to playing an optimal action in the limit, regardless of social setting (See Theorem 2 below, which is a trivial consequence of Theorem 2.2 in Agrawal (1995)). Finally, UCB algorithms not only find optimal actions asymptotically, but when employed in isolation, they actually minimize regret at *each* stage of inquiry (Auer et al. 2002).

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 $\omega \in \Omega$:

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

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, SA-strategies and wRL 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 ϵG, SA, and $\delta\epsilon$ strategies are always (i.e. in every learning problem)* IC. *Similarly for wRL methods in the learning problems in which they are well-defined. Finally, 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.

Here, when we speak of "some" or "all" RL strategies, we mean "some" or "all" specifications of the parameters $\langle w_a\rangle_{a\in A}$ in the definitions of the algorithm. Similarly for the remaining algorithms. The second type of convergence that we consider is that of an individual learner in a network of other, not necessarily similar, learners. This notion, which we call "universal consistency", requires that the 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 $g \in G$:

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

By definition, any UC strategy is IC, since the isolated network is a strategic network. Now 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. But in fact, one class of the algorithms above both incorporates data from one's neighbors and is also UC:

Theorem 2 *[Agrawal (1995)] UCB algorithms are always* UC.

The proof of theorem 2 is obtained by simply rewriting Agrawal's proof in our notation, and so is omitted. However, none of the remaining strategies considered above are always UC:

Theorem 3 *In all difficult learning problems, there are RL, wRL, 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 ϵG strategies first. Recall, in such strategies, the ϵ function specifies the probability with which a learner will experiment/explore seemingly inferior actions. This function must 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.

How so? Consider a learner g who employs an ϵG strategy as follows. After n stages of learning, g randomly explores some seemingly suboptimal action with probability $\frac{1}{n^{x/y}}$, where x is the total number of actions that g has observed (including those of his neighbors) and y is the total number of actions that g herself has performed. Thus, our learner very reasonably adjusts the rate at which she explores new actions as a function of the amount of acquired evidence. Yet this is equivalent to adopting an experimentation rate of $1/n^{|N_G(g)|}$, i.e. to adjusting one's experimentation rate on the basis of one's neighborhood size. So our learner experiments with rate $\frac{1}{n}$ in isolation, but with rate no greater than $\frac{1}{n^2}$ when embedded in a larger network. By the Borel Cantelli Lemma, the former rate is large enough to ensure that each action is played infinitely often (and thus the space of actions is explored), whereas the latter is not. So some ϵG strategies (and hence, some $\delta\epsilon$ strategies) are IC but not UC.

Similar remarks apply to SA and wRL strategies. For instance, recall that SA strategies are driven by a "cooling schedule", which dictates the probability with which one switches from one action to another. Just as the ϵ function in ϵG methods can be adjusted to drive exploration in isolation but not in a social context, so can the cooling schedule of SA strategies be defined so that, in social settings, learners adopt some suboptimal action early and, with non-zero probability, never abandon it.

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 suboptimal actions with positive probability in the limit.

One might wonder if this result is at all robust, as the networks considered here are *unweighted*. That is, perhaps the fact the RL methods are not UC depends crucially upon the assumption that agents trust all of their neighbors equally. Instead, one might imagine that an agent assigns real-valued "weights" to each of her neighbors, where such weights might represent how much the agent trusts her neighbors. If agents weight their neighbors in such a way, one might hypothesize that a learner employing an RL method cannot be misled. This hypothesis is false, so long as the weights that

agents assign to neighbors are constant. After proving Theorem 3 in the Appendix, we show a similar result holds for RL methods even in weighted networks.

The argument that RL strategies fail to be UC, however, does rely heavily the existence of learners who fail in the limit to find the optimal action. This might occur because the RL learner is surrounded by individuals intent on deceiving her, or alternatively because she is surrounded by individuals who have a different utility function over outcomes. One might exclude this possibility. When only RL methods are present in a network, then Theorem 4 below shows that, under most assumptions, every learner is guaranteed to find optimal actions. That is, RL methods work well together as a *group*. We will return to this point after introducing more general definitions of group rationality. Before doing so, however, we explain why UCB algorithms are uniquely UC among the methods we have considered.

Like ϵG and SA methods, UCB algorithms are designed to balance "exploration and exploitation" and so one might wonder why the upper-confidence bound functions cannot be tuned so as to yield UCB functions that are IC but not UC in the same way one can for ϵG and SA methods. The difference is that, in UCB algorithms, the confidence bound u_a associated with a as a function of both (i) the total number of actions taken, and (ii) the number of times a in particular has been played. Importantly, the bound associated with a cannot decrease unless a is played again (by property (2)), and so the action a must be explored until its inferiority is nearly certain. This is not true of the ϵG, SA, wRL, and RL methods considered above, as their rates of experimentation do not depend upon how often any particular action has been employed.

