Theoretical Economics 14 (2019), 135–172 1555-7561/20190135

Learning dynamics with social comparisons and limited memory[1]

Juan I. Block
Faculty of Economics, University of Cambridge

Drew Fudenberg
Department of Economics, Massachusetts Institute of Technology

David K. Levine
Department of Economics, European University Institute and Department of Economics, Washington University in St. Louis

We study models of learning in games where agents with limited memory use social information to decide when and how to change their play. When agents observe only the aggregate distribution of payoffs and recall only information from the last period, aggregate play comes close to Nash equilibrium for generic games, and pure equilibria are generally more stable than mixed equilibria. When agents observe both the payoff distribution of other agents and the actions that led to those payoffs, and can remember this for some time, the length of their memory plays a key role: With short memories, aggregate play may not come close to Nash equilibrium unless the game satisfies an acyclicity condition. When agents have sufficiently long memory, generically aggregate play comes close to Nash equilibrium. However, unlike in the model where social information is solely about how well other agents are doing, mixed equilibria can be favored over pure ones.

Keywords. Social learning, Nash equilibrium, best-response dynamics, equilibrium selection.

JEL classification. C72, C73.

1. Introduction

In this paper, we develop and compare two models of learning in which agents who have access to social information and have limited memory decide when to adjust their strategies based on social comparisons. We consider large populations of agents who play the same game repeatedly but are "strategically myopic," meaning that they do not

Juan I. Block: jb2002@cam.ac.uk
Drew Fudenberg: drew.fudenberg@gmail.com
David K. Levine: david@dklevine.com

[1]This paper previously circulated under the title "Learning dynamics based on social comparisons."

We thank Glenn Ellison, Johannes Hörner, George Mailath, Salvatore Modica, Hamid Sabourian, and Larry Samuelson for helpful conversations, as well as Harry Pei for detailed comments. We acknowledge support from the Cambridge–INET Institute, the EUI Research Council, and NSF Grant 1643517.

© 2019 The Authors. Licensed under the Creative Commons Attribution-NonCommercial License 4.0. Available at http://econtheory.org. https://doi.org/10.3982/TE2626

attempt to influence how their current opponents will play in the future. We characterize conditions on social information and memory under which the learning procedures lead to Nash equilibria and provide equilibrium selection results in generic games.

Our interest in agents who use social information is motivated by the fact that individuals may learn from media sources, colleagues, and friends as well as from their personal experience. Of course, observing other people's financial success, social status, and even whether they are happy is a long way from observing the cardinal value of their realized utility, but we simplify here and suppose that the agents observe the payoffs of the best-performing other agents, perhaps because the payoffs in question are derived from underlying monetary income. We postulate that agents reexamine their strategies according to a simple social comparison rule: if they are getting close to the current highest payoff realized in their own population, then they are content and continue to play the same action; otherwise, they are discontent and experiment at random with different actions in hopes of doing better.

In addition to the random play of discontent agents, we impose three other behavioral assumptions prevalent in the literature. First, we assume that content agents tremble (or make a mistake) with a small probability; when this happens, they experiment with other actions and become discontent. As in Foster and Young (1990), Kandori et al. (1993), and Young (1993), we focus on the limits of the ergodic distributions of the system induced by our learning procedures as this mistake probability goes to 0. Second, we assume that agents only reassess their play with a probability bounded away from 1, as in Nöldeke and Samuelson (1993) and Samuelson (1994). Finally, as in Sandholm (2012), we assume there is a small (relative to the population) number of committed agents who play specific actions regardless of what they observe. This ensures that every action has positive probability even when there are no trembles.[2]

In our *low-information* model, agents observe the highest current utility realized in their own population without observing the corresponding actions. This restricted form of information seems plausible in many real world social interactions. It is also of relevance to laboratory experiments on large extensive form games, such as infinitely repeated games: Here it is feasible to tell participants the payoffs that other participants received in past plays of the repeated game, but not to tell them the exact strategies used by the participants who obtained high payoffs.[3] In this model, agents remember just whether they are content, and if so, what they did last period, and discontent agents randomize uniformly over all actions.[4] We show that all of the states with positive probability in the limit ergodic distribution—that is, the "stochastically stable" states—are

[2]Committed agents are also used in Heller and Mohlin (2017). We discuss this modeling choice further in Section 8.1.

[3]See, for example, the infinitely repeated prisoner's dilemma experiments surveyed in Dal Bó and Fréchette (2018). In many of these, a substantial minority of participants defect most or all of the time and receive a much lower overall payoff than participants who appear to be "conditionally cooperative," which raises the question of what would happened if participants were told something about the payoffs that others have received in previous plays of the repeated game.

[4]Fudenberg and Peysakhovich (2014) find experimentally that last period experiences have a larger impact on behavior than do earlier observations, and that individuals approach optimal strategies when provided with summary statistics (see also Erev and Haruvy (2016)).

those where all but the committed agents are getting about the same payoff in their own population. Since some committed agents always obtain the highest possible payoff, in generic games these limit states must be approximate Nash equilibria.

Under these assumptions, our equilibrium selection result is that the stochastically stable states are those where the largest number of shocks is required to lead the system to another equilibrium state; these numbers are the "radii" of the equilibria (Ellison (2000)). The intuition of this result is that once a sufficiently large number of trembling agents destabilize an equilibrium, some agents enter a search mode that eventually pushes everyone into search due to the limited social information and memory, allowing any equilibrium to be reached without further trembles. In other words, equilibrium selection boils down to the difficulty of escaping from each equilibrium. We find that while the radius of a pure strategy Nash equilibrium is generally large, growing linearly with the size of the population, every mixed equilibrium has radius 1, which in turn implies that stochastically stable states correspond to pure equilibria if they exist; otherwise, they correspond to all of the mixed equilibria. Moreover, using the methodology of Levine and Modica (2016), we are able to characterize the relative amount of time spent at different equilibria, and show that in large populations, the system spends less time at mixed strategy equilibria than at any of the pure strategy equilibria, including those that are not stochastically stable, and even when the mixed equilibrium gives the players a higher payoff.[5] Furthermore, we show that the same conclusion can hold when the noise in the system comes from noisy observation of others' payoffs as opposed to exogenous trembles.

In our *high-information* model, agents have more social information: they observe the highest current payoff in their own population together with the actions that yielded such payoff.[6] The agents also have more memory. Specifically, they recall the best-performing actions in the last T periods and their most recent action, in addition to whether they are content. When people can remember the strategies that worked well for others, it seems natural to mimic those strategies. We assume that discontent agents experiment randomly over remembered best responses and last period action instead of over all actions. If agents recall best responses only in the last period, then agents' behavior reduces to playing a myopic best response, and we show that our social learning process and the standard best response with inertia dynamic (Samuelson (1994)) predict the same stochastically stable set. In particular, both models can have stochastically stable cycles, and only pure Nash equilibria as stochastically stable states if the game is acyclic (Young (1993)).

We characterize the role of memory on stochastic stability of Nash equilibria by identifying a new acyclicity condition, which nests Young's widely used definition of acyclicity. We show that only Nash equilibria are stochastically stable if agents can recall best

[5] Fudenberg and Imhof (2006) characterize the relative frequencies of various homogeneous steady states in a family of imitation processes, but these processes can in some games spend most of their time near non-Nash states. As in our model, Levine and Modica (2013) examine the relative amount of time spent at different steady Nash states, but the dynamic is driven by group conflict rather than learning errors.

[6] We consider environments in which people have access to public records that aggregate information about payoffs as well as behavior. However, many institutions delete all records after a fixed period of time (due to storage costs or law), and record-keeping devices may deteriorate or become obsolete.

responses in the last $k \times l$ periods when the game is $k \times l$ *acyclic*, meaning that from any strategy profile there is a best-response path to a $k \times l$ CURB block. Because every game is $k \times l$ acyclic for k and l at least as large as the action spaces, when memory is sufficiently long, the high-information model leads to stochastic stability of approximate Nash equilibria in any game, even games that have only mixed equilibria. To understand this, note that play will eventually enter a CURB block of some specific size because the game is $k \times l$ acyclic. If some agents are not playing best responses, then all agents are pushed into search mode with positive probability. Since discontent agents recall and experiment among recent best responses, they eventually play the actions corresponding to a mixed equilibrium in the CURB block. Finally, we show by example that in the high-information model, mixed equilibria can be favored over pure equilibria.

The main methodological contribution of this paper is to characterize learning dynamics by combining the standard theory of perturbed Markov chains with the method of circuits, adapting Levine and Modica (2016) Theorem 9 to the case in which there is a single circuit. To illustrate the complementarity between this approach and past work, we show how to find the stochastically stable set by constructing circuits of circuits and, alternatively, by using Ellison's (2000) radius–co-radius theorem. Our results also contribute to the longstanding debate about pure versus mixed equilibria, and provide a clear connection between what agents observe and equilibrium selection.

Related literature

Our paper belongs to the larger literature that uses non-equilibrium adaptive processes to understand and predict which Nash equilibria are most likely to be observed. The literature on stochastic fictitious play (Fudenberg and Kreps (1993), Fudenberg and Levine (1998), Benaïm and Hirsch (1999), Hofbauer and Sandholm (2002)) concludes that stable equilibria can be observed while unstable equilibria cannot be, but also concludes there can be stable cycles.[7] Roughly speaking, stochastic fictitious play does not generate enough experimentation for only Nash equilibria to be stochastically stable. The same is true in the literature that studies deterministic best-response-like procedures perturbed with small random shocks (Foster and Young (1990), Kandori et al. (1993), Young (1993), Samuelson (1994)), although that literature does generically provide a way to select between strict equilibria in specific classes of games.

The idea that players observe outcomes and update play with probability less than 1 appears in the Nöldeke and Samuelson (1993) analysis of evolution in games of perfect information; our model differs in that agents are able to observe the average payoff and/or action distribution, not the outcomes of all matches for the current round of play.[8] The ideas that agents only change their actions if they are "dissatisfied" and/or

[7] Most of this work studies models where there is a single agent per side or where all agents receive the same information. Unlike these papers, Fudenberg and Takahashi (2011) study stochastic fictitious play in a model where, as here, there are many agents in each population and each agent has his/her own personal experience.

[8] As in our model, this stochastic observation technology means that every sequence of one-move-at-a-time intentional adjustments has positive probability; they use this to show that if a single state is selected as noise goes to 0, it must be a self-confirming equilibrium (Fudenberg and Levine (1993)).

that they have information about the distribution of payoffs have also been explored in the literature; these papers (Björnerstedt and Weibull (1996), Binmore and Samuelson (1997)) have assumed that agents receive information about the actions or strategies used by agents they have not themselves played. Our committed agents resemble the "nonconventional" agents proposed by Myerson and Weibull (2015) in that committed agents consider a (strict) subset of actions.

More recent literature (Foster and Young (2003, 2006), Young (2009), Pradelski and Young (2012)) considers learning procedures where players randomize when they are "dissatisfied."[9] The closest of these to our work is Pradelski and Young (2012), who consider a model in which agents can be watchful, content, hopeful, or discontent, and keep track of an aspiration utility based on their personal observation. Unlike in our models, this paper assumes that agents are more likely to adopt a new action the larger is the payoff improvement or payoff level. Their learning process selects an efficient equilibrium when pure equilibria exist; however, it can spend almost all the time at non-Nash efficient outcomes when they do not.

Our study of social comparisons is also related to Babichenko (2018) and Pradelski (2015), who analyze models of social influence in which agents' payoffs depend on an aggregate statistic, and agents observe and best respond to information about the actions played. Pradelski (2015) considers a model with heterogeneous agents playing a two-action coordination game and compares the stochastically stable sets when agents respond to the current distribution of actions to that when they respond to the time-averaged distribution. Similarly, Babichenko (2018) studies the long-run stability of approximate pure Nash equilibria when players choose a distorted best response to the current average of players' actions in continuous action games.

Our high-information procedure is related to Young (1993, 1998), and Hurkens (1995), who consider models with one agent in each player role, where the players observe and best respond to a random sample of the actions recently played, though these papers studies models with only one agent in each player role, a setting in which the strategic myopia assumption is harder to justify.[10] In these models, the observation of a random sample from the past actions serves as a source of noise, while in our high-information model, all actions for a fixed past horizon are observed. Their results differ substantially from ours when minimal CURB blocks contain only mixed equilibria or when they correspond to the entire game. More recently, Oyama et al. (2015) study a continuous time model with a continuum of agents, where agents respond to a finite and possibly small sample of current play. Most of their results concern games that can be solved by the iterated elimination of strategies that are not contained in the minimal p-dominant set for some $p > 1/2$ (Tercieux (2006)).

[9]These papers focus on when the procedure converges to Nash equilibrium with probability 1. Hart and Mas-Colell (2006) and Fudenberg and Levine (2014) have the same focus, but do not use the device of "dissatisfied states."

[10]The two papers differ in whether sampling is with or without replacement and in how beliefs are related to the sample that is observed. The role of large populations in justifying strategic myopia for learning in games was first mentioned by Fudenberg and Kreps (1993), who, however, analyzed a model with a single agent per side.

2. The population game

Let $G = ((u^j, A^j)_{j=1,2})$ be a two-player normal-form game where A^j is the finite set of actions for player j, $u^j : A^j \times A^{-j} \to \mathbb{R}$ is the utility function for player j, and $u^j(a^j, a^{-j})$ is player j's utility when choosing action $a^j \in A^j$ against the opponent playing $a^{-j} \in A^{-j}$. For any finite set X, we let $\Delta(X)$ denote the space of probability distributions over X. We extend u^j to mixed strategy profiles $\alpha \in \Delta(A^j) \times \Delta(A^{-j})$ in the usual way. For $\zeta \geq 0$, we say that $\hat{a}^j \in A^j$ is a *ζ-best response to* $\alpha^{-j} \in \Delta(A^{-j})$ if $u^j(\hat{a}^j, \alpha^{-j}) + \zeta \geq u^j(a^j, \alpha^{-j})$ for all $a^j \in A^j$.

We are interested in the *population* game generated when G is played by agents in two populations, indexed by i. Agent i of each population j chooses an action $a_i^j \in A^j$. There are N agents in each population, and agents are matched round-robin[11] against each agent of the opposing population. For any integer K and any set X, let $\Delta^K(X)$ be the subset of $\Delta(X)$ where each coordinate is an integer multiple of $1/K$. We want to deal with the population fractions that play different actions. We call $\Delta^N(A^j)$ the *grid for population j*; the *grid* is the product space $\Delta^N(A) = \Delta^N(A^1) \times \Delta^N(A^2)$. We also make use of the grids for subsets of the population. Aggregate play in population j can be represented by a mixed strategy $\alpha^j \in \Delta^N(A^j)$, where $\alpha^j(a^j)$ is the proportion of agents i playing $a_i^j = a^j$. The utility of agent i is $u_i^j(a_i^j, \alpha^{-j})$ since he plays each opponent in the opposing population in turn.

We make the following assumptions about payoffs.

Assumption 1. *For each player j and every $\alpha^{-j} \in \Delta^N(A^{-j})$, $\arg\max_{a^j \in A^j} u^j(a^j, \alpha^{-j})$ is a singleton.*

This assumption holds for generic payoff functions. It implies, in particular, that there is a unique best response to any pure action. Throughout the paper, we maintain this and all other numbered assumptions from the point they are stated.

Assumption 2. *No player j has an action $\hat{a}^j \in A^j$ such that $u^j(\hat{a}^j, \alpha^{-j}) \geq \max_{a^j \in A^j} u^j(a^j, \alpha^{-j})$ for all $\alpha^{-j} \in \Delta^N(A^{-j})$.*

This condition rules out games where one player has a pure strategy that weakly dominates all others.

Since a unique best response must be strict and action sets are finite, we may define $g > 0$ as the smallest difference between the utility of the best response and second best response to any pure strategy. Note that if $\zeta < g$, then there is a unique ζ-best response to every pure strategy. Moreover, in conjunction with Assumption 2, $\zeta < g$ also implies that there is no approximately dominant strategy. Formally, there is no player j and action $\hat{a}^j \in A^j$ such that $u^j(\hat{a}^j, \alpha^{-j}) + \zeta \geq \max_{a^j \in A^j} u^j(a^j, \alpha^{-j})$ for all $\alpha^{-j} \in \Delta^N(A^{-j})$.

[11] Equivalently, we may think of each agent playing against an average of the opposing population. This can be thought of as an approximation to a situation where each agent is randomly matched against the opposing population a substantial number of times. See Ellison et al. (2009) for conditions under which this approximation is valid.

3. Low-information social learning

In the low-information learning model, agents have no direct information about the behavior of others, but observe only the frequency of utilities in their own population. This model also assumes that agents remember only whether they are content and if so what action they played in the previous period.

The population game described above is played in every period $t = 0, 1, 2, \ldots$. An agent's *type* at the start of period t is $\theta_t^j \in \Theta^j \equiv A^j \cup \{0\} \cup \Xi^j$. In each population there is a fixed set Ξ^j of *committed* agents. An agent $\xi^j \in \Xi^j$ is committed to the action $a^j(\xi^j) \in A^j$. We assume that there is at least one agent committed to each action. We refer to the other agents as *learners* with type $A^j \cup \{0\}$. If $\theta_{it}^j \in A^j$, the learner is *content* with the action θ_{it}^j, and if $\theta_{it}^j = 0$, the learner is *discontent*. The process begins with an exogenous initial distribution of these types.

Committed agents play the action they are committed to and never change their type. Each learner trembles with independent probability ϵ, randomizing uniformly over all actions. A discontent learner randomizes uniformly even if he does not tremble, while a content learner who does not tremble plays $a_{it}^j = \theta_{it}^j$.[12] We assume the action choice made by trembling and discontent agents at the start of the period is held fixed throughout the round-robin.

A learner who trembled is discontent, $\theta_{it+1}^j = 0$. Each nontrembling learner has an independent probability $1 > p > 0$ of being *active* and a complementary probability $1 - p$ of being *inactive*. Inactive learners preserve their type, $\theta_{it+1}^j = \theta_{it}^j$. Given the population play α_t, let $U^j(\alpha_t^{-j}) \in \mathbb{R}^N$ denote the vector of utilities in population j corresponding to $u_i^j(a_{it}^j, \alpha_t^{-j})$ for each $a_{it}^j \in A^j$, and let $\phi^j(\alpha_t) \in \Delta^N(U^j(\alpha_t^{-j}))$ be the frequency of utilities of population j.[13] Let $\overline{u}^j(\phi^j(\alpha_t))$ be the highest time-t utility received in population j.[14] Let $\nu \geq 0$ be the *social comparison* parameter. If $u_i^j(a_{it}^j, \alpha_t^{-j}) > \overline{u}^j(\phi^j(\alpha_t)) - \nu$, the active learner is content, $\theta_{it+1}^j = a_{it}^j$; otherwise, he is discontent, $\theta_{it+1}^j = 0$.[15] Note that this social comparison allows the learner to determine whether he is playing a ν-best response, since there is always a committed agent playing a ν-best response. However, the learner cannot identify which actions are ν-best responses, as he does not see the actions played by others. Instead we assume that if an agent learns he is not playing a ν-best response, he chooses an action uniformly at random.

In summary, the play of the learners is governed by three parameters: the probability ϵ of trembling, the probability p of being active, and the tolerance for getting less than the current highest possible payoff ν.

[12] Notice that learners tremble whether they are discontent or not, but discontent learners play the same way whether they tremble or not.

[13] Let $\mathbf{A}^j(u^j, \alpha_t^{-j}) \subseteq A^j$ be the (possibly empty) subset of actions a_{it}^j for which $u_i^j(a_{it}^j, \alpha_t^{-j}) = u^j$. Then the time-$t$ frequency of utility level u^j is $\phi^j(\alpha_t)[u^j] = \sum_{a_{it}^j \in \mathbf{A}^j(u^j, \alpha_t^{-j})} \alpha_t^j(a_{it}^j)$.

[14] Agents observe the average payoff frequency of actions played, not the payoff frequency across matches.

[15] This is a very naive and non-Bayesian form of learning: active agents acquire information passively and make no effort to observe anything else.

4. Aggregate dynamics with low information

The behavior of individual agents gives rise to a Markovian dynamic. Let $\Phi_t^j \in \Delta^N(\Theta^j)$ be a vector of population shares of the player j types in period t. Define the (finite) *aggregate state space* $Z = \Delta^N(\Theta^1) \times \Delta^N(\Theta^2)$ to be the set of vectors $z = (\Phi^1, \Phi^2)$. We derive the exact formula for the *aggregate transition probabilities* $P_\epsilon(z_{t+1}|z_t)$ in Appendix A.1. Our interest is in studying this Markov process and how it depends on ϵ, the tremble probability of each learner. Let $\alpha^j(z) \in \Delta^N(A^j)$ be any feasible action profile of population j in state z. Since content learners all use pure strategies, but not all of those strategies need be the same, we say that a state z is *pure for population* j if all learners in population j have the same type, and that the state is *pure* if it is pure for both populations. Otherwise, we refer to it as a *mixed* state.

We start by identifying those states that correspond to approximate Nash equilibria and that are robust to the play of the committed agents. We refer to these as ζ-robust states.

Definition 1. For any number $\zeta \geq 0$, a state z is *ζ-robust* if all the learners from each population j are content and playing a ζ-best response to $\alpha^{-j}(z)$.

From now on, we set the tolerance level ζ to equal the social comparison parameter ν. Notice that the fact that in a ν-robust state the learners are playing a ν-best response to $\alpha^{-j}(z)$ and that the committed agents are playing their committed actions means that aggregate play of the learners corresponds to a ν-Nash equilibrium. However, the finiteness of the populations implies that some mixed equilibria can only be approximated and do not exactly correspond to the play of learners in a mixed ν-robust state, for example those mixed equilibria with irrational mixing probabilities. Define $M \equiv \max\{\#\Xi^1, \#\Xi^2\}$ to be the maximum number of committed agents in the two populations. To ensure the existence of ν-robust states, it suffices for the social comparison parameter ν to be greater than 0 and for M to be small relative to N. The proofs of the results for this section can be found in Appendix B.1 in the Supplemental Material, available in a supplementary file on the journal website, http://econtheory.org/supp/2626/supplement.pdf.

Lemma 1. *If $\nu > 0$, there is an η such that if $N/M > \eta$, a ν-robust state exists.*

We assume that the tolerance level ν is sufficiently small relative to g so that the ν-best responses indicated by the social comparison satisfy the implications of our assumptions on payoffs.

Assumption 3. *We have $\nu < g$.*

The next lemma says that the implications of our assumptions on (approximate) best responses still hold in the population game with committed agents, provided there are not too many of them relative to N.

LEMMA 2. *There is an η such that if $N/M > \eta$ and a^j is a strict best response to $a^{-j} \in A^{-j}$, then a^j is a strict best response to all $\alpha^{-j} \in \Delta^N(A^{-j})$ satisfying $\alpha^{-j}(a^{-j}) > 1 - M/N$. In particular, if a^j is the only ν-best response to $a^{-j} \in A^{-j}$ and $\nu < g$, then it is a strict best response to a^{-j}, so the same conclusion obtains.*

ASSUMPTION 4. *We have $N/M \geq \eta$, where η is large enough that Lemmas 1 and 2 hold.*

This assumption and Assumption 3 yield the following result.

LEMMA 3. *In any 0-robust state, the action profile of the learners must be a pure strategy Nash equilibrium, and any pure strategy Nash equilibrium corresponds to the play of learners in some 0-robust state.*

As shown by Lemma B.1 (in the Supplemental Material), P_ϵ is irreducible and aperiodic. This implies that for $\epsilon > 0$, the long-run behavior of the system can be described by a unique invariant distribution $\mu^\epsilon \in \Delta(Z)$ satisfying $\mu^\epsilon P_\epsilon = \mu^\epsilon$. We denote by μ^ϵ_z the (ergodic) probability assigned to state z. To characterize the support of the ergodic distribution as $\epsilon \to 0$, we use the concept of the *resistance* of the various state transitions. Because $P_\epsilon(z'|z)$ is a finite polynomial in ϵ for any z, z', P_ϵ is *regular,* meaning that $\lim_{\epsilon\to 0} P_\epsilon = P_0$ exists, and if $P_\epsilon(z'|z) > 0$ for $\epsilon > 0$, then for some nonnegative number $r(z,z')$, we have that $\lim_{\epsilon\to 0} P_\epsilon(z'|z)\epsilon^{-r(z,z')}$ exists and is strictly positive. We then write $P_\epsilon(z'|z) \sim \epsilon^{r(z,z')}$; let $r(z,z') \in [0,\infty]$ denote the resistance of the transition from z to z'. Moreover if $P_\epsilon(z'|z) = 0$, then this transition is not possible and we set $r(z,z') = \infty$, while if $P_0(z'|z) > 0$, we have $r(z,z') = 0$. A path $\mathbf{z}$ is a finite sequence of at least two not necessarily distinct states $(z_0, z_1, \ldots, z_t)$ and its resistance is defined as $r(\mathbf{z}) = r(z_0,z_1) + \cdots + r(z_{t-1},z_t)$. Notice that we allow for *loops* where some states are revisited along the path and that some transition probabilities are bounded away from zero independent of ϵ.

5. ANALYSIS OF THE LOW-INFORMATION MODEL

Our main goal is to characterize the long-run behavior of aggregate play. We show that in games with pure strategy equilibria, the ν-robust states with the largest radius (in the sense of Ellison (2000)) are most likely to be observed in the long run, and in games without pure strategy equilibria, all ν-robust states are about equally likely to be observed. To do this we first identify which transitions between states are most likely.

5.1 *Characterizing the least resistance paths*

The next lemma shows that if all the learners are currently playing a ν-best response, there is a zero resistance path to a ν-robust state in which they play the same way. To state this precisely, we define a partial ordering $\succeq$ over states. Let $D^j(z)$ be the number of discontent learners of player j in state z. Let $\overline{\alpha}^j(z) \in \Delta^{N-D^j(z)}(A^j)$ be the action profile corresponding to the content and committed types in z. We write $z \succeq z'$ if for $j = 1, 2$, $D^j(z) \geq D^j(z')$ and $\overline{\alpha}^j(z)$ is consistent with $\overline{\alpha}^j(z')$ in the sense that $(N - D^j(z'))\overline{\alpha}^j(z') = (N - D^j(z))\overline{\alpha}^j(z) + (D^j(z) - D^j(z'))\tilde{\alpha}^j$ for some action profile $\tilde{\alpha}^j \in \Delta^{D^j(z)-D^j(z')}(A^j)$. This says that we can get from z' to z by making some learners discontent.

LEMMA 4. *If $z \succeq \hat{z}$ and $\hat{z}$ is ν-robust, then there exists a zero resistance path (of length 1) $\mathbf{z}$ from z to $\hat{z}$.*

The proofs of this and the next two results can be found in Appendix A.2. The next lemma says that in calculating least resistance paths we may assume that discontent learners remain discontent. We refer to it as the *free to stay discontent* principle.

LEMMA 5. *For any path $\mathbf{z} = (z_0, z_1, \ldots, z_t)$ starting at any z_0, there is a path $\tilde{\mathbf{z}} = (\tilde{z}_0, \tilde{z}_1, \ldots, \tilde{z}_{\tilde{t}})$ with $\tilde{z}_0 = z_0$ and $r(\tilde{\mathbf{z}}) \leq r(\mathbf{z})$ satisfying the property that $\tilde{z}_\tau \succeq \tilde{z}_{\tau-1}$ and $\tilde{z}_{\tilde{t}} \succeq z_t$ for all $1 \leq \tau \leq t$.*

These two lemmas combined enable us to compute least resistance paths between ν-robust states by determining how many content learners must switch actions to move from one to the other and then computing the resistance to making those learners discontent. In effect it enables us to compute least resistance by counting the least number of trembles. We introduce a concept that captures the support of mixed strategy profiles that correspond to the play of content learners. More precisely, the *j width* of a state z denoted $w^j(z) \in \mathbb{Z}_+$ is the number of distinct types for content learners of player j. The *width* of a state z is $w(z) = w^1(z) + w^2(z)$. Observe that pure ν-robust states z have $w(z) = 2$.

Define a *proto-ν-robust* state z as a state in which all content learners from each population j are playing a ν-best response to $\alpha^{-j}(z)$. We divide these into three types: a *totally discontent* state in which $w(z) = 0$, so all learners of both players are discontent; a *semi-discontent* state in which all learners of one player are discontent but $w(z) > 0$, so at least one learner of the other players is content; a *standard* state in which at least one learner of each population is content. The next result characterizes transitions between states that involve proto-ν-robust states with the property that paths have no resistance.

LEMMA 6. *(i) If z is totally discontent, there is a zero resistance path to every ν-robust state.*

(ii) If z is proto-ν-robust but not totally discontent, there is a zero resistance path to a ν-robust state $\hat{z}$, and if z is standard, we can choose $\hat{z}$ so that $w(z) \geq w(\hat{z})$.

(iii) If z is not proto-ν-robust, there exists a zero resistance path to a state $\tilde{z}$ with $w(z) > w(\tilde{z})$.

5.2 *Absorbing states and approximate Nash equilibria*

Our first theorem shows that when there are no trembles, the ν-robust states are exactly the absorbing states, with all other states being transient.

THEOREM 1. *If $\epsilon = 0$, then every ν-robust state is absorbing and all other states are transient.*

PROOF. Start in a ν-robust state $z_t = z$. Since $\epsilon = 0$, no learner trembles and all agents do not change play. All learners remain content, regardless of being active or not, so $\theta^j_{it+1} = a^j_{it}$. The committed agents always keep their type by assumption. This implies that $z_{t+1} = z_t$ with probability 1, so z is absorbing.

If states are proto-ν-robust, there is a zero resistance path to a ν-robust state by Lemma 6(ii); otherwise, by Lemma 6(iii), there is a zero resistance path to a state with strictly less width. As long as the system does not reach a proto-ν-robust state, it has positive probability of moving along zero resistance paths to states with strictly lower width, applying parts (ii) and (iii) of Lemma 6, until it visits a proto-ν-robust state with $w > 0$ or reaches a totally discontent state, from which it has a positive probability of being absorbed at a ν-robust state as established in Lemma 6(i). □

Since there are a finite number of states, every state is either recurrent or transient, and when $\epsilon = 0$ the system will eventually be absorbed at a ν-robust state (and thus at a ν-Nash equilibrium). We consider both pure equilibria (as in Kandori et al. (1993) and Young (1993) Young (2009)) and mixed equilibria (similar to Foster and Young (2006), Hart and Mas-Colell (2006), Pradelski and Young (2012), and Fudenberg and Levine (2014)), though unlike those papers, we also characterize the stochastically stability of approximate equilibria.