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. We 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 $N \cong M$.

In other words, an N network is a network in which all and only the methods in N are employed.

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 $\omega \in \Omega$:

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

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 $\omega \in \Omega$:

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

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 $\mathbf{M}$ be a set of methods. Then $\mathbf{M}$ is GIC (respectively, GUC) if for for every sequence of methods M such that $ran(M) = \mathbf{M}$, the sequence M is GIC (respectively, GUC).

So a set $\mathbf{M}$ is GIC if, for all connected networks that have only methods in $\mathbf{M}$ and each method in $\mathbf{M}$ occurs at least once in the network, each learner in the network converges to playing optimal actions. A set $\mathbf{M}$ is GUC if, for all networks in which each method in $\mathbf{M}$ is represented at least once and those employing $\mathbf{M}$ are connected by paths of learners using $\mathbf{M}$, each agent in the subnetwork employing $\mathbf{M}$ 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 as 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. Thus, 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 4 *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.

The first half of Theorem 4 asserts that RL methods do, as we expected, perform well as a group when no other agents are present. However, the proof of the second half of the theorem shows that just as a single learner employing an RL method can be misled by his or her neighbors, so can arbitrarily large groups of RL methods be swayed by actions in the ambient network.

One might wonder whether, like RL methods, all IC methods form GIC groups. That conjecture is false:

Theorem 5 *In difficult learning problems, there are sequences M (respectively sets) of ϵG (and hence, $\delta\epsilon$) 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 methods, and for wRL methods in the learning problems in which they are well-defined.*

Notice that, with respect to ϵG, SA, and $\delta\epsilon$ methods, Theorem 5 entails Theorem 3. In fact, the sketch of proof Theorem 3 given above is also a sketch of proof of Theorem 5: in networks consisting of exclusively ϵG or SA learners, agents may acquire information too quickly thereby failing to explore the space of actions sufficiently.

Thus far, we have investigated the relationship between (i) IC and UC methods, and (ii) IC and GIC sets of methods. And 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). So what is the relationship between IC and GUC sets of methods?

In general, not all sequences (respectively sets) of methods that are GIC or GUC need to be composed entirely of IC methods. Consider, for instance, the set of strategies consisting of one UC method, and another method that "imitates" the best strategy amongst one's neighbors (other than oneself) and defaults to some fixed action if one has no neighbors. Such imitators will fail to converge in isolation, as in isolation, they may default to some fixed inferior action. However, if a *connected* network consists of at least one UC method and such imitators, then all agents will always play EU-maximizing actions in the limit, and hence, the sequence is GIC. Why? Since there is a learner g employing a UC method, he or she will play EU-maximizing actions with probability one in the limit. All of g's neighbors, by definition, either imitate g or employ the same method as g, and therefore, they also play EU maximizing actions with probability one in the limit. Thus, neighbors of neighbors of g also play EU maximizing actions with probability one in the limit. And so on. Because the network is connected, g's optimal behavior cascades through the entire network.

This argument, however, trades on the fact that at least one learner in the network employs a UC strategy. Surprisingly, there are sequences and sets of strategies that are GUC, but such that no strategy itself is IC (let alone UC).

Theorem 6 *In non-trivial learning problems, there are sequences and sets of $\delta\epsilon$ methods M such that M is* GUC*, but no m in M is* IC*.*

The proof of Theorem 6 is nearly identical to the above argument concerning the network of "imitators." For simplicity, consider the class of $\delta\epsilon$ methods for which ϵ is the constant function 0. Like the imitators, these $\delta\epsilon$ methods mimic the behavior of their most successful neighbors with high probability, and otherwise, they default to some preferred set of actions with probability δ. For concreteness, assume that δ assigns unit probability to some fixed action a if a appears to be optimal, and probability $\frac{1}{n}$ otherwise, where n is the number of stages of learning that have elapsed. In isolation, such a $\delta\epsilon$ method may fail to explore the space of actions sufficiently because it may default to its preferred action a over and over again.