5.3 *Characterization of the limit invariant distribution*

Our next result is a corollary that characterizes the relative frequency of different ν-robust states. Because the transition kernel P_ϵ is regular, Young (1993, Theorem 4) implies that as $\epsilon \to 0$, the ergodic distributions μ^ϵ have a unique limit distribution μ, which is one of the possibly many invariant distributions for P_0. A state z is *stochastically stable* if $\lim_{\epsilon\to 0} \mu^\epsilon_z > 0$ (Foster and Young (1990)). By Theorem 1, when ϵ is small but positive, the invariant distribution μ^ϵ puts almost all the probability on one or more ν-robust states. The *basin* of the ν-robust state z is the set of states for which there is a zero resistance path to z and no zero resistance path to some other ν-robust state z'.[16] We let r_z denote the *radius* of the ν-robust state z; this is defined to be the least resistance of paths from z to states out of its basin.

In characterizing the ergodic distribution μ_ϵ for small ϵ, we combine some standard technical tools and the more recent method of circuits developed by Levine and Modica (2016). Let $R(z, z')$ denote the least resistance of any path that starts at z and ends at z'. We say a set of ν-robust states Ω is a *circuit* if for any pair $z, z' \in \Omega$, there exists a least resistance *chain*, meaning a sequence $\mathbf{z} = (z_0, z_1, \ldots, z_t)$ with $z_0 = z$ to $z_t = z'$ with $z_k \in \Omega$ and $R(z_k, z_{k+1}) = r_{z_k}$ for $k = 0, \ldots, t-1$. That is, one of the most likely (lowest order of ϵ) transitions from z_0 is to z_1, one of the most likely transitions from z_1 is to z_2, and so forth. The next corollary follows directly from Theorem 9 in Levine and Modica (2016), which is specialized to the case where the only recurrent classes when $\epsilon = 0$ are singletons and there is a single circuit.

[16]Equivalently, the basin of the ν-robust state z is the set of starting states that lead to state z with probability 1 according to P_0.

COROLLARY 1. *If all ν-robust states are in the same circuit, then $\mu_z^\epsilon/\mu_{z'}^\epsilon \sim \epsilon^{r_{z'}-r_z}$ and, in particular, the set of stochastically stable states is exactly the ν-robust states with the largest radius.*

We sketch two proofs of the corollary to illustrate the complementarity between the methodologies of Ellison (2000) and Levine and Modica (2016). First we use Ellison's approach to show that the stochastically stable states are those that have the largest radius. For any target $z = z_t$, define the *modified resistance from* $z' = z_0$ to be $\mathrm{mr}(z', z) = \min_{\mathbf{z}=(z',z_1,\ldots,z)} R(z', z_1) + \cdots + R(z_{t-1}, z) - r_{z_1} - \cdots - r_{z_{t-1}}$ and the *modified co-radius* as $c_z = \max_{z'} \mathrm{mr}(z', z)$. If S is a union of recurrent classes, then the radius r_S is the least resistance path from S out of the basin of S, that is, to states where there is a positive probability of being absorbed outside of S. Define the *modified co-radius* c_S of a set of recurrent classes S to be the minimum over $z \in S$ of c_z. Ellison shows that a sufficient condition for a set S of ν-robust states to be stochastically stable is that $r_S > c_S$. If we let $\bar{r}$ denote the largest radius of any ν-robust state, then the set S of ν-robust states with radius $\bar{r}$ itself has radius r_S at least equal to $\bar{r}$. By assumption, all ν-robust states are in the same circuit, so we can compute an upper bound on c_S by considering, for each state $z' \notin S$, a least resistance chain from z' to z, meaning a sequence of states for which the resistance $R(z_k, z_{k+1}) = r_{z_k}$. The modified resistance of this chain is $\mathrm{mr}(z', z) = r_{z'}$ and since $r_{z'} < \bar{r} = r_S$, the conclusion follows.

For the sharper result that $\mu_z^\epsilon/\mu_{z'}^\epsilon \sim \epsilon^{r_{z'}-r_z}$, we use the method of Levine and Modica (2016). For any ν-robust state z, we consider *trees with root* z, where the nodes of the tree are all of the ν-robust states and the resistance of the tree is the sum of all the $R(z_k, z_{k+1})$, where z_{k+1} is the successor of z_k. Using the Markov chain tree formula (see, for example, Bott and Mayberry (1954)) it follows, as noted by Freidlin and Wentzell (1998), that $\log(\mu_z^\epsilon/\mu_{z'}^\epsilon)/\log\epsilon$ converges to the difference in resistance between the least resistance tree with root z and that with root z'. Notice that since each ν-robust state must be in the tree, the resistance of connecting that node is at least r_{z_k}, so that the least resistance tree cannot have less resistance than the sum of the radii of all nodes except the root. We now show there is a tree with exactly that resistance by building it recursively. Place the root z first. There must be some remaining node that can be connected to the tree at resistance equal to the radius because all stable states are in the same circuit. Add that node to the tree with that resistance. Continuing in this way, we eventually construct a tree in which the resistance is exactly the sum of radii of all but the root node. It follows that the difference in resistance between the least resistance tree with root z and root z' is exactly the difference in the radii, which is what is asserted in the corollary.

5.4 *Exact pure strategy equilibria and stochastic stability*

In this subsection, we characterize the stochastic stability of pure strategy Nash equilibria. To this end, we assume that pure strategy Nash equilibria exist and we set the social comparison parameter $\nu = 0$. Recall that every pure strategy Nash equilibrium corresponds to the play of learners in a 0-robust state.

Learners play a fundamental role in determining least resistance paths. On a path that moves away from a 0-robust state, content learners must tremble, and so the path

has positive resistance. In addition to the random mistakes, every active learner who is not playing a best response transitions to discontentment with no resistance irrespective of her current type. For each 0-robust state z, we define $r_z^j \in \mathbb{Z}_+$ for player j to be the least number of learners of player $-j$ that need to deviate for there to be a learner of player j that is not using a best response. Then in finding least resistance paths out of the basin of a 0-robust state z, we establish that the critical threshold to be considered is the smaller of r_z^1, r_z^2. We use this to characterize the radius of a 0-robust state z and we show that the minimum resistance to any other 0-robust state z' is the same for every z'.

THEOREM 2. *(i) If z is a 0-robust state, its radius is $r_z = \min\{r_z^1, r_z^2\} > 0$. Moreover, for any 0-robust state $\overline{z} \neq z$, there is a path from z to $\overline{z}$ that has resistance r_z.*

(ii) If z and z' are 0-robust states, then $\mu_z^\epsilon / \mu_{z'}^\epsilon \sim \epsilon^{r_{z'} - r_z}$ and, in particular, those states with the largest radius are stochastically stable.

PROOF. Consider a least resistance path $\mathbf{z}$ from a 0-robust state z to any 0-robust state $\overline{z}$. From Lemma 5 we know that there exists a path $\tilde{\mathbf{z}} = (\tilde{z}_0, \tilde{z}_1, \ldots, \tilde{z}_t)$ from $\tilde{z}_0 = z$ with $r(\tilde{\mathbf{z}}) \leq r(\mathbf{z})$ and $\tilde{z}_t \succeq \overline{z}$. Since $\tilde{z}_t \succeq \overline{z}$ and $\overline{z}$ is 0-robust, by Lemma 4 there is a zero resistance path from $\tilde{z}_t$ to $\overline{z}$. Then it is sufficient to compute $r(\tilde{\mathbf{z}})$ so as to obtain the radius of z.

We begin by characterizing the basin of the 0-robust state z. Lemma 5 implies it suffices to consider $D^j(\tilde{z}_\tau)$ for $\tau \leq t$, since discontent learners stay discontent on the path $\tilde{\mathbf{z}}$. If $D^j(\tilde{z}_\tau) < r_z$ for both players j, we show that $\tilde{z}_\tau$ is in the basin of z. Suppose that no learner trembles, and that discontents play the unique best response a^j in each population j, are active, and become content. This transition has no resistance. In the resulting state, all learners are content and playing a^j, the unique best response to α^{-j}; that is, the state is z. Hence, we have a zero resistance path back to z. However, to be in the basin, there must not be a zero resistance path to some different 0-robust state $\hat{z}$. We show that any such path starting at $\tilde{z}_\tau$ has a resistance of at least 1. Moving along any such path requires that all content learners of at least one player j play $\hat{a}^j \neq a^j$, by Assumption 1. Since $D^j(\tilde{z}_\tau) < r_z$ for $j = 1, 2$, all content learners are playing the best response, which implies that any transition $(\tilde{z}_\tau, z')$ on the path to $\hat{z}$ requires that $D^j(z') > D^j(\tilde{z}_\tau)$ for at least one player j. But in this transition, at least one content learner who is playing the best response becomes discontent, so this transition has resistance at least 1.

Next we establish that any path from z to any other 0-robust state $\hat{z}$ has resistance r_z. We show that if $D^j(\tilde{z}_\tau) \geq r_z^{-j}$ for either player j, then there exists a zero resistance path to any 0-robust state. Suppose that $D^j(\tilde{z}_\tau) \geq r_z^{-j}$ for one player j. Then consider a transition where no learner trembles and the action profile α is such that all content learners in population $-j$ are active and observe a better response played by a committed agent, and so become discontent, while learners in population j are inactive. This transition has zero resistance. The next transition has all learners not trembling and an action profile α so that contents in j are active, get a signal about a better response provided by a committed agent, and become discontent, while learners in $-j$ are inactive. It follows that this transition has no resistance. By Lemma 6 there is a zero resistance path to any 0-robust state. Hence part (ii) follows directly from Corollary 1. □

Intuitively, as long as the system remains within the basin of a pure equilibrium, not too many discontent agents are experimenting with new strategies, and the rest of the learners are content and playing a best response. Thus, from states in this basin the discontent learners are likely to find their way back to equilibrium. Interestingly, we find that once the system leaves the basin of a pure equilibrium, there must be lots of agents trying new strategies, which in turn pushes everyone into the state of searching. One key observation is that random search is relatively as likely to find one equilibrium as another, so all that matters is the rate at which the different basins are left. We show that this corresponds to the difference in radii, so that the system spends approximately $\epsilon^{r_{z'}-r_z}$ times as much time at the equilibrium z as at the equilibrium z'. Note that our equilibrium selection result is simple in that it requires only one to compute the radius r_z of each equilibrium.

5.5 *Stability of approximate and mixed strategy equilibria*

We now analyze the ergodic distributions and stochastically stable states in general finite two-player games, where pure strategy equilibria need not exist. We provide the complete structure of the transitions between equilibria. We show that the system starting at a mixed equilibrium either moves with resistance 1 toward mixed equilibria with smaller supports or transitions along resistance 1 paths to every equilibrium. Alternatively, we establish that if the system begins at pure equilibria, it transitions to every equilibrium.

We now set the social comparison parameter $\nu > 0$, since exact mixed strategy equilibria need not be attainable by population play represented on the grid $\Delta^N(A)$. In this case Lemma 2 ensures that a ν-robust state exists. As we observed above, in ν-robust states, aggregate play corresponds (modulo the play of the committed types) to an approximate equilibrium. That is, in any ν-robust state z, the action profile of the learners $\tilde{\alpha}$ is such that for every learner in each population j, $u_i^j(\tilde{a}_i^j, \alpha^{-j}) \geq u_i^j(a_i^j, \alpha^{-j}) - \nu$ for each $\tilde{a}_i^j$ in the support of $\tilde{\alpha}^j$ and all $a_i^j \in A^j$.

Note that, starting at a pure Nash equilibrium $\hat{a}$, as the play of learners in population $-j$ shifts to put increasingly more weight on actions other than $\hat{a}^{-j}$, eventually two things happen to the learners j best responses. First, additional actions may become ν-best responses to the play of population $-j$ in addition to $\hat{a}^j$ and, second, $\hat{a}^j$ will eventually no longer be a ν-best response to the play of the opposing population $-j$. When $\nu = 0$, the assumption of unique best responses on the grid (Assumption 1) assures that these two changes take place for exactly the same play of population $-j$. However, with $\nu > 0$, in general additional ν-best responses arise before $\hat{a}^j$ is no longer a ν-best response. Then it is possible that the system leaves a pure Nash equilibrium by having content agents trembling in both populations so that learners in each population have additional ν-best responses. When $\nu = 0$, this possibility does not exist: Since the point at which a critical number of learners in one population have a different best response already pushes the system out of the basin, it cannot be the least resistance path for content learners in both populations to tremble so that content agents in both populations

are not playing a best response. In the following discussion, we impose an assumption to rule out this possibility when $\nu > 0$ as well.

To rigorously describe the structure of the basin for any pure ν-robust state z with content actions corresponding to a given pure action a, we define $\underline{r}_z^j \in \mathbb{Z}_+$ for player j to be the least number of learners of player $-j$ that need to deviate so that a^j is no longer the *only* ν-best response to any feasible play of population $-j$. Similarly, let $\overline{r}_z^j \in \mathbb{Z}_+$ be the least number of learners of player $-j$ who must deviate for there to be a learner of player j who is not playing a ν-best response. Observe that $\overline{r}_z^j \geq \underline{r}_z^j$ and $N - \#\Xi^{-j} \geq \overline{r}_z^j, \underline{r}_z^j \geq 0$ for all j. For both j, if $\overline{r}_z^j > \underline{r}_z^1 + \underline{r}_z^2$, then "sidewise" escape from the equilibrium where learners tremble in both populations will have lower resistance than "direct" escape where learners tremble in only one population. However, since, for both j, $|\overline{r}_z^j - \underline{r}_z^j| \to 0$ as $\nu \to 0$, we can find conditions under which there is no sidewise escape with lower resistance than any direct escape.

LEMMA 7. *There is a $\chi > 0$ and $\gamma > 0$ with $N/M > \gamma$ and $\nu < \chi$ such that for every pure ν-robust state z, there is at least one j such that $\overline{r}_z^j \leq \underline{r}_z^1 + \underline{r}_z^2$, and $\underline{r}_z^j \geq 1$ for both j.*

We prove this result in Appendix B.3. We assume that the parameters ν and N/M are such that Lemma 7 holds, and we establish a separation between pure ν-robust states so that direct escape always has lower resistance.

ASSUMPTION 5. *We have $\nu < \chi$ and $N/M \geq \gamma$, where γ is large enough and χ is small enough that Lemma 7 holds.*

We next characterize the least resistance to leave the basin of a pure ν-robust state z in terms of the thresholds $\overline{r}_z^1$ and $\overline{r}_z^2$. Note that if a is a pure equilibrium, the least number of learners who must deviate before the original actions fail to be a best response increases linearly with N.

LEMMA 8. *The radius of a pure ν-robust state z is $r_z = \min\{\overline{r}_z^1, \overline{r}_z^2\}$, and if $\overline{z}$ is any ν-robust state, there is a path from z to $\overline{z}$ with resistance equal to r_z.*

The proof of this and the next lemma can be found in Appendix A.3. We introduce a notion that captures the largest mass of learners in the support of the current frequency of content actions: for any state z, let the *height* $h(z) \in \mathbb{Z}_+$ be the largest number of learners playing an action in the support of $\bar{\alpha}(z)$, the action profile that corresponds to the aggregate play of contents and committed agents.

We now determine the least resistance to leave the basin of mixed ν-robust states. In general, there are multiple mixed approximate equilibria in a neighborhood of mixed equilibria, so one might expect to move between those mixed approximate equilibria through one agent changing play at a time. The next lemma shows that, unlike the case of pure ν-robust states, the radius of mixed ν-robust states is 1 regardless of N, and that once the process leaves the basin of a mixed ν-robust state, it can move either with resistance 1 to another mixed ν-robust state with weakly smaller support or with resistance 1 to a pure ν-robust state.

LEMMA 9. *The radius of a mixed ν-robust state z is $r_z = 1$, and there is a path with resistance 1 either to every ν-robust state $\overline{z}$ or to a ν-robust state $\tilde{z}$ with $w(\tilde{z}) \leq w(z)$, and either $w(\tilde{z}) < w(z)$ or $h(\tilde{z}) > h(z)$.*

Equipped with these lemmas, we can determine which states are stochastically stable.

THEOREM 3. *For every pair z, z' of ν-robust states, $\mu_z^\epsilon/\mu_{z'}^\epsilon \sim \epsilon^{r_{z'}-r_z}$, so in particular the stochastically stable states are those with the largest radii.*

PROOF. The fact that all ν-robust states are connected by least resistance paths follows from Lemmas 8 and 9. The first conclusion follows from Corollary 1 and the second follows immediately from the first. □

A key implication of our characterization is the analysis of the relative likelihood of pure and mixed approximate equilibria.

COROLLARY 2. *If z is a mixed ν-robust state, z' is a pure ν-robust state, and N is large enough that $r_{z'} > 1$, then $\frac{\mu_z^\epsilon}{\mu_{z'}^\epsilon} \to 0$ as $\epsilon \to 0$.*

PROOF. Since $r_{z'}$ increases linearly with N by Lemma 8, choose N so that $r_{z'} > 1$. From Lemma 9 it follows $r_z = 1$. Then $\epsilon^{r_{z'}-r_z} \to 0$ as $\epsilon \to 0$. □

Thus, for large populations of interacting agents, we can conclude that in games with pure equilibria, the stochastically stable states must be pure ν-robust and, hence, the pure equilibria will be selected over mixed equilibria in the long run. Applying Theorem 8 of Levine and Modica (2016), we can also conclude that the system will spend on average more time in the part of the basin of the stochastically stable pure Nash equilibrium that excludes the equilibrium itself than it will at any nonstochastically stable Nash equilibrium. To see this, suppose that there exists a pure ν-robust state z. Note that moving from z to a state where one agent is discontent has resistance 1. During the period of time at the ν-robust state z before reaching another ν-robust state z', the ratio of time spent with one agent being discontent to the ν-robust state z is roughly ϵ, as there is zero resistance from one discontent agent to z and with large population N, the radius of z is larger than 1 so the bounds in Theorem 8 of Levine and Modica (2016) are tight. Consider another ν-robust state z' with smaller radius $r_{z'} < r_z$: the ratio of time spent at z' to the other stochastically stable state z is approximately $\epsilon^{r_z - r_{z'}}$, which is much smaller than ϵ since, again with large population, the difference in radii is considerably greater than 1.

6. HIGH-INFORMATION SOCIAL LEARNING

Our low-information learning procedure focused on agents who have very limited cross-section social information and very limited memory. Despite these limitations, we

found that learning leads agents to Nash equilibrium and provides an equilibrium selection result. We now consider the high-information model where agents both observe and remember more. In particular, we are interested in the interplay between cross-section social information about behavior and memory, and how it affects the learning dynamics.

6.1 *The learning procedure*

For ease of exposition, we present the high-information model by describing only the modifications we make to the low-information learning procedure. Regarding cross-section social information, now we assume that in every period t, learners observe the joint frequency of utilities and actions played in their own population:[17]

$$\Upsilon^j(\alpha_t)\big[u^j, a^j\big] = \begin{cases} \alpha_t^j(a_{it}^j) & \text{for } a_{it}^j \in \mathbf{A}^j\big(u^j, \alpha_t^{-j}\big) \\ 0 & \text{for } a_{it}^j \notin \mathbf{A}^j\big(u^j, \alpha_t^{-j}\big). \end{cases}$$

Concerning recall, we now assume that at the beginning of a period, learners can recall which actions were ν-best responses during the last finite $T \geq 1$ periods, and that all learners recall the last action they played, not only those learners who are content.[18] In this model, an agent's type is $\theta_t^j \in \Theta_T^j \equiv (A^j \times \{0,1\} \times T^{A^j}) \cup \Xi^j$. There is a given initial type distribution. We first describe the learners' type $A^j \times \{0,1\} \times T^{A^j}$. We use 0 to indicate the learner is discontent and use 1 to indicate the learner is content.[19] Thus, $A^j \times \{0,1\}$ gives the previous action taken a_{it-1}^j for both content and discontent learners. The final part that characterizes the learner's type $T_{it}^j[a^j]$ is the amount of time since each action a^j was observed to be a ν-best response to α_{t-1}^{-j}: $T_{it}^j[a^j] = 0$ if a^j was a ν-best response to α_{t-1}^{-j}; otherwise $T_{it}^j[a^j] = \min\{T, T_{it-1}^j[a^j] + 1\}$. Since this is the same for all learners of player j, we refer to this as the *common memory* of player j and refer to the actions a^j for which $T_{it}^j[a^j] < T$ as the players' *common memory set*, which we denote by $\mathbf{A}_T^j$. Let T be the *memory length*. This simplifies the description of the aggregate state of the resulting system, since we can use a single memory that is relevant for all learners of a player.[20] The *individual memory set* of learner i of player j is the common memory set and the last action that agent played, that is, $\mathbf{A}_i^j = \mathbf{A}_T^j \cup \{a_{it-1}^j\}$.

The impact of the memory set is only on the behavior of discontent learners. We assume they play uniformly only over their individual memory set $\mathbf{A}_i^j$ rather than over all actions A^j. Even though the behavior of the learners in this model can depend on their

[17] Recall that $\mathbf{A}^j(u^j, \alpha_t^{-j}) \subseteq A^j$ is the possibly empty subset of actions a_{it}^j for which $u_i^j(a_{it}^j, \alpha_t^{-j}) = u^j$.

[18] Note that actions that are best responses are those with the highest utilities, since agents observe the payoff to each possible action given the presence of committed types.

[19] Despite this modification, the transition rules concerning contentment do not change.

[20] We think of this common memory set as the amount of public information available to each population. As we discussed, the bounded memory assumption is motivated by the limitation of record-keeping devices: borrower's credit history is limited, insurance companies have access only to the most recent driving records that are cleared after a certain number of years, and information in informal markets is usually transmitted through word of mouth that naturally fades away.

memory, we use the same definition of a ν-robust state: It is a state where each learner is content and playing a ν-best response to the aggregate play of the other population. Therefore, the pure ν-robust states will still be the pure strategy Nash equilibria, and the mixed equilibria will correspond to a mixed approximate equilibrium.

Notice that our procedure differs from the formulation of Young (1993) where, in every period, only one agent per player role moves at that period and takes a size K random sample of play from the last T periods without replacement. Given this sample, certain actions are best responses, and only those actions have positive probability of being played. In contrast, our model allows agents to choose actions from the last T periods that were ν-best responses in the period in which they were used based on that period's cross-section information. Our model also differs from Young (1993, 1998), Hurkens (1995), Oyama et al. (2015), and related papers in that in our model agents do not take random samples.

6.2 *Equivalence between $T = 1$ and the best-response dynamics*

In this subsection, we restrict attention to the case at $T = 1$ and $\nu = 0$. The latter implies that learners choose exact best responses. Here Assumption 1 implies that for each population j, there is a single action a^j with $T^j_{it}[a^j] = 0$ and all the other actions $\tilde{a}^j \neq a^j$ have $T^j_{it}[\tilde{a}^j] = 1$: all actions except the best responses to the previous period's opposing population play are forgotten. Because $T = 1$, discontent learners randomize over their last period action and the current best response. This is similar to the two population version of the best-response-plus-mutation dynamic in Kandori et al. (1993). The specific version of their model that we focus on is called *best response with inertia* (see Samuelson (1994)): It assumes that in each period, with probability $1 > \lambda > 0$, each agent independently continues to play the same action as in the previous period, with probability $1 - \lambda - \epsilon$, they play a best response to the population distribution of opponent's actions, and with probability ϵ, they choose randomly over all possible actions. While in the one population case, the assumption that $\lambda > 0$ plays little role, as Kandori et al. (1993) show by example, it can lead to better behaved and more sensible dynamics in the two population case with results similar to those with one population.[21]

To make this comparison formally, we must extend the state space to incorporate the current population play. Let $\Phi^j_t \in \Delta^N(\Theta^j_T \times A^{-j})$ be a vector of population shares of the player j types in period t, which includes the play of the opposing population α^{-j}_{t-1} in period $t-1$. Both our dynamic and the best response with inertia dynamic are Markov processes on this extended state space. For compatibility with past work, we assume here (only) that $\#\Xi^j = 0$ for each population, that is, there are no committed agents, but rather that learners directly observe which actions are best responses. Since we assume that N/M is large, this is a reasonable approximation.

We now show that when $T = 1$ and $\nu = 0$, the high-information social learning and best response with inertia dynamic have the same recurrent classes and the same resistance transitions from one recurrent class to another, which in turn implies that the

[21] In the study of Markov chains this sort of inertia is called laziness and is used to turn periodic irreducible chains into aperiodic ones; it serves the same purpose here by ruling out limit cycles.

stochastically stable set and (for $\varepsilon > 0$) the ergodic probabilities of the recurrent classes are the same. In particular, from Lemma B.2 in Appendix B.4 (see also Samuelson (1994)), we know that only Nash equilibria are stochastically stable for acyclic games and $T = 1$.[22]

THEOREM 4. *High-information social learning with $T = 1$ is equivalent to best response with inertia in the sense that they have the same recurrent classes and the same least resistance between any pair of such classes.*

PROOF. Define z to be equivalent to z' if they have the same action distribution, and consider the equivalence classes $\{z\}$. In the best response with inertia dynamic, the non-action part of the state (subtypes and common memory sets) never changes, so given the initial condition, there is a unique point in each $\{z\}$ that will occur. This in turn implies that along the least resistance path from that unique point in $\{z_t\}$ to the unique point in $\{z_{t+1}\}$, the least resistance is given by taking all the actions that are not best responses to α_{t-1}^{-j} and the increase in the number of agents playing those actions by j summed for $j = 1, 2$. In high-information social learning with $T = 1$ dynamic regardless of the starting point in $\{z_t\}$, the least resistance over all targets in $\{z_{t+1}\}$ is exactly the same since agents who are not playing a best response to α_{t-1}^{-j} must have trembled: content and discontent agents play the unique best response to α_{t-1}^{-j}. Hence, if we have a recurrent class with respect to best response with inertia dynamics, a subset of the equivalence classes of states in that recurrent class are a recurrent class with respect to high-information social learning with $T = 1$ dynamics, and the least resistance between recurrent classes is the same for both dynamics. □

6.3 *Learning dynamics with T limited memory*

We next consider special classes of games in which the stochastic stability of Nash equilibria depends on the memory length. As the amount of memory increases, we can show stochastic stability of Nash equilibria under less restrictive conditions on the game, and if memory is long enough, we obtain stochastic stability for generic games.

Given the learners' memory set, it is convenient to define a *block* to be any set $W = W^1 \times W^2$ with nonempty subsets of actions $W^j \subseteq A^j$ for $j = 1, 2$ and define the associated *block game* G^W to be the original game G restricting payoffs and actions to the block W. A block W is closed under rational behavior (CURB) if $\arg\max_{a^j \in A^j} u^j(a^j, \alpha^{-j}) \subseteq W^j$ for every action profile $\alpha \in \Delta(A)$, where $\alpha^j(a^j) = 0$ for $a^j \notin W^j$, and every player j (Basu and Weibull (1991)). That is, a set of action profiles is CURB if it contains all best responses to itself. A CURB block is *minimal* if it contains no smaller CURB block. Define a *best-response path* to be a sequence of action profiles $(a_1, a_2, \ldots, a_t) \in (A^1 \times A^2)^t$ in which, for each successive pair of action profiles (a_k, a_{k+1}), only one player changes action, and each time the player who changes chooses a best response to the action the opponent played in the previous period. We now develop a notion of acyclicity in the spirit of Young (1993), but for movement to CURB blocks.

[22] Samuelson (1994) does not provide a proof of this, so we give one for completeness.

DEFINITION 2. A game G is $k \times l$ *acyclic* if, for every action profile $a \in A$, there exists a best-response path starting at a and leading to a CURB block W, with $\#W^1 = k$ and $\#W^2 = l$.

Notice that every game is $\#A^1 \times \#A^2$ acyclic since the entire game is a CURB block and that any 1×1 acyclic game is *acyclic* (Young (1993)). The following game is 2×2 acyclic but is not acyclic:[23]

	H	T	U	D
H	2, 0	0, 2	0, 0	0, 0
T	0, 2	2, 0	0, 0	0, 0
U	0, 0	0, 0	5, 5	8, 2
D	0, 0	0, 0	9, 1	2, 8

A more general class of $k \times l$ acyclic games that includes this example consists of $\#A^1 \times \#A^2$ games, where $\#A^1 = n \times k$ and $\#A^2 = m \times l$, with $k \times l$ blocks along the diagonal in which payoffs are strictly positive and, in each block, there is a unique mixed strategy equilibrium, and all other payoffs are zero. This class is similar to coordination games but with mixed equilibria on the blocks along the diagonal instead of pure strategy equilibria.

We next show that our learning procedure leads agents to equilibrium if more memory is combined with our weaker notion of $k \times l$ acyclicity; note that as memory grows, the requirement of $k \times l$ acyclicity is weakened. If we consider memory length equal to the largest CURB block, we obtain that high-information social learning without trembling converges with probability 1 to a Nash equilibrium for generic two-player games.

THEOREM 5. *If the game G is $k \times l$ acyclic, then, with memory length $T \geq k \times l$ and $\epsilon = 0$, ν-robust states are absorbing and other states are transient.*

PROOF. Starting at a ν-robust state z, since all learners are playing a ν-best response, all content learners remain content with their action, so such states are absorbing. We next prove that from any non-ν-robust state, there is a zero resistance path to a ν-robust state.

Pick any state z_t and suppose it is not ν-robust. Then there is zero resistance to a state z_{t+1} in which all learners of one population, say j, play the same action and are inactive, while one committed agent in population $-j$ plays the ν-best response a^{-j} to α_t^j, and all learners of population $-j$ are active and those learners who are not playing a ν-best response become discontent. From z_{t+1} there is zero resistance to a state z_{t+2} where learners of population j are inactive and hold their actions fixed, while all learners of population $-j$ play the same ν-best response a^{-j} to α_{t+1}^j in the common memory set. We proceed similarly, starting at z_{t+2} and moving without resistance to z_{t+3}. We assume learners in population $-j$ hold their play fixed and are inactive, whereas one committed agent in population j plays the ν-best response a^j to α_{t+2}^{-j}, and learners of player j are

[23]The game is not acyclic because there are two best-response cycles, but is 2×2 acyclic since from any action profile, either CURB block $\{H, T\} \times \{H, T\}$ or $\{U, D\} \times \{U, D\}$ can be reached along a best-response path.

all active and those not playing a ν-best response become discontent. Consider the zero resistance transition to state z_{t+4} in which learners in population $-j$ play the previous action and are inactive, while learners in population j all play the same best response a^j to a^{-j} in the memory set and are inactive. The resulting state z_{t+4} is pure.