However, consider a set M of *different* $\delta\epsilon$ methods, where each method has a *different* favored action. Suppose, for example, the set of methods are sufficiently diverse so that every action is favored by at least one method in the set. Now consider a network consisting of such methods. Whatever the true state of the world, there is some agent g in the network who favors an optimal action by assumption. Thus, if δ is defined as above (so that the chance of playing one's favored action is always at least $\frac{1}{n}$), then g will play his favored action infinitely often. So both g and his neighbors will learn that said action is optimal, and because $\delta\epsilon$ learners imitate the best available

action, both g and his neighbors employing $\delta\epsilon$ methods will perform his favored action (or some other optimal action) with unit probability in the limit. By the same reasoning, neighbors of neighbors of g will also adopt g's behavior (or some optimal action). And so on. If each of the $\delta\epsilon$ learners are connected by a path in the network, g's optimal behavior will cascade through the entire network among such learners. So sets of $\delta\epsilon$ methods can be GUC, though no individual method in the set is IC. This argument provides further evidence for the the broader thesis that a "diversity" of learning methods or problem-solving approaches can increase the performance of a group, even when each individual's approach is not particularly fruitful (Hong and Page 2001, 2004).

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 between the four types of convergence that is not discussed in the above corollary is the relationship between GUC and GIC. But recall that Theorem 4 asserts that some GIC sets are not GUC, and we note the reverse implication is trivial.

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.

Our analysis, however, is only a first step in understanding the connections between individual and group rationality in learning. Our work ought to be extended in at least five ways.

First, 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) illustrate how a method which is not IC, UC, GIC, or GUC nonetheless converges in a *particular* type of network. Additional investigation into more permissive notions of

individual and group rationality are likely to illustrate the virtues of other boundedly rational learning rules, and may potentially reveal further conceptual distinctions.

Second, the algorithms we consider do not directly assess the reliability and/or trust-worthiness of data acquired from other learners in the network. It is obvious that "real-world" learners do assess and weigh competing information received from others. Thus, one natural extension of our work is to modify common learning algorithms, like the six classes considered above, so that they employ such "social" information about others' reliabilities. One could then investigate whether such modified algorithms also exhibited divergent behavior in isolation versus social contexts.

Third, although we distinguish four criteria of individual and group rationality, we focus exclusively on *asymptotic* convergence to optimal behavior. In contrast, important work in machine-learning (see, for example, Auer et al. (2002)) aims to find algorithms that minimize regret *uniformly* at each stage of learning. It is important to characterize whether or not such stricter criteria—which essentially require agents to learn *quickly*—yield divergent judgments of individual and group rationality in the same way we have found with asymptotic criteria.

Fourth, in addition to considering different methods/algorithms and alternative criteria of rationality within the bandit-problem framework, our results 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.

In particular, one natural question is how our results extend to the "adversarial" multi-armed bandit problem, in which a forecaster competes with a player who chooses the observed outcomes so as to maximize the forecaster's predictive loss. This model can be understood as representing the task of learning other players' strategies in a competitive game, or as statistical inference from non-stationary processes. How such a model should be generalized to incorporate network learning (e.g. Do all players in a network compete agains a common player? Can the common opposing player choose different outcomes for different learners? etc.) is an open question.

Fifth, there are a range of natural applications for our results. As already suggested, understanding how various boundedly rational strategies perform in a multi-armed bandit problem has important implications for a variety of different economic phenomena, and in particular, for models of the influence that social factors on learning. 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.

Appendix A: Formal definitions

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. 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 sequence is not an initial segment of the latter. Given a network G and a group $G' \subseteq G$ and any $n \in \mathbb{N}$, we let $H_{G',n}$ denote sequences of $H_{G'}$ of length n. Because (i) the sets of actions, outcomes, and agents 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 $\mathcal{A}_{g,n}(h)$ to denote the action taken by g on the nth stage of inquiry in h, and $\mathcal{O}_{g,n}(h)$ 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), however, it will be easier to simply write $\mathcal{A}_g(h)$ and $\mathcal{O}_g(h)$ 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), then we write $\mathcal{A}_n(h)$ and $\mathcal{O}_n(h)$ to denote the action and outcome respectively taken/obtained at stage n in h. Finally, for any group history $h \in H_{G'}$, define $Q_{G'}(h, a) := \{\langle n, g\rangle \in \mathbb{N} \times G' : \mathcal{A}_{g,n}(h) = a\}$, to bet the set of ordered pairs $\langle n, g\rangle$ such that g plays a at stage n in h.

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 $\mathbf{H}_{G'}$. Define complete individual histories $\mathbf{H}_g$ similarly. By abuse of notation, we let $\mathbf{H}_G = \cup_{G'\subseteq G}\mathbf{H}_{G'}$ 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:

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

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, and so let τ_G be the topology generated by sets of the form $[h]$, i.e. τ_G is arbitrary unions of sets of the form $[h]$, where $G' \subseteq G$ and $h \in H_{G'}$. Let $\mathbb{F}_G$ be the Borel algebra generated by τ_G.

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

1. $[\mathcal{A}_{g,n} = a] = \{\mathbf{h} \in \mathbf{H}_G : \mathcal{A}_{g,n}(\mathbf{h}_n) = a\}$ *for fixed $a \in A$ and $g \in G$*
2. $[G'$ *plays* A' *infinitely often*$] = \{\mathbf{h} \in \mathbf{H}_G : \forall n \in \mathbb{N}\exists k \geq n\exists g \in G'(\mathcal{A}_{g,k}(\mathbf{h}_k) \in A')\}$ *for fixed $A' \subseteq A$ and $G' \subseteq G$*
3. $[\lim_{n\to\infty} \hat{\mu}^a_{g,n} = r\ ] = \{\mathbf{h} \in \mathbf{H}_G : \lim_{n\to\infty} \hat{\mu}^a_{g,n}(\mathbf{h}_n) = r\}$ *for fixed $a \in A$, $g \in G$, and $r \in \mathbb{R}$.*
4. $[\lim_{n\to\infty} m(h_n)(A_\omega) = 1] = \{\mathbf{h} \in \mathbf{H}_G : \lim_{n\to\infty} m(\mathbf{h}_n)(A_\omega) = 1\}$*, where ω is a fixed state of the world such that A_ω is finite, and m is a fixed method.*

Hence, if one considers $\mathcal{A}_{g,n} : \mathbf{H}_G \to A$ as the function $\mathbf{h} \mapsto \mathcal{A}_{g,n}(\mathbf{h}_n)$, then by Theorem 2.1, the mapping $\mathcal{A}_{g,n}$ is a random variable with respect to the power set algebra on A (recall, A is countable). Similar remarks apply to $\mathcal{O}_{g,n}$, $\mathcal{O}^a_{g,n}$, $\mathcal{W}^a_{g,n}$, $\mathcal{R}^a_{g,n}$ etc. when considered as functions from complete histories to the set of outcomes (or sums of outcomes). In general, we will use calligraphic letters to denote random variables.

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 $h \in H_{G',n}$.

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

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 $\langle \mathbf{H}_G, \mathbb{F}_G\rangle$ such that

$$p^S_\omega([h]) = p^S_{G',\omega,n}(h) \text{ for all } G' \subseteq G \text{ and all } h \in H_{G'}.$$

The measure p^S_ω exists and is unique by Caratheodory's Extension theorem. Details are available in Mayo-Wilson et al. (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.

Basic Lemmas

Lemma 3 *Let $S = \langle G, M\rangle$ be any strategic network, $g \in G$, and $a \in A$. Then for all $\omega \in \Omega$:*

$$p^S_\omega\left(\lim_{n\to\infty} \hat{\mu}^a_{g,n} = \mu^a_\omega \mid N_G(g) \textit{ plays a infinitely often}\right) = 1$$

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

Proof Fix $g \in G$ and consider the random variables $\mathcal{O}^a_{N_G(g),n}$ restricted to the subalgebra of $\mathbb{F}_G$ generated by the event $E = [N_G(g)$ plays a infinitely often] (i.e. consider them as maps to the set of outcomes, endowed with power set algebra, from the set $[N_G(g)$ plays a infinitely often] endowed with the σ-algebra $\mathbb{F}_G \cap E := \{F \cap E : F \in \mathbb{F}_G\}$). These random variables are independent by construction of p^S_ω. They are identically distributed on the subalgebra generated by the event E because the action a is played infinitely often by definition in E, and they have mean μ^a_ω by definition. Hence, by the strong law of large numbers:

$$\lim_{n\to\infty} \hat{\mu}^a_{g,n} := \lim_{n\to\infty} \frac{\sum_{j\le n} \mathcal{O}^a_{N_G(g),j}}{n} = \mu^a_\omega$$

with probability one under the measure $p^S_\omega(\cdot|N_G(g)$ plays a infinitely often). The result follows. □

Lemma 4 *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 $\lim_{n\to\infty} p^S_\omega(\mathcal{A}_{g,n} \in A_\omega) = 1$.*

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

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

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

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

Notice that $\mathbf{E}_{1,\epsilon} \subseteq \mathbf{E}_{2,\epsilon} \subseteq \ldots$, and so it follows that $\lim_{n\to\infty} p^S_\omega(\mathbf{E}_{n,\epsilon})$ $= p^S_\omega(\cup_{n\in\mathbb{N}}\mathbf{E}_{n,\epsilon})$. Now by assumption:

$$p^S_\omega\left(\lim_{n\to\infty} m_g(h_n)(A_\omega) = 1\right) = 1$$

Notice that

$$\left[\lim_{n\to\infty} m_g(h_n)(A_\omega) = 1\right] = \cap_{\delta\in\mathbb{Q}\cap(0,1)} \cup_{n\in\mathbb{N}} \mathbf{E}_{n,\delta}.$$