Take any pure state z_t. Since the game is finite and $k \times l$ acyclic, the best-response path from this state goes to a $k \times l$ CURB block W in a finite number of steps. Notice that in the following transitions, when moving along the best-response path, we use only best responses to play in the previous period, so it suffices to have $T = 1$. First, a committed agent in one population, say j, plays a ν-best response a^j to the population play $-j$, all other learners play their previous actions, and all learners from population j are active, so those not playing a^j become discontent. In the next transition, all discontent learners of population j (who played the ν-best response a^j that belongs to the common memory set $\mathbf{A}_T^j$) are inactive. All learners in population $-j$ play the same actions as in the previous period and are active, and there is a committed agent in population $-j$ whose committed action a^{-j} is a ν-best response to the population j play α^{-j}, so the active learners in population $-j$ become discontent. We continue until the state is such that population play of learners corresponds to the $k \times l$ CURB block.

Start at z_t, where the population play of learners lies in a $k \times l$ CURB block W, and pick any $\mathbf{A}_T^j \subseteq W^j$ for each j with $T = k \times l$. In each population j, if all content learners are playing a ν-best response and for each j the common memory set $\mathbf{A}_T^j$ contains only actions that are ν-best responses to any feasible $\alpha^{-j}(z_t)$, then there is zero resistance to discontents choosing $a_{it}^j \in \mathbf{A}_T^j$: all learners are active and become or stay content, hence reach a ν-robust state. Otherwise, there exists at least one learner in one of the populations who is not playing a ν-best response to any feasible $\alpha^{-j}(z_t)$. Consider the transition where all learners play the same previous action, and in one population j those learners who are not playing a ν-best response are active and become or stay discontent because they observe a ν-better response played by some committed agent, which implies that $\#\mathbf{A}_T^j$ increases by 1 and that $\mathbf{A}_T^j \subseteq W^j$. If there are learners in population $-j$ who are not playing a ν-best response, we proceed to repeat the argument, which results in a larger memory set $\mathbf{A}_T^{-j} \subseteq W^{-j}$. Eventually, after $k \times l$ steps, we have not lost any relevant memory since $T = k \times l$, so all learners are discontent and we have expanded each memory set $\mathbf{A}_T^j$ to include all actions in the $k \times l$ CURB block W, which contains a ν-Nash equilibrium by definition. From there, there is zero resistance to a state where all discontents play the action profile corresponding to such equilibrium, and all learners are active and become content, therefore reaching the corresponding ν-robust state. □

The necessity of longer memory for aggregate behavior to approach Nash equilibria in acyclic games is familiar from stochastic best-response dynamics in which agents play a best response to samples from their memory, as in Young (1993); but this is not, however, sufficient to ensure convergence to equilibria in generic games. Intuitively, the relationship between $k \times l$ acyclicity and memory length captures that discontent learners randomize among recent best responses, which leads play to enter a CURB block of the appropriate size, which in turn leads to the result of stability of mixed equilibrium.

The following result shows that the radii of mixed ν-robust states can increase with N under high-information social dynamic, and that the support of those ν-robust states belongs to a CURB block that does not include all equilibria.

LEMMA 10. *If a CURB block does not contain all Nash equilibria, then there exists a constant $\kappa > 0$ such that the radius of the set of ν-robust states for which content learners play entirely within the CURB block is at least κN.*

In particular, this applies to a CURB block that does not contain all Nash equilibria, but does contain only one completely mixed equilibrium. We can also conclude from this lemma that for intermediate values of T, the equilibrium selection problem has two levels: first, selection among CURB blocks and then selection within the CURB blocks.

Two learning processes that examine the stochastically stability of set-valued equilibrium notions in generic games are related to our high-information model with long memory. Similar to Young (1993), Hurkens (1995) considers a learning model where a single player in each population is randomly selected to play the game, draws a sample of K observations with replacement from the set of the last T actions played by the opponent, and plays a best response to the sample. He shows that when T is large, ergodic sets correspond to minimal CURB blocks, and that if the game has a unique strict equilibrium, then for large enough histories, the probability that players are playing the equilibrium tends to 1. Yet if the two-player game has multiple minimal CURB blocks, trembles will not select a particular one. Building on this idea, Young (1998) extends Young's (1993) learning procedure to generic games. He finds that absorbing states correspond to minimal CURB blocks if K/T is sufficiently small, and that in the limit as ϵ vanishes, stochastically stable minimal CURB blocks are those that have minimal stochastic potential. However, the predictive power of such learning models is limited because minimal CURB blocks can be very large (i.e., the full strategy space) in many games.

7. EXAMPLES

In this section, we compare the equilibrium selection of high- and low-information models in two examples. We observe that when there are no committed agents (i.e. $\#\Xi^j = 0$ for $j = 1, 2$) and agents are able to directly observe the best responses, the computation of the radius and co-radius with the high-information model is exactly the same as for the best response with inertia dynamic.

EXAMPLE 1. Our first example illustrates that the low-information dynamic can select different equilibria than the high-information dynamic with low memory. Consider the game G_1:

	A	B	C	D
A	5, 5	0, 0	0, 0	0, 0
B	0, 0	10, 10	0, 9	9, 0
C	0, 0	9, 0	10, 10	0, 9
D	0, 0	0, 9	9, 0	10, 10

This game is 1×1 acyclic (Young (1993)), so from Theorem 4 and Lemma B.1, the limit invariant distribution for the high-information model with $T = 1$ contains only singleton pure Nash equilibria. There are four pure strategy equilibria (A, A), (B, B), (C, C), and (D, D). Initially we consider $\nu = 0$. We will show that (A, A) has the largest radius, so it is stochastically stable in the low-information model, yet in the high-information dynamic with $T = 1$, the equilibria (B, B), (C, C), and (D, D) are stochastically stable and have equal ergodic probability by symmetry.

We start by observing that to escape from (A, A) requires about $N/3$ of one population to tremble, say to B, so that is the radius of (A, A). Alternatively, to escape from (B, B), (C, C), or (D, D) requires only about $N/11$ of one population to tremble, so those are the radii of (B, B), (C, C), and (D, D). Hence with the low-information dynamic, (A, A) is stochastically stable according to Theorem 2, as it has the largest radius among pure strategy equilibria. That equilibrium would also be selected under under Young's dynamics. To analyze the best response with inertia dynamic, define S to be the union of the three equilibria (B, B), (C, C), and (D, D). The radius r_S of S is at least $N/2$ since if $1/2$ of one population is playing in either of the three equilibria (B, B), (C, C), or (D, D), one of those strategies must earn at least $(1/2)(6 + 1/3)$, while playing (A, A) yields no more than $5/2$. Alternatively, the co-radius of S is approximately $N/3$, since the maximum distance between any state and the basin of S is indeed the distance from (A, A) to the basin of S, because (A, A) is the only pure Nash equilibrium outside of S. Hence, by Ellison's theorem, the radius of S is larger than the co-radius, so S contains all stochastically stable states. In Appendix B.5 we use the results of Levine and Modica (2016) to show that this is still true when $T > 16$ and $\nu > 0$.

One of the reasons that the set S is stochastically stable under the best response with inertia dynamic is that when agents are at the equilibrium (B, B) and enough opponents switch to strategy C, agents' behavior gradually adjusts toward (C, C) because they can see that choosing C is the optimal strategy. Observing other agents' payoffs, but not their actions, allows the system to move from (B, B) to (A, A), and once it arrives at (A, A) to stay there for a long time.[24] ◊

EXAMPLE 2. In this example, we focus on how the stability of mixed equilibria depends on information conditions. Consider the game G_2:

	H	T	P
H	5, 3	3, 5	1, 1
T	2, 5	5, 2	1, 1
P	1, 1	1, 1	2, 2

This game is 3×3 acyclic and has three equilibria: the strict equilibrium (P, P), and two mixed equilibria $((\frac{3}{5}H, \frac{2}{5}T), (\frac{2}{5}H, \frac{3}{5}T))$ and $((\frac{3}{19}H, \frac{2}{19}T, \frac{14}{19}P), (\frac{2}{19}H, \frac{3}{19}T, \frac{14}{19}P))$. Suppose that there is a population of $N > 13$ agents of which 3 are committed to each action. Let $0 < \nu < 1$. We first observe that the set of action profiles for ν-robust states consists

[24] In Appendix B.5, we illustrate that the low-information dynamic can also predict a different equilibrium even when the best response with inertia dynamic has a singleton stochastically stable set.

of the state in which learners play (P, P), along with the sets of mixed approximate equilibrium profiles $\mathcal{B}$ and $\mathcal{C}$.[25]

Lemma 9 shows that ν-robust states that correspond to either $\mathcal{B}$ or $\mathcal{C}$ move along a path of resistance 1 to any other ν-robust state. We also know from Lemma 8 that ν-robust states in which learners play (P, P) may transition to any ν-robust state along a path of resistance $\lceil (N(1+\nu)-8)/5 \rceil$. Our characterization of the relative likelihood of different equilibria (Corollary 2) enables us to conclude that relatively $\epsilon^{1-\lceil (N(1+\nu)-8)/5 \rceil}$ times as long is spent at the pure ν-equilibrium as at either mixed ν-equilibrium. Since $N > 13$ and all mixed equilibria have a radius of 1, Corollary 2 says that the pure equilibrium is far more likely than the mixed equilibria in the long run: The fact that the mixed equilibria have radius 1 means a single experiment can shift the population away from them, and $N > 13$ implies that once a pure equilibrium is reached, it is relatively likely to stick.

Next consider the predictions of the high-information model with memory $T = 1$. We denote the block $\{H, T\} \times \{H, T\}$ by HT. Here we can easily show from the radius–co-radius argument that the block HT contains the stochastically stable set.[26] Within this set, play follows a deterministic best-response cycle; in particular, each outcome of the block game G_2^{HT} will have equal weight in the limit invariant distribution μ. By continuity of μ in λ, the agents' time average payoff for small λ is approximately 15/4. Consequently, agents obtain less than their minmax payoff, which is not a desirable property of a learning procedure (see, for example, Fudenberg and Kreps (1993), Fudenberg and Levine (1995)). Note that the game has two minimal CURB blocks HT and (P, P), so under the dynamics considered by Hurkens (1995) and Young (1998) both CURB blocks are ergodic sets. Yet Hurkens's dynamic does not provide a selection result, and the deterministic best-response cycle in the block HT is the unique stochastically stable set under Young's dynamic since it has minimal stochastic potential.

Finally, we consider the high-information model with large $T > 9$. The block HT still contains the stochastically stable set by the radius–co-radius argument as well as the equilibrium $((\frac{3}{5}H, \frac{2}{5}T), (\frac{2}{5}H, \frac{3}{5}T))$. By Theorem 5, the stochastically stable set is a subset of ν-robust states in a neighborhood of $((\frac{3}{5}H, \frac{2}{5}T), (\frac{2}{5}H, \frac{3}{5}T))$ and, therefore, the mixed equilibrium would be selected. Importantly, the limit set does not contain deterministic best-response cycles and agents receive more than the minmax payoff in the long run. ◊

[25] We have $\mathcal{B} = \{\alpha : |N^{-1}(N-3)(3\tilde{\alpha}^1(T) - 2\tilde{\alpha}^1(H)) + N^{-1}| < \nu, |N^{-1}(N-3)(3\tilde{\alpha}^2(H) - 4\tilde{\alpha}^2(T)) + N^{-1}| < \nu, \tilde{\alpha}^j(P) = 0\}$ and $\mathcal{C} = \{\alpha : |N^{-1}(N-3)(3\tilde{\alpha}^1(T) - 2\tilde{\alpha}^1(H)) + N^{-1}| < \nu, |N^{-1}(N-3)(2\tilde{\alpha}^2(H) + 4\tilde{\alpha}^2(T) - \tilde{\alpha}^2(P)) + 5N^{-1}| < \nu, |N^{-1}(N-3)(3\tilde{\alpha}^2(H) - 4\tilde{\alpha}^2(T)) + N^{-1}| < \nu, |N^{-1}(N-3)(4\tilde{\alpha}^2(H) + 2\tilde{\alpha}^2(T) - \tilde{\alpha}^2(P)) + 5N^{-1}| < \nu\}$, where $\tilde{\alpha}^j$ corresponds to the population play of content agents in j.

[26] To see this, the radius of the block HT is at least $2N/3$ because if 2/3 of one population is playing in the block HT, any of these strategies must earn at least $(2/3)(3 + 4/5)$, while playing (P, P) yields at most 4/3. But the co-radius of the block HT is about $N/5$, since at least 1/5 of one population has to mutate to escape from (P, P), which is the only pure Nash equilibrium outside of the block HT. Since the radius of the block HT is larger than its co-radius, the stochastically stable states are contained in the block HT.

8. Discussion and extensions

8.1 *Committed agents and round-robin*

We have assumed that there is a small fixed number of committed agents, but enough to play every action. We have also assumed agents play each other in a round-robin rather than a random matching process. We now discuss the role of these two assumptions and the consequences of relaxing them.

With respect to committed agents, as argued by Ellison (1994), it is plausible that, in large populations, some agents fail to learn either because they have limited cognitive capacity or idiosyncratic preferences. Moreover, as argued by Binmore and Samuelson (1997), one should not expect all kinds of mistakes to be negligible. Similarly, Sandholm (2012) argues that if there are agents for whom the environment is of particularly minor importance or who are engaged in multiple activities that place great demands on their attention or reasoning capacities, it seems reasonable to expect that some of these agents may not bother to consider switching strategies at all. More recently, Heller and Mohlin (2017) emphasize the importance of introducing committed agents into the study of community enforcement mechanisms.

With our assumptions on what agents observe, the committed agents act as a "flashlight" for the learners who are not playing a ν-best response by shining a light on the fact that there is some other strategy out there that gives a higher payoff. The flashlight enables learners to learn and become discontent when they are not playing a best response. Our results need this learning to happen substantially more often than learners becoming discontent when they are playing best responses. If the committed agents played randomly or all agents had a probability of experimentation bounded away from zero, this would eventually generate the necessary information. However, random play also makes payoffs a stochastic function of the intended play of the other population, which we avoid by having a fixed number of committed agents who each play a specific action.

In a similar way, if we replaced round-robin with some other random matching process, this would also make payoffs a stochastic function of the intended play of the other population. Moreover, random experimentation, random numbers of committed agents, or random matching are only some of the many possible sources of noise. We examine more carefully the role of the different types of noise next.

8.2 *Noisy information*

We start by focusing on a simple model to analyze different types of noise. In the analysis so far, there is no error in observing ν-best responses. To study the effect of noisy observation, we first provide a simplified variation on the original low-information model and we then modify the simplified model by introducing noisy observation of utility.

In the *simplified low-information* model, with probability p, one single learner of each population is chosen to be active and matched with one randomly chosen *comparison agent* from the same population after the round-robin. We now assume that after being matched with the comparison agent, the active learner has independent probability ϵ of being discontent. With complementary probability $1-\epsilon$, the active learner is

content with the current action if the comparison agent has lower utility and otherwise is discontent.

This model allows the possibility that an active learner who is not currently playing a best response becomes content without resistance if the comparison agent is playing a relatively worse action. However, as this probability will be bounded away from 1 due to the presence of committed agents, there is no cost to staying discontent, so the *free to stay discontent* principle still applies. (This follows since one cannot lower the resistance of the path constructed in the proof of Lemma 5 by having one learner accidentally become content.) Thus, a learner playing a best response becomes discontent with resistance ϵ in both the original model and this variation, so the stochastically stable set does not change.[27]

We now modify the simplified model to allow noisy observation of the comparison payoffs. We eliminate the exogenous probability that the active learner becomes discontent. In the population game, we replace the single round-robin tournament with K rounds of round-robin tournament against the opposing population, still holding fixed the actions of all agents. We assume that the selected active learners and their comparison agents are held fixed throughout the K rounds.