So it follows that

$$\begin{aligned}
1 &= p^S_\omega\left(\lim_{n\to\infty} m_g(h_n)(A_\omega) = 1\right)\\
&= p^S_\omega(\cap_{\delta\in\mathbb{Q}\cap(0,1)} \cup_{n\in\mathbb{N}} \mathbf{E}_{n,\delta})\\
&\leq p^S_\omega(\cup_{n\in\mathbb{N}} \mathbf{E}_{n,\epsilon})
\end{aligned}$$

$$= \lim_{n\to\infty} p^S_\omega(\mathbf{E}_{n,\epsilon})$$
$$\leq \frac{1}{1-\epsilon} \cdot \lim_{n\to\infty} p^S_\omega(\mathcal{A}_{g,n+1} \in A_\omega) \text{ by the argument above}$$

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

Lemma 5 *Let $S = \langle G, M\rangle$ be a strategic network, $g \in G$, $A' \subseteq A$, and $\omega \in \Omega$. Suppose $\lim_{n\to\infty} p^S_\omega(\mathcal{A}_{g,n} \in A') = 1$. Then:*

$$p^S_\omega(g \textit{ plays } A' \textit{ infinitely often }) = 1.$$

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 $] = \cup_{n\in\mathbb{N}} \cap_{k\geq n} [\mathcal{A}_{g,k} \notin A']$, and so (by countable additivity), there is some $j \in \mathbb{N}$ such that $p^S_\omega(\cap_{k\geq j}[\mathcal{A}_{g,k} \notin A']) = r > 0$. It follows that $p^S_\omega(\mathcal{A}_{g,k} \in A') \leq 1 - r$ for all $k \geq j$. Hence, $\lim_{n\to\infty} p^S_\omega(\mathcal{A}_{g,n} \in A') \leq 1 - r < 1$. □

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*

$$p^S_\omega\left(\bigcap_{k>n} [\mathcal{A}_{g,k} \notin A_\omega]\right) > 0$$

Then $\lim_{n\to\infty} p^S_\omega(\mathcal{A}_{g,n} \in A_\omega) < 1$.

Appendix: Proofs of major Lemmas

In the following two propositions, let m_ϵ be ϵG method defined as follows. Let $\langle w_a\rangle_{a\in A}$ be a vector of strictly positive real numbers, and $\epsilon : H \to \mathbb{R}^+ \cup \{0\}$ be the function $\epsilon(h) = \frac{1}{|h|^{|h_1|}}$. Let m_ϵ be the ϵG method defined as follows:

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

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

Proof Let S_{m_ϵ} be the isolated network consisting of one learner g employing m_ϵ. Let $a \in A$ and $n \in \mathbb{N}$. Define:

$$E_n = [\mathcal{A}_n = a]$$

Then by definition of the method m_ϵ action a is chosen on stage n is always played with probability at least $\frac{1}{|A|\cdot n} > 0$, regardless of history, by assumption the initial

weights are positive. It follows that: $p_\omega^{S_{m_\epsilon}}(E_n \mid \cap_{k\leq j<n} E_j^c) \geq \frac{1}{|A|\cdot n}$ for any pair of natural numbers n and k such that $k < n$. Hence, for all $k \in \mathbb{N}$:

$$\sum_{n>k} p_\omega^{S_{m_\epsilon}}\left(E_n \mid \cap_{k\leq j<n} E_j^c\right) = \infty$$

By the Borel Cantelli-Lemma, it follows that $p_\omega^{S_{m_\epsilon}}(E_n \text{ infinitely often }) = 1$. 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 3, g's estimates of the expected utility of each action approach the true expected utility in every possible state of the world with probability one, i.e.,

$$p_\omega^{S_{m_\epsilon}}\left((\forall a \in A)\lim_{n\to\infty} \hat{\mu}_{g,n}^a = \mu_\omega^a\right) = 1$$

Because m_ϵ plays the (estimated) EU maximizing actions with probability approaching one in every state of the world, it follows that:

$$p_\omega^{S_{m_\epsilon}}\left(\lim_{n\to\infty} m_\epsilon(h_n)(A_\omega) = 1\right) = 1.$$

□

By Lemma 4, 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) $p_{\omega_1}^S(h) > 0$. Suppose h has length n. Define:

$$E = [h] \cap \bigcap_{g\in G}\bigcap_{j>n} [\mathcal{A}_{g,j} \notin A_{\omega_1}]$$

$$F = [h] \cap \bigcap_{g\in G}\bigcap_{j>n} [\mathcal{A}_{g,j} \in A_{\omega_2}]$$

$$F_k = [h] \cap \bigcap_{g\in G}\bigcap_{n<j<n+k} [\mathcal{A}_{g,j} \in A_{\omega_2}]$$

Notice first that $F \subseteq E$, and so $p_{\omega_1}^S(F) \leq p_{\omega_1}^S(E)$. Thus, it suffices to show that $p_{\omega_1}^S(F) > 0$. Next notice that $F_1 \supseteq F_2 \supseteq \ldots F$, and so $\lim_{k\to\infty} p_{\omega_1}^S(F_k) = p_{\omega_1}^S(F)$.