We introduce noise by assuming that in each round $\tau = 1, 2, \ldots, K$, the active learner i of player j with the comparison agent k observes his/her own utility $u^j(a^j_{it}, \alpha_t^{-j})$ and a noisy signal of the utility of the comparison agent. In this model, we view $\nu > 0$ as the parameter of a *perception threshold* that is captured by the function ψ. The threshold function is $\psi(u^j(a^j_i, \alpha^{-j}), u^j(a^j_i, \alpha^{-j}), \nu) = u^j(a^j_i, a^j_k)$ for $|u^j(a^j_i, \alpha^{-j}) - u^j(a^j_k, \alpha^{-j})| < \nu$ and is $\psi(u^j(a^j_i, \alpha^{-j}), u^j(a^j_k, \alpha^{-j}), \nu) = u^j(a^j_k, \alpha^{-j})$ otherwise. That is, the learner cannot distinguish any utility level that is within ν of his/her current payoff.[28] Then the noisy signal of agent k's utility is given by $\psi(u^j(a^j_{it}, \alpha_t^{-j}), u^j(a^j_{kt}, \alpha_t^{-j}), \nu) + \eta^j_{\tau t}$. This form of the noise corresponds to the case where the noise impacts the learner after application of the perception threshold, so that the probability an agent who is playing ν-best responses gets a false signal of a ν-better action is independent of just how close to the maximum payoff that agent's payoff is. As we explain below, the model is different if the noise is applied to the observation before testing if it passes the perception threshold, so that the agent's signal is $\psi(u^j(a^j_{it}, \alpha_t^{-j}), u^j(a^j_{kt}, \alpha_t^{-j}) + \eta^j_{\tau t}, \nu)$ (instead of $\psi(u^j(a^j_{it}, \alpha_t^{-j}), u^j(a^j_{kt}, \alpha_t^{-j}), \nu) + \eta^j_{\tau t}$).

We assume that the random shocks $\eta^j_{\tau t}$ are independent and identically distributed (i.i.d.) over time and across actions, independent of the payoffs, with zero mean, have support on the entire real line, and have a twice continuously differentiable moment generating function. Moment generating functions are always log concave; we

[27] Although only one learner can be active at a time in the new model, the number of active learners plays no role in the original setup. The fact that a learner who becomes discontent plays uniformly at the beginning of the next period likewise plays no role. Hence, while the resistance of paths can be different in the two models, the resistance of many events defined in terms of collections of states remains unchanged; particularly, the resistance of the ratios of ergodic probabilities described in Theorem 2 as well as the waiting times described in Levine and Modica (2016).

[28] There is a large literature in psychology that studies agents who have limited ability to perceive small differences. In economics, the idea of perception thresholds has been explored by, for example, Fishburn and Trotter (1999), Rayo and Becker (2007), and Salant and Rubinstein (2008).

strengthen this slightly by assuming that the second derivative of the logarithm of the moment generating function is strictly negative. We also assume that both populations face the same distribution of payoff shocks.

Finally, the active learners' type is determined by a modified pairwise comparison: the active learners compare their own utility with that of the average comparison signal over the K rounds. That is, for $\nu > 0$, the active learner is discontent if

$$u^j\big(a_{it}^j, \alpha_t^{-j}\big) + \nu < \frac{1}{K}\sum_{\tau=1}^{K}\big(\psi\big(u^j\big(a_{it}^j, \alpha_t^{-j}\big), u^j\big(a_{kt}^j, \alpha_t^{-j}\big), \nu\big) + \eta_{\tau t}^j\big)$$

and is content otherwise. This can be written as

$$\frac{1}{K}\sum_{\tau=1}^{K}\eta_{\tau t}^j > \nu + u^j\big(a_{it}^j, \alpha_t^{-j}\big) - \psi\big(u^j\big(a_{it}^j, \alpha_t^{-j}\big), u^j\big(a_{kt}^j, \alpha_t^{-j}\big), \nu\big),$$

that is, the active learner is discontent if the observation error is large relative to the perceived utility difference.

Our interest is in the case where K is large, so the probability of being discontent when playing a ν-best response is small. To this end, take $K = -\log\epsilon$. We must now determine the least resistance to a learner who is playing a ν-best response being discontent for each given configuration. Let $\mathcal{L}[x]$ denote the logarithm of the moment generating function of $\eta_{\tau t}^j$.[29] Assume that learner i is playing a ν-best response and let $r(a_{it}^j, a_{kt}^j, \alpha_t^{-j})$ denote the resistance of an active learner playing a_{it}^j for whom the comparison agent is playing a_{kt}^j, both against α_t^{-j} to being discontent. The large deviations theorem from probability theory (see Theorem I.4 of Den Hollander (2008)) shows that

$$r\big(a_{it}^j, a_{kt}^j, \alpha_t^{-j}\big) = \min_x\big[\mathcal{L}[x] - \big(\nu + u^j\big(a_{it}^j, \alpha_t^{-j}\big) - \psi\big(u^j\big(a_{it}^j, \alpha_t^{-j}\big), u^j\big(a_{kt}^j, \alpha_t^{-j}\big), \nu\big)\big)x\big].$$

Since the learner is playing a ν-best response $\nu + u^j(a_{it}^j, \alpha_t^{-j}) - \psi(u^j(a_{it}^j, \alpha_t^{-j}), u^j(a_{kt}^j, \alpha_t^{-j}), \nu) \geq 0$, this resistance is minimized over the comparison agents' actions when $u^j(a_{it}^j, \alpha_t^{-j}) - u^j(a_{kt}^j, \alpha_t^{-j}) = 0$, so that the least resistance to a learner becoming discontent when playing a ν-best response is $\min_x[\mathcal{L}[x] - \nu x]$, a positive constant.[30]

Since which events have zero resistance and which have positive resistance have not changed, the *free to stay discontent* principle still applies, and so all equilibria still lie in a single circuit. Moreover, their relative ergodic resistances are still computed by differences in the radii between the equilibria. Observe that a constant probability of becoming discontent gives the same quantitative result as the original model, that is, the radius is computed by counting the least number of deviations to reach the edge of the basin for each population. Thus, the conclusion of Theorem 3 still applies with the new values of the radii. Here the radii of mixed equilibria are bounded uniformly in the

[29]Note that in the case of normal errors, this function is quadratic.

[30]Note that the least resistance is positive, because of our assumption that $\nu > 0$, but sufficiently small. If $\nu = 0$ and the observational errors have symmetric distribution around zero, then the probability that observational error exceeds zero is 1/2, so people would be discontent almost all the time.

population size, while the radii of the pure equilibria are linear in population, so in large populations only pure equilibria are stochastically stable.

This analysis of noisy payoff observation relies on the assumption that the noise is independent of the state, so that the resistance to being discontent when playing a ν-best response is independent of the state as well. There are at least three reasons this might not be true:

(i) The matching may not be exactly a round-robin. For example, with uniform random matching, a learner will receive noisy observations unless all of their opponents play the same action.[31]

(ii) If the number or the play of committed agents is random, the variance of the payoffs that a learner observes will depend on the state.

(iii) If the learner's perception threshold is random or if noise impacts the signal before the threshold test is applied, then learners who obtain payoffs that are farther from the threshold are more likely to get false signals. In this case, the probability of becoming discontent depends on the distribution of payoffs among ν-best responders in the particular state.

Our computation of modified radii depends on the circuit structure. The three sorts of complications above can change the resistances, and noise that breaks ties in the resistances can change the circuit structure. Nevertheless, the modified radii are robust in the sense that they are a continuous function of the underlying resistances.[32] Thus, we expect that our results will extend if the overall noise from all of these sources is approximately independent of action and population. As Ellison et al. (2009) pointed out, the order of limits does matter. For our results to go through, we consider first sending the matching procedure noise, committed agents noise, and threshold effect noise to zero (the rate or order in which we do this does not matter). Then taking the limit as ϵ (either the probability of trembling or the analogous measure in the noisy simplified low-information model) becomes vanishingly small. In particular with this order of limits it is still the case that in sufficiently large populations, the pure equilibria are stable and the mixed ones are not. This may not be true if we take $\epsilon \to 0$ before one of the other limits.

8.3 *Performance of the learning rules*

We conclude by showing that the learning rules we study do well in environments in which the system spends most of the time at some approximate Nash equilibrium.

[31] For a formal result about when this sampling error has negligible impact on the stochastically stable set in a related model, see Ellison et al. (2009).

[32] To see this we observe that the modified radius of a stable state can also be computed as the resistance of the least resistance tree for which the root is that stable state: the resistance of a least resistance tree must be continuous in the underlying resistances regardless of whether the structure of the tree changes when the underlying resistances change. That is, there is a finite number of trees over which the minimum is taken, and the resistance of each of those trees is continuous in the underlying resistances; hence, the minimum is continuous as well.

Specifically, in such environments, no agent could improve his expected time average payoff by more than ν by using a different learning procedure, given the play of the other agents. This is true even when the alternative learning procedures use any amount of information, including knowing in advance what the agents of the other player are going to do.[33]

Formally, in a state z, agent i's learning rule gives expected utility $U_i(z)$ that depends only on z. Given the state z, there is a unique probability distribution $\pi^{-j}(z)[\alpha^{-j}]$ over $\alpha^{-j} \in \Delta^N(A^{-j})$. Suppose that action distributions α^{-j} of the opposing population are drawn from $\pi^{-j}(z)$, and that the agent i observes the outcome α^{-j} and chooses a best response to it. Let $V_i(z)$ be the corresponding expected utility with respect to $\pi^{-j}(z)$.[34] Let $\overline{u}$ denote the largest difference between any two utilities in the game. Taking expectations with respect to P_ϵ and letting S denote the stochastically stable set, we compute

$$\limsup_{\epsilon\to 0}\limsup_{\tau\to\infty}\frac{1}{\tau}\mathbb{E}\sum_{t=1}^{\tau}\big(V_i(z_t)-U_i(z_t)\big)\leq \nu+\big(1-\mu_S^\epsilon\big)\overline{u}.$$

The reason for this is simply that z_t is at a ν-robust state except for a fraction of the time $(1-\mu_S^\epsilon)$, and when it is at a ν-robust state, $U_i(z_t)$ cannot do more than ν-worse (for the learners) than any strategy regardless of how it is learned. Put differently, if the agent knew that agents of the other population were going to follow a stationary strategy for very long periods of time τ (where τ depends on ϵ) and that committed agents in their own population were going to reveal what the agents of the other population are doing, despite their limited memory and information, the agent could not do much better than either our low-information learning procedure or our high-information learning procedure with large T.

9. Conclusion

In many settings, people have aggregate information about the payoffs and/or behaviors of others, and may use this information to help select their strategies. Most people also have bounded memory. We have considered two learning models that incorporate these ideas, showed that aggregate play comes to approximate Nash equilibria, and related the amount of social information and memory used to which equilibria we should expect to see in the long run.

In our low-information model, agents observe aggregate information about how well others are doing, but not how they obtained those payoffs, so agents are not able to directly imitate successful actions. Here we assume that agents use their limited memory to remember their own most recent action and its payoff, together with a "search state" that indicates that there might be better actions with which to experiment. We demonstrated that pure strategy equilibria should be expected to be seen a larger fraction of

[33]This is not a "universal consistency property" (see, e.g., Hannan (1957), Fudenberg and Levine (1995), and Hart and Mas-Colell (2000)) since it depends on the fact that the other agents are also using the same learning procedure.

[34]No learning rule using any information can do better than this.

the time than mixed strategy equilibria when people cannot easily see which actions did well. We then used several examples to compare the predictions of our learning model to those of the best response with inertia dynamic.

Our high-information social learning model supposes that people observe aggregate information about how well and what others did, which might describe some sorts of consumption and financial decisions, and that when people experiment, they use actions that performed well recently. When people recall only the last action and approximate best responses, we found that our learning dynamic predicts the same stochastically stable states as best response with inertia, and so can be trapped in cycles in the long run. When agents have more memory, cycles become improbable, and mixed strategy equilibria can be relatively more stable than pure strategy equilibria.

If we think of greater information and greater memory as corresponding to greater sophistication, we can summarize our results in the following way: In a game with both mixed and pure equilibria, low sophistication leads to pure equilibria, while high sophistication can lead to either pure or mixed equilibrium, depending on the game. Intermediate degrees of sophistication may not lead to any equilibrium at all.

Which of these models is a better description for how people learn to play Nash equilibria will, of course, depend on the information available to the agents and to the cognitive effort they put into processing it. Neither model should be expected to apply literally to a wide spectrum of situations, but we hope they will provide a useful complement to the widely used best-response dynamic in making predictions about long-run social outcomes. We believe that it would be interesting to explore our learning models in controlled laboratory experiments because our results establish sharp predictions depending on observability and memory.

Appendix

A.1 *Description of the aggregate state process*

The Markov process describes the evolution of the states z that correspond to population shares of types. This aggregate-level process is generated by a micro-level process that describes the evolution of the agent states that correspond to the types of individual agents. Define the (finite) *agent state* $x = (x^1, x^2) \in (\Theta^1)^N \times (\Theta^2)^N$ to be an assignment of types to agents. An agent state x induces population shares of player types (Φ^1, Φ^2); it is *consistent* with a state z if the shares match those in z, in which case we write $x \in X(z)$.

To determine the aggregate transition probability $P_\epsilon(z_{t+1}|z_t)$ from z_t to z_{t+1} start by choosing an agent state $x_t \in X(z_t)$. For any $x_{t+1} \in X(z_{t+1})$, we define the *agent–state transition probability* $P_\epsilon(x_{t+1}|x_t)$ and we then compute $P_\epsilon(z_{t+1}|z_t) \equiv \sum_{x_{t+1}\in X(z_{t+1})} P_\epsilon(x_{t+1}|x_t)$.[35] Let $D^j(x_t)$ be the number of discontent learners of population j in x_t and let $\mathcal{C}(x_t)$ be the set of content learners in x_t. Let $\mathcal{T}^j$ denote the trembling learners of player j and let $\mathcal{N}^j$ be the nontrembling learners. Let $\mathcal{R}^j \subseteq \mathcal{N}^j$ be the active learners. Denote an assignment of actions to all agents by $\sigma^j \in (A^j)^N$.

[35]This is well defined since while $P_\epsilon(x_{t+1}|x_t)$ depends on which $x_t \in X(z_t)$ is chosen, the sum does not. If we permute the names in x_t and the names in x_{t+1} the same way, then the agent–state transition probability is unchanged.

LEMMA A.1. *The aggregate transition probabilities are given by*

$$P_\epsilon(z_{t+1}|z_t) = \sum_{x_{t+1}\in X(z_{t+1})} \sum_{\mathcal{T},\sigma,\mathcal{R}} \underbrace{\prod_{j=1,2} \epsilon^{\#\mathcal{T}^j}(1-\epsilon)^{\#\mathcal{N}^j}\left(\frac{1}{\#A^j}\right)^{\#(\mathcal{T}^j\cap\mathcal{C}(x_t))+D^j(x_t)} p^{\#\mathcal{R}^j}(1-p)^{\#\mathcal{N}^j-\#\mathcal{R}^j}}_{\equiv P(\mathcal{T},\sigma,\mathcal{R},x_{t+1}|x_t)}$$

if σ^j is feasible with respect to $\mathcal{T}^j$ and x_t, that is, if it is consistent with the play of the nontrembling content and committed types, and if $x_{t+1}\in X(z_{t+1})$; otherwise $P_\epsilon(z_{t+1}|z_t)=0$.

PROOF. The determination of $P_\epsilon(x_{t+1}|x_t)$ has several steps involving interim variables. The probability of a given set of tremblers and nontremblers is $\epsilon^{\#\mathcal{T}^j}(1-\epsilon)^{\#\mathcal{N}^j}$. Choose any $\sigma^j\in(A^j)^N$. Such an action assignment has probability defined as $\Gamma^j(x_t,\mathcal{T}^j)[\sigma^j]$ that is calculated below. Given σ^j and the corresponding α_t, we compute the frequency of payoffs $\phi^j(\alpha_t)$. For the nontremblers $\mathcal{N}^j$ and each subset $\mathcal{R}^j\subseteq\mathcal{N}^j$ of active nontremblers, there is probability $p^{\#\mathcal{R}^j}(1-p)^{\#\mathcal{N}^j-\#\mathcal{R}^j}$ that this subset of learners is active.

Now we use these interim variables to compute the transition probabilities. If $i\notin\mathcal{R}^j$, then $\theta^j_{it+1}=\theta^j_{it}$. If $i\in\mathcal{R}^j$ and $u^j_i(a^j_{it},\alpha_t^{-j})>\overline{u}^j(\phi^j(\alpha_t))-\nu$, then $\theta^j_{it+1}=a^j_{it}$; otherwise $\theta^j_{it+1}=0$. We also compute feasible strategy profiles conditional on $\mathcal{T}^j$. Let $\overline{\alpha}^j(x_t,\mathcal{T}^j)\in\Delta^{\#\Xi^j+\#(\mathcal{N}^j\cap\mathcal{C}(x_t))}(A^j)$ be the strategy profile corresponding to the play of the committed and nontrembling content types in x_t.[36] A strategy profile $\alpha^j\in\Delta^N(A^j)$ in x_t is *feasible with respect to* $\mathcal{T}^j$ if $N\alpha^j=(\#\Xi^j+\#(\mathcal{N}^j\cap\mathcal{C}(x_t)))\overline{\alpha}^j(x_t,\mathcal{T}^j)+(D^j(x_t)+\#(\mathcal{T}^j\cap\mathcal{C}(x_t)))\tilde{\alpha}^j$ for some strategy profile $\tilde{\alpha}^j\in\Delta^{D^j(x_t)+\#(\mathcal{T}^j\cap\mathcal{C}(x_t))}(A^j)$.[37] In particular, , $\overline{\alpha}^j(z_t)\equiv\overline{\alpha}^j(x_t,\varnothing)$ be the strategy profile corresponding to the aggregate play of content and committed agents in state z_t which is well defined since $\overline{\alpha}^j(x_t,\varnothing)$ is independent of $x_t\in X(z_t)$, and define $\mathcal{A}^j(z_t)$ to be the set of all corresponding feasible α^j. Finally, let $\mathcal{T}=(\mathcal{T}^1,\mathcal{T}^2)$, $\mathcal{R}=(\mathcal{R}^1,\mathcal{R}^2)$, and $\sigma=(\sigma^1,\sigma^2)$.

We compute the joint conditional probability $P(\mathcal{T},\sigma,\mathcal{R},x_{t+1}|x_t)$ of the terminal agent state x_{t+1} and the interim variables $\mathcal{T}$, σ, and $\mathcal{R}$ considering two sets of events. In the first case, if σ^j is not feasible given $\mathcal{T}^j$ and x_t or if $x_{t+1}\notin X(z_{t+1})$, this probability is zero. Observe that the nontrembling content learners play the last period action and all other learners play uniformly; this implies that $\Gamma^j(x_t,\mathcal{T}^j)[\sigma^j]=(1/\#A^j)^{D^j(x_t)+\#(\mathcal{T}^j\cap\mathcal{C}(x_t))}$. Then for the other case, the probability is given by

$$P(\mathcal{T},\sigma,\mathcal{R},x_{t+1}|x_t)=\prod_{j=1,2}\epsilon^{\#\mathcal{T}^j}(1-\epsilon)^{\#\mathcal{N}^j}\left(\frac{1}{\#A^j}\right)^{D^j(x_t)+\#(\mathcal{T}^j\cap\mathcal{C}(x_t))}p^{\#\mathcal{R}^j}(1-p)^{\#\mathcal{N}^j-\#\mathcal{R}^j}.$$

Now we can compute $P_\epsilon(x_{t+1}|x_t)=\sum_{\mathcal{T},\sigma,\mathcal{R}}P(\mathcal{T},\sigma,\mathcal{R},x_{t+1}|x_t)$. □

[36]Where the nontrembling content agents play the action corresponding to their type and the committed types play their committed action.

[37]That is, if it is consistent with the play of the nontrembling content and committed types.

Next we formally show that an active learner who is doing well will never get a signal that suggests he is doing poorly, so these learners become discontent only when they tremble.

LEMMA A.2. *If σ^j is feasible with respect to $\mathcal{T}^j$, and some content learner $i \in \mathcal{R}^j$ is playing an a^j_{it} that is a ν-best response to α_t^{-j} and $\theta^j_{it+1} \neq a^j_{it}$ in x_{t+1}, then $P(\mathcal{T}, \sigma, \mathcal{R}, x_{t+1}|x_t) \leq \epsilon$.*

PROOF. Since $i \in \mathcal{R}^j$ is content and playing a ν-best response to α_t^{-j}, it cannot be that $u^j_i(a^j_{it}, \alpha_t^{-j}) \leq \overline{u}^j(\phi^j(\alpha_t)) - \nu$. Hence, learner i must either remain content with a^j_{it} or must have trembled: in the latter case, the whole transition has probability at most ϵ. □

A.2 *Proofs for Section 5.1*

Since $P_\epsilon(z'|z)$ is defined as a sum, and the terms in the sum are of the form $P(\mathcal{T}, \sigma, \mathcal{R}, x_{t+1}|x_t)$, it is sufficient when analyzing resistance to look for a target $x_{t+1} \in X(z_{t+1})$ and realizations $\mathcal{T}$, σ, and $\mathcal{R}$ for which the probability $P(\mathcal{T}, \sigma, \mathcal{R}, x_{t+1}|x_t)$ has the least resistance. Denote this resistance as $r(x_t, x_{t+1})$. For it to be finite, σ^j must be feasible given $\mathcal{T}^j$ for $j = 1, 2$, in which case the resistance is equal to the number of trembles, $r(x_t, x_{t+1}) = \#\mathcal{T}^1 + \#\mathcal{T}^2$. In particular, to show that the aggregate resistance is zero, it is sufficient to find an agent state resistance for the transition that has resistance zero.

PROOF OF LEMMA 4. Let $x_t \in X(z)$ and $z_t = z$. Since $z \succeq \hat{z}$ and $\hat{z}$ is ν-robust, we have for each j that $N\bar{\alpha}^j(\hat{z}) = (N - D^j(z))\bar{\alpha}^j(z) + D^j(z)\tilde{\alpha}^j$ for some $\tilde{\alpha}^j \in \Delta^{D^j(z)}(A^j)$. This implies that $\mathcal{A}^j(\hat{z}) \subseteq \mathcal{A}^j(z)$; hence, if $\alpha^j_t \in \mathcal{A}^j(\hat{z})$, then $\alpha^j_t \in \mathcal{A}^j(z)$, and $\alpha^j_t \in \mathcal{A}^j(\hat{z})$ implies that all learners are playing ν-best responses in α^j_t. Then there is zero resistance to none of the learners trembling and all learners being active, so all become or stay content with a^j_{it}. The resulting agent state x_{t+1} therefore satisfies $x_{t+1} \in X(\hat{z})$ and, by construction, the resistance of this transition is zero. □

The next lemma will be used in the proof of Lemma 5.

LEMMA A.3. *If $\mathbf{z} = (z_0, z_1, \ldots, z_t)$ is a path, then there exists a path $\tilde{\mathbf{z}} = (\tilde{z}_0, \tilde{z}_1, \ldots, \tilde{z}_{\tilde{t}})$ with $\tilde{z}_0 = z_0$ and $\tilde{z}_{\tilde{t}} = z_t$ with $r(\tilde{\mathbf{z}}) \leq r(\mathbf{z})$, and agent states $\tilde{x}_\tau \in X(\tilde{z}_\tau)$ for $\tau = 0, 1, \ldots, \tilde{t}$ that have transitions between $\tilde{x}_{\tau-1}$ and $\tilde{x}_\tau$ in which no discontent learner trembles and every content learner, including those who tremble, plays the action with which they are content.*

PROOF. First observe that we can replace the discontent learners who tremble with discontent learners who play the same way and who are inactive, and strictly lower the resistance, so there is a path to the target with no greater resistance if no discontent learner ever trembles. To show that we can have every content learner playing the same action, we replace each transition $z_\tau, z_{\tau+1}$ with two transitions $z_\tau, \tilde{z}_{2\tau+1}, z_{\tau+1}$. Let $x_\tau \in X(z_\tau)$ together with $\mathcal{T}_\tau, \sigma_\tau, \mathcal{R}_\tau, x_{\tau+1} \in X(z_{\tau+1})$ have resistance $r(z_\tau, z_{\tau+1})$. For the transition $z_\tau, \tilde{z}_{2\tau+1}$, choose the same x_τ, set $\tilde{\mathcal{T}}_\tau = \mathcal{T}_\tau$, and set $\tilde{\sigma}_\tau$ such that all content learners play the remembered action, $\tilde{\sigma}^j_t$ is consistent with $\overline{\alpha}^j(x_\tau, \varnothing)$, and all learners are inactive.

Then $r(x_\tau, \tilde{x}_{2\tau+1}) = r(x_\tau, x_{\tau+1})$ so that $r(z_\tau, \tilde{z}_{2\tau+1}) \leq r(x_\tau, x_{\tau+1}) = r(z_\tau, z_{\tau+1})$. For the transition $\tilde{z}_{2\tau+1}, z_{\tau+1}$ take $\tilde{\mathcal{T}}^j_{2\tau+1} = \varnothing$, $\tilde{\sigma}_{2\tau+1} = \sigma_\tau$, and $\tilde{\mathcal{R}}_{2\tau+1} = \mathcal{R}_\tau$ so that the terminal state is $x_{\tau+1} \in X(z_{\tau+1})$ and $r(\tilde{x}_{2\tau+1}, x_{\tau+1}) = 0$, implying $r(\tilde{z}_{2\tau+1}, z_{\tau+1}) = 0$ and concluding that $r(z_\tau, \tilde{z}_{2\tau+1}) + r(\tilde{z}_{2\tau+1}, z_{\tau+1}) \leq r(z_\tau, z_{\tau+1})$. □

Proof of Lemma 5. If $r(\mathbf{z}) = \infty$, for any $\tilde{x}_0 \in X(z_0)$ and any $\tilde{t} = 1$, take $\tilde{x}_1$ to have all learners discontent, $D^j(\tilde{z}_\tau) = N - \#\Xi^j$ for both j, and note that $r(\tilde{\mathbf{z}}) < \infty$ since we may have all learners tremble. It follows that $\tilde{z}_1 \succeq \tilde{z}_0, z_t$.

Next suppose that $r(\mathbf{z}) < \infty$. We may assume from Lemma A.3 that in $\mathbf{z}$, the least resistance transitions have agent transitions in which no discontent learner trembles and every content learner plays the action they recall. We will now find a path with $\tilde{t} = t$ and prove that if $\tilde{z}_\tau \succeq z_\tau$, we can find a state satisfying $\tilde{z}_\tau \succeq z_\tau, \tilde{z}_{\tau-1}$ and $r(\tilde{z}_\tau, \tilde{z}_{\tau+1}) \leq r(z_\tau, z_{\tau+1})$. To do this, use the fact that $\tilde{z}_\tau \succeq z_\tau$ to order the learners of each player j so that the first $N - D^j(\tilde{z}_\tau) - \#\Xi^j$ agents in $\tilde{x}_\tau \in X(\tilde{z}_\tau)$ have exactly the same type as the first $N - D^j(\tilde{z}_\tau) - \#\Xi^j$ learners in $x_\tau \in X(z_\tau)$. Observe that $r(z_\tau, z_{\tau+1})$ is determined by a particular target $x_{\tau+1} \in X(z_{\tau+1})$ and realizations $\mathcal{T}_\tau$, σ_τ, and $\mathcal{R}_\tau$, and that $r(z_\tau, z_{\tau+1}) = \#\mathcal{T}^1_\tau + \#\mathcal{T}^2_\tau$ since σ_τ is feasible as we have assumed a finite resistance path. Denote by $\mathcal{A}^j(z)$ the set of feasible $\alpha^j \in \Delta^N(A^j)$ such that $N\alpha^j = (N - D^j(z))\bar{\alpha}^j(z) + D^j(z)\tilde{\alpha}^j$ for some action profile $\tilde{\alpha}^j \in \Delta^{D^j(z)}(A^j)$. Because $\tilde{z}_\tau \succeq z_\tau$, we have $\mathcal{A}^j(z_\tau) \subseteq \mathcal{A}^j(\tilde{z}_\tau)$ and the realization σ_τ is feasible for $\tilde{x}_\tau$, so we set $\tilde{\sigma}_\tau = \sigma_\tau$. We also define $\tilde{\mathcal{R}}_\tau$ to be $\mathcal{R}_\tau$ applied only to those learners who are content in $\tilde{x}_\tau$, that is, discontent learners are inactive, but content learners are active if and only if the corresponding learner was active in $\mathcal{R}_\tau$. Now let $\tilde{\mathcal{T}}_\tau$ be $\mathcal{T}_\tau$ applied to those learners who are content in $\tilde{x}_\tau$. Given $\tilde{\sigma}_\tau$, $\tilde{\mathcal{T}}_\tau$, and $\tilde{\mathcal{R}}_\tau$, take $\tilde{x}_{\tau+1} \in X(\tilde{z}_{\tau+1})$ to be the corresponding agent state. Then $r(\tilde{z}_\tau, \tilde{z}_{\tau+1}) = \#\tilde{\mathcal{T}}^1_\tau + \#\tilde{\mathcal{T}}^2_\tau \leq \#\mathcal{T}^1_\tau + \#\mathcal{T}^2_\tau = r(z_\tau, z_{\tau+1})$ since $\mathcal{T}_\tau$ applies to every learner to whom $\tilde{\mathcal{T}}_\tau$ applied. By construction no learner is content in $\tilde{x}_{\tau+1}$ unless she has the same type as in $\tilde{x}_\tau$ so certainly $\tilde{z}_{\tau+1} \succeq \tilde{z}_\tau$. Also by construction, every learner who is content in $\tilde{x}_{\tau+1}$ has the same type as the corresponding learner in $x_{\tau+1}$, so indeed $\tilde{z}_{\tau+1} \succeq z_\tau$. □

Proof of Lemma 6. Suppose $z_t = z$ is totally discontent and $\hat{z}$ is ν-robust. Take $x_t \in X(z)$ and action assignment σ_t in which $\alpha^j_t \in \mathcal{A}^j(\hat{z})$. This is feasible since $\mathcal{A}^j(\hat{z}) \subseteq \mathcal{A}^j(z)$ for $j = 1, 2$. Suppose next that the transition does not involve any learner trembling and has all learners being active. Since $\hat{z}$ is ν-robust the learners are all playing a ν-best response and, hence, have zero resistance to becoming content. The resulting state $x_{t+1} \in X(\hat{z})$, so the process reaches $\hat{z}$ with zero resistance and showing part (i).