Because m_ϵ chooses actions in $A \setminus \mathcal{A}^*_{g,n}(h)$ with probability at most $\frac{1}{|h|^2}$, it is easy to check, by induction on k, that

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

□

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

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

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. Fix $a_0 \in A$, and choose $q_{a_0,a} > 0$ however one pleases so that $\sum_{a\in A} q_{a_0,a} = 1$. Set $q_{a,a'} = q_{a_0,a'}$ for all $a, a' \in A$. Set the cooling schedule $T : H \to \mathbb{R}$ to be 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 analogous to those of Proposition 1 and 2. See Mayo-Wilson et al. (2010) for details. □

Proposition 4 *Let $\langle \Omega, A, O, p\rangle$ be a difficult learning problem with only finitely many actions. Let $\epsilon(h) = \frac{1}{|h|^{|h_1|}}$. Suppose $\phi(x) = x^r$, where $r \geq 2$, or $\phi(x) = \frac{1}{\lambda}e^{\lambda x}$, where $\lambda > 0$. Then the wRL method $m_{\phi,\epsilon}$ is* IC *but not* GIC.

Proof The proof that the wRL method $m_{\phi,\epsilon}$ is IC follows directly from Theorem 6.9 in Cesa-Bianchi and Lugosi (2006). To show that it is not GIC, let S be the connected strategic network consisting of exactly two learners g_1 and g_2 employing $m_{\phi,\epsilon}$. We show that the probability that an optimal action is *never* played $p^S_\omega((\forall n \in \mathbb{N})(\forall g \in G)\mathcal{A}_{g,n} \in A \setminus A_\omega)$ is strictly positive. Let g be either of the learners. Then:

$$\begin{aligned}
&p^S_\omega(\mathcal{A}_{g,n+1} = a)\\
&\quad= \sum_{h\in H_{g,n}} p^S_\omega(h) \cdot m_{\phi,\epsilon}(h)(a)\\
&\quad= \sum_{h\in H_{g,n}} p^S_\omega(h) \cdot \left[\frac{1}{|A|\cdot|h|^2} + \left(1 - \frac{1}{|A|\cdot|h|^2}\right)\frac{\phi'(\mathcal{R}^a_{g,|h|}(h))}{\sum_{b\in A}\phi'(\mathcal{R}^b_{g,|h|}(h))}\right]\\
&\quad\leq \sum_{h\in H_{g,n}} p^S_\omega(h) \cdot \left[\frac{1}{|A|\cdot|h|^2} + \left(1 - \frac{1}{|A|\cdot|h|^2}\right)\right] \text{as} \frac{\phi'(\mathcal{R}^a_{g,|h|}(h))}{\sum_{b\in A}\phi'(\mathcal{R}^b_{g,|h|}(h))} \leq 1
\end{aligned}$$

$$= \sum_{h \in H_{g,n}} p^S_\omega(h) \cdot \left[\frac{1}{|A| \cdot n^2} + \left(1 - \frac{1}{|A| \cdot n^2}\right)\right] \text{ by definition of } H_{g,n}$$

$$= \left(1 - \frac{1}{n^2}\right)\left(1 - \frac{1}{|A|}\right) \sum_{h \in H_{g,n}} p^S_\omega(h)$$

$$= \left(1 - \frac{1}{n^2}\right)\left(1 - \frac{1}{|A|}\right) \quad \text{as} \sum_{h \in H_{g,n}} p^S_\omega(h) = 1$$

Routine calculation then shows that:

$$p^S_\omega((\forall n \in \mathbb{N})(\forall g \in G)\mathcal{A}_{g,n} \in A \setminus A_\omega) \geq \Pi_{n \in \mathbb{N}} |A \setminus A_\omega| \cdot \left[\left(1 - \frac{1}{n^2}\right)\left(1 - \frac{1}{|A|}\right)\right]^2 > 0$$

□

In the following two propositions, let $m^{\delta\epsilon}_a$ be the $\delta\epsilon$ method determined by the quadruple $\langle w = \langle w_a \rangle, A' = \{a\}, \epsilon, \delta \rangle$ such that $\epsilon(h) = 0$, and $w_{a'} = 0$ for all $a' \in A$, and

$$\delta(h) = \begin{cases} 1 & \text{if } a \in \mathcal{A}^*_{g,n}(h) \\ \frac{1}{|h|} & \text{otherwise} \end{cases}$$

Proposition 5 *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 $[\mathcal{A}_n \notin A_\omega] \subseteq [\mathcal{A}_n = a]$. Define E to be the set of histories along which only the action a is played, i.e., $E = \bigcap_{n \in \mathbb{N}} [\mathcal{A}_n = a]$. 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. $a \in \mathcal{A}^*_{g,0}(\langle - \rangle)$. 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 6 *Let $\langle \Omega, A, O, p \rangle$ be any learning problem, and $M = \langle m^{\delta\epsilon}_a \rangle_{a \in A}$, where $m^{\delta\epsilon}_a$ is defined as in Proposition 5. 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 $E_n = [\mathcal{A}_{g,n} = a]$, so that $p^S_\omega(E_n \mid \cap_{k \leq j < n} E^c_j) \geq \frac{1}{n}$ for any pair of

natural numbers $k < n$ (by definition of $m_a^{\delta\epsilon}$). By the Second Borel-Cantelli Lemma, it follows that

$$p_\omega^S(g \text{ plays } a \text{ infinitely often }) = 1.$$

□

By Lemma 3, 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 $a \in A_\omega$, by the definition of the strategies $\{m_{a'}^{\delta\epsilon}\}_{a' \in A}$ and Lemma 4, it then follows that, almost surely, every learner in $N_G(g) \cap G'$ plays actions in A_ω with probability approaching one.

Continuing, by Lemma 5, 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.

Let $g'' \in G'$ be an agent such that g'' is a neighbor of some neighbor g' of g. Now we can repeat the argument above. Since g' plays some optimal action $a_{g'} \in A_\omega$ infinitely often almost surely, then by Lemma 3, 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 $\{m_{a'}^{\delta\epsilon}\}_{a' \in A}$ and Lemma 4, it then follows that g'' plays 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 $n \in \mathbb{N}$, 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, for all $g' \in G$, there is a finite path between g and g', and so we're done.

Appendix B: Proofs of Theorems

Proof of Theorem 1 That all RL strategies are IC under the assumptions of the theorem follows from Theorem 4, which is a trivial generalization of the proof of Theorem 1 in Beggs (2005). The proof that the wRL method $m_{\phi,\epsilon}$ of Proposition 4 is IC follows directly from Theorem 6.9 in Cesa-Bianchi and Lugosi (2006). 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, and for conditions characterizing when a wide class of SA methods are IC, see Bertsimas and Tsitsiklis (1993).

Proof of Theorem 3 This is an immediate consequence of Theorems 5 and 4.

Proof of Theorem 4 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 $\lim_{n\to\infty} p^S_\omega(\mathcal{A}_{g,n}(h) \in A_\omega) = 1$.

To do so, we adopt the proof of Theorem 1 in Beggs (2005) in the following way. As in Beggs' proof, it suffices to consider the 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 $\mathcal{W}^{a_i}_{g,n}$ and $\mathcal{W}^{a_i}_{g,n} - \mathcal{W}^{a_i}_{g,n-1}$. Because neighborhoods contain only finitely many agents by assumption, the assumptions of the theorem imply that the two variables 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 $\lim_{n\to\infty} p^S_\omega(\mathcal{A}_{g,n} \in A_\omega) \neq 1$.

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

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

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.

Weighted networks example: In this example, we show that the above proof that RL methods are not UC extends to weighted networks, so long as the weights remain

constant. Let S be any strategic network, and fix an agent g in S. Assume that g assigns each of her neighbors $f \in N_G(g)$ some normalized weight s_f so that $0 \leq s_f < 1$ and $\sum_{f \in N_G(g)} s_f = 1$. We assume that $s_g < 1$ so that the agent g does not completely disregard her neighbors findings. A similar example can be given when weights are not normalized.

Suppose g employs a RL method, and let $W = \sum_{a \in A} w_a$ be the sum of all initial weights assigned to each action according to g's RL method. The assumption that the network is weighted plays a role in g's calculations of reinforcements as follows. If $g's$ neighbor f obtains outcome o from playing action a on stage n, then instead of simply adding o to the reinforcement $\mathcal{W}^a_{g,n}$, the agent g adds $s_f \cdot o$, i.e., she weights the outcome. Let ω be any state of the world, and let k_2 be the value of any optimal action. Let $a_0 \notin A_\omega$ be any non-optimal action. As in the previous proposition, we assume payoffs are bounded from below by some constant $k_1 > 0$, and in particular, the payoff of a_0 is at least k_1 on each stage. Finally, suppose that all of g's neighbors always play the action a_0. By the following calculation, it follows that, for large enough n, the probability $p_n(a_0)$ that g plays action a_0 on stage n is bounded away from zero in ω, and since $a_0 \notin A_\omega$, it follows that RL methods are not UC in weighted networks.

$$\begin{aligned} p_n(a_0) &\geq \frac{w_a + k_1 n(1 - s_g)}{W + k_2 n s_g + k_1 n(1 - s_g)} \\ &\geq \frac{k_1 n s_g}{W + k_2 n s_g + k_1 n(1 - s_g)} \end{aligned}$$

For large enough n, we have $k_2 n(1 - s_g) > W$, and so:

$$\begin{aligned} p_n(a_0) &\geq \frac{k_1 n(1 - s_g)}{2k_2 n s_g + k_1 n(1 - s_g)} \\ &= \frac{k_1(1 - s_g)}{2k_2 s_g + k_1(1 - s_g)} \\ &> 0 \text{ as } s_g < 1 \end{aligned}$$

Proof of Theorem 5 Immediate from Propositions 1, 2, 3, and 4.

Proof of Theorem 6 Immediate from Propositions 5 and 6.

References

Argiento R, Pemantle R, Skyrms B, Volkov S (2009) Learning to signal: analysis of a micro-level reinforcement model. Stoch Process Appl 119(2):373–390

Agrawal R (1995) Sample mean based index policies with $O(\log n)$ regret for the multi-armed bandit problem. Adv Appl Probab 27(4):1054–1078

Auer P, Cesa-Bianchi N, Fischer P (2002) Finite-time analysis of the multiarmed bandit problem. Mach Learn 47:235–256

Bala V, Goyal S, (2008) Learning in networks. In: Benhabib J, Bisin A, Jackson MO (eds) Handbook of mathematical economics. Princeton University Press, Princeton

Baron J, Ritov I (2004) Omission bias, individual differences, and normality. Org Behav Hum Decis Process 94:74–85
Beggs A (2005) On the convergence of reinforcement learning. J Econ Theory 122:1–36
Berry DA, Fristedt B (1985) Bandit problems: sequential allocation of experiments. Chapman and Hall, chris
Bertsimas D, Tsitsiklis J (1993) Simulated annealing. Stat Sci 8(1):10–15
Bolton P, Harris C (1999) Strategic experimentation. Econometrica 67(2):349–374
Branke J, Meisel S, Schmidt C (2008) Simulated annealing in the presence of noise. J Heuristics 14(6):627–654
Cesa-Bianchi N, Lugosi G (2006) Prediction, learning, and games. Cambridge University Press, chris
Ellison G, Fudenberg D (1993) Rules of thumb for social learning. J Polit Econ 101(4):612–643
Hong L, Page S (2001) Problem solving by heterogeneous agents. J Econ Theory 97(1):123–163
Hong L, Page S (2004) Groups of diverse problem solvers can outperform groups of high-ability problem solvers. Proc Natl Acad Sci 101(46):16385–16389
Hopkins E (2002) Two competing models of how people learn in games. Econometrica 70(6):2141–2166
Hopkins E, Posch M (2005) Attainability of boundary points under reinforcement learning. Games Econ Behav 53(1):1105
Huttegger S (2011) Carnapian inductive logic for two-armed bandits
Huttegger S, Skyrms B (2008) Emergence of information transfer by inductive learning. Studia Logica 89:2376
Keller G, Rady S, Cripps M (2005) Strategic experimentation with exponential bandits. Econometrica 73(1):39
Kuhlman MD, Marshello AF (1975) Individual differences in game motivation as moderators of preprogrammed strategy effects in prisoner's dilemma. J Pers Soc Psychol 32(5):922–931
Mayo-Wilson C, Zollman K, Danks D (2010) Wisdom of the crowds vs. groupthink: connections between individual and group epistemology. Carnegie Mellon University, Department of Philosophy. Technical Report No. 187
Roth A, Erev I (1995) Learning in extensive-form games: experimental data and simple dynamic models in the intermediate term. Games Econ Behav 8:164–212
Skyrms B, Pemantle R (2004) Network formation by reinforcement learning: the long and medium run. Math Soc Sci 48(3):315–327
Stanovich KE, West RF (1998) Individual differences in rational thought. J Exp Psychol Gen 127(2):161–188
Zollman K (2009) The epistemic benefit of transient diversity. Erkenntnis 72(1):17–35