Now consider a proto-ν-robust state $z_t = z$ that is not totally discontent with $w(z) > 0$. Let population j have at least one content learner so $w^j(z) \geq 1$. Since z is proto-ν-robust and $w^j(z) \geq 1$, one content learner in j plays an action $\hat{a}^j$ that is a ν-best response to $\alpha^{-j}(z)$. Take any $x_t \in X(z)$, and consider the following zero resistance transition to z': In population j, learners do not tremble and are active, content learners play the last period action, and discontent learners play the action $\hat{a}^j$; in population $-j$ learners do not tremble, play the same actions as the previous period, and are inactive. For the next transition, we consider two cases. Suppose first that there is no content learner in population $-j$, that is, $w^{-j}(z) = w^{-j}(z') = 0$. By Lemma 2, there is an M/N

such that $\hat{a}^{-j}$ is a strict best response to $\alpha^j(z')$ with $\alpha^j(\hat{a}^j) > 1 - M/N$. Along the transition from z' to $\hat{z}$ suppose in population j nobody trembles and all learners are inactive, while in population $-j$ all learners do not tremble, and discontent learners play $\hat{a}^{-j}$ and are active. In the resulting state $\hat{z}$ all learners are content and playing a ν-best response, and $w(\hat{z}) > w(z)$. If instead $w^{-j}(z) = w^{-j}(z') = 1$, the content learner in $-j$ is playing the ν-best response $\hat{a}^{-j}$ to $\alpha^j(z')$. Then, in the transition from z' to $\hat{z}$, assume learners in population j do not tremble and are inactive, and all learners in population $-j$ do not tremble, and discontent learners play $\hat{a}^{-j}$ and are active. The resulting state $\hat{z}$ is ν-robust with $w(z) \geq w(\hat{z})$. By construction, unless z was semi-discontent, we did not increase the width, which is claimed in part (ii).

Finally, to show part (iii), suppose that $z_t = z$ is not proto-ν-robust with $w(z) > 0$. Then in at least one population j there is at least one content learner with a_i^j that is not a ν-best response to some $\alpha^{-j} \in \mathcal{A}^{-j}(z)$. Pick any $x_t \in X(z)$. There is zero resistance to having population $-j$ play α^{-j} if no learner trembles, all learners are inactive, and discontent learners play the same action; it does not also add resistance to this transition if one committed agent of player j plays a ν-better response than a^j and play of population j corresponds to some $\alpha^j \in \mathcal{A}^j(z)$. Moreover, there is zero resistance when all learners of player j do not tremble so the learners in state a^j become discontent. Then $x_{t+1} \in X(z_{t+1})$ with $w(z_{t+1}) < w(z)$. □

A.3 *Proofs for Section 5.5*

Proof of Lemma 8. Let $\mathbf{z}$ be a least resistance path from a pure ν-robust state z to any ν-robust state $\overline{z}$. Lemma 5 implies there is a path $\tilde{\mathbf{z}} = (\tilde{z}_0, \tilde{z}_1, \ldots, \tilde{z}_t)$ from $\tilde{z}_0 = z$ with $r(\tilde{\mathbf{z}}) \leq r(\mathbf{z})$. Moreover, since $\tilde{z}_t \succeq \overline{z}$ and $\overline{z}$ is ν-robust, there is a zero resistance path from $\tilde{z}_t$ to $\overline{z}$ by Lemma 4. Hence the radius of z may be computed as the resistance of $\tilde{\mathbf{z}}$. Let a^j, a^{-j} be the profile of content actions corresponding to z.

Suppose for player j that $\overline{r}^j \leq \underline{r}_z^1 + \underline{r}_z^2$ and $\overline{r}^j \leq \overline{r}^{-j}$. It suffices to consider the case where $D^j(\tilde{z}_\tau) < \underline{r}^{-j}$ and $\underline{r}^j < D^{-j}(\tilde{z}_\tau) < \overline{r}^j$. In population $-j$, content learners are playing a ν-best response while discontent learners need not be. Consider the transition in which no learner trembles, and discontent learners play a^{-j} and are active. This transition has no resistance. If discontent learners in population j do not play a ν-best response, are active, and no learner trembles, we reach this transition with zero resistance. In the former case, since $\tilde{z}_\tau \succeq \tilde{z}_{\tau-1}, z_\tau$ for all τ, the number of discontent learners in population j can be increased only if $D^j(\tilde{z}_\tau) < \underline{r}^{-j}$ increases, and since $\underline{r}^{-j} \geq 1$, this requires at least one content learner who is playing a ν-best response to become discontent so this transition has resistance at least 1 by Lemma A.2. This characterizes the basin of z. Next, we show that as long as we leave the basin, we can reach any other ν-robust state. Assume $D^{-j}(\tilde{z}_\tau) \geq \overline{r}_z^j$. Then player j content learners are not playing a ν-best response to some feasible profile of actions $\alpha^{-j} \in \mathcal{A}^{-j}(\tilde{z}_\tau)$. Let them be active and let no learners tremble. This transition has no resistance. In the following state, suppose that all the discontent learners in j induce a feasible action so that content learners in $-j$ are not playing a ν-best response. Then discontent learners in j and $-j$ are inactive, content learners in $-j$ are active and observe that they are not playing a ν-best response, and

there are no trembles. This zero resistance transition results in a state where all learners are discontent. By Lemma 6(i), there is a zero resistance path to any ν-robust state. □

Proof of Lemma 9. By Lemma 6 it suffices to consider paths $\mathbf{z}$ from z to any proto-ν-robust state z'. Because z is ν-robust, all learners are content and play a ν-best response. Hence, any transition from z to some other proto-ν-robust state $\hat{z}$ has $r(z,\hat{z}) \geq 1$, since by Lemma A.2 at least one content learner who is playing a ν-best response must tremble for the system to leave z. We apply the following algorithm to construct least resistance paths between ν-robust states. In z, identify an action $\tilde{a}^j$ for one player j that is played by the largest number of learners in $\operatorname{supp}(\overline{\alpha}^j(z))$. Suppose that in the transition from z to z' one content player j agent in state $a^j \in A^j$ trembles and become discontent, while all the other content agents are inactive and do not tremble. This implies that $r(z,z'') = 1$, and $w(z'') \leq w(z)$ by construction. If z'' is proto-ν-robust, consider the transition from z'' to z''' where the unique discontent learner plays the action $\tilde{a}^j \neq a^j$ (notice that $\tilde{a}^j \in \operatorname{supp}(\overline{\alpha}^j(z''))$), is inactive, and does not tremble, while the rest of the learners do not tremble and are inactive. Thus z''' is ν-robust and $h(z''') > h(z)$. Otherwise, z'' is not proto-ν-robust, so there is a zero resistance path $\mathbf{z}$ from z'' to a state $\tilde{z}$ with $w(\tilde{z}) < w(z'')$ by Lemma 6. If $\tilde{z}$ is a proto-ν-robust state, we are done. If $\tilde{z}$ is not a proto-ν-robust state, we proceed as in the last step. By repeatedly applying Lemma 6, we construct a zero resistance path $\mathbf{z}'$ from $z'_0 = \tilde{z}$ to another state $z'_t = \overline{z}$ with $w(z_{\tau+1}) < w(z_\tau)$ for $t \geq \tau \geq 0$ until we reach a proto-ν-robust state $\overline{z}$ (which could be totally discontent or not). By Lemma 6, from a totally discontent state we can reach any ν-robust state. □

A.4 *Proof of Lemma 10*

Assume there is a CURB block W that does not contain all Nash equilibria and that there is a set of ν-robust states S where the support of the action profiles corresponding to content learners lies entirely on W. For S, define κ_S^j to be the least fraction of learners from population $-j$ that must play $a^{-j} \in A^{-j} \setminus W^{-j}$ so that there is a learner in population j who is not using a ν-best response. Let $\kappa_S = \min\{\kappa_S^1, \kappa_S^2\}$. Note that $\kappa_S > 0$ as at least one Nash equilibrium is not in W and let $\hat{z}$ be any ν-robust state not in S. Any z' such that for either population, $D^j(z') < \kappa_S N$ belongs to the basin of S since the system returns to S with probability 1. This is because $\mathbf{A}_T^j \subseteq W^j$ for both j, which in turn implies that discontent learners eventually choose a ν-best response, and when active become content, returning to S. Also, any transition from z' to $\hat{z}$ requires at least one content learner to tremble from at least one population j for a ν-best response a^{-j} corresponding to $\hat{z}$ to be part of $\mathbf{A}_T^{-j}$.

References

Babichenko, Yakov (2018), "Fast convergence of best-reply dynamics in aggregative games." *Mathematics of Operations Research*, 43, 333–346. [139]

Basu, Kaushik and Jörgen W. Weibull (1991), "Strategy subsets closed under rational behavior." *Economics Letters*, 36, 141–146. [153]

Benaïm, Michel and Morris W. Hirsch (1999), "Mixed equilibria and dynamical systems arising from fictitious play in perturbed games." *Games and Economic Behavior*, 29, 36–72. [138]

Binmore, Kenneth and Larry Samuelson (1997), "Muddling through: Noisy equilibrium selection." *Journal of Economic Theory*, 74, 235–265. [139, 159]

Björnerstedt, Jonas and Jörgen W. Weibull (1996), "Nash equilibrium and evolution by imitation." In *The Rational Foundation of Economic Behavior* (Kenneth J. Arrow, Enrico Colombatto, Mark Perlman, and Christian Schmidt, eds.), 155–171, London: MacMillan. [139]

Bott, Raoul and James P. Mayberry (1954), "Matrices and trees." In *Economic Activity Analysis* (Oskar Morgenstern, ed.), 391–400, John Wiley and Sons, Inc., New York, New York. [146]

Dal Bó, Pedro and Guillaume R. Fréchette (2018), "On the determinants of cooperation in infinitely repeated games: A survey." *Journal of Economic Literature*, 56, 60–114. [136]

Den Hollander, Frank (2008), *Large Deviations*, volume 14. American Mathematical Society. [161]

Ellison, Glenn (1994), "Cooperation in the Prisoner's Dilemma with anonymous random matching." *Review of Economic Studies*, 61, 567–588. [159]

Ellison, Glenn (2000), "Basins of attraction, long-run stochastic stability, and the speed of step-by-step evolution." *Review of Economic Studies*, 67, 17–45. [137, 138, 143, 146]

Ellison, Glenn, Drew Fudenberg, and Lorens A. Imhof (2009), "Random matching in adaptive dynamics." *Games and Economic Behavior*, 66, 98–114. [140, 162]

Erev, Ido and Ernan Haruvy (2016), "Learning and the economics of small decisions." In *The Handbook of Experimental Economics, Volume 2* (John H. Kagel and Alvin E. Roth, eds.), 638–702, Princeton University Press. [136]

Fishburn, Peter C. and William T. Trotter (1999), "Split semiorders." *Discrete Mathematics*, 195, 111–126. [160]

Foster, Dean P. and H. Peyton Young (1990), "Stochastic evolutionary game dynamics." *Theoretical Population Biology*, 38, 219–232. [136, 138, 145]

Foster, Dean P. and H. Peyton Young (2003), "Learning, hypothesis testing, and Nash equilibrium." *Games and Economic Behavior*, 45, 73–96. [139]

Foster, Dean P. and H. Peyton Young (2006), "Regret testing: Learning to play Nash equilibrium without knowing you have an opponent." *Theoretical Economics*, 1, 341–367. [139, 145]

Freidlin, Mark I. and Alexander D. Wentzell (1998), *Random Perturbations of Dynamical Systems*, Second edition. Springer, New York. [146]

Fudenberg, Drew and Lorens A. Imhof (2006), "Imitation processes with small mutations." *Journal of Economic Theory*, 131, 251–262. [137]

Fudenberg, Drew and David M. Kreps (1993), "Learning mixed equilibria." *Games and Economic Behavior*, 5, 320–367. [138, 139, 158]

Fudenberg, Drew and David K. Levine (1993), "Self-confirming equilibrium." *Econometrica*, 61, 523–545. [138]

Fudenberg, Drew and David K. Levine (1995), "Consistency and cautious fictitious play." *Journal of Economic Dynamics and Control*, 19, 1065–1089. [158, 163]

Fudenberg, Drew and David K. Levine (1998), *The Theory of Learning in Games*. MIT Press, Cambridge, Massachusetts. [138]

Fudenberg, Drew and David K. Levine (2014), "Recency, consistent learning, and Nash equilibrium." *Proceedings of the National Academy of Sciences*, 111, 10826–10829. [139, 145]

Fudenberg, Drew and Alexander Peysakhovich (2014), "Recency, records and recaps: Learning and non-equilibrium behavior in a simple decision problem." Proceedings of the Fifteenth ACM Conference on Economics and Computation, 971–986. [136]

Fudenberg, Drew and Satoru Takahashi (2011), "Heterogeneous beliefs and local information in stochastic fictitious play." *Games and Economic Behavior*, 71, 100–120. [138]

Hannan, James (1957), "Approximation to Bayes risk in repeated play." In *Contributions to the Theory of Games Volume III* (Melvin Dresher, Albert William Tucker, and Philip Wolfe, eds.), 97–139, Princeton University Press. [163]

Hart, S. and A. Mas-Colell (2000), "A simple adaptive procedure leading to correlated equilibrium." *Econometrica*, 68, 1127–1150. [163]

Hart, Sergiu and Andreu Mas-Colell (2006), "Stochastic uncoupled dynamics and Nash equilibrium." *Games and Economic Behavior*, 57, 286–303. [139, 145]

Heller, Yuval and Erik Mohlin (2017), "Observations on cooperation." (Forthcoming). [136, 159]

Hofbauer, J. and W. H. Sandholm (2002), "On the global convergence of stochastic fictitious play." *Econometrica*, 70, 2265–2294. [138]

Hurkens, Sjaak (1995), "Learning by forgetful players." *Games and Economic Behavior*, 11, 304–329. [139, 152, 156, 158]

Kandori, Michihiro, George J. Mailath, and Rafael Rob (1993), "Learning, mutation, and long run equilibria in games." *Econometrica*, 61, 29–56. [136, 138, 145, 152]

Levine, D. K. and S. Modica (2016), "Dynamics in stochastic evolutionary models." *Theoretical Economics*, 11, 89–131. [137, 138, 145, 146, 150, 157, 160]

Levine, David K. and Salvatore Modica (2013), "Conflict, evolution, hegemony, and the power of the state." Working paper, NBER Working Paper 19221. [137]

Myerson, R. and J. W. Weibull (2015), "Tenable strategy blocks and settled equilibria." *Econometrica*, 83, 943–976. [139]

Nöldeke, Georg and Larry Samuelson (1993), "An evolutionary analysis of backward and forward induction." *Games and Economic Behavior*, 5, 425–454. [136, 138]

Oyama, Daisuke, William H. Sandholm, and Olivier Tercieux (2015), "Sampling best response dynamics and deterministic equilibrium selection." *Theoretical Economics*, 10, 243–281. [139, 152]

Pradelski, Bary S. R. (2015), "The dynamics of social influence." Working paper, University of Oxford Discussion Paper Series Number 742. [139]

Pradelski, Bary S. R. and H. Peyton Young (2012), "Learning efficient Nash equilibria in distributed systems." *Games and Economic Behavior*, 75, 882–897. [139, 145]

Rayo, Luis and Gary S. Becker (2007), "Evolutionary efficiency and happiness." *Journal of Political Economy*, 115, 302–337. [160]

Salant, Yuval and Ariel Rubinstein (2008), "(A, f): Choice with frames." *The Review of Economic Studies*, 75, 1287–1296. [160]

Samuelson, Larry (1994), "Stochastic stability in games with alternative best replies." *Journal of Economic Theory*, 64, 35–65. [136, 137, 138, 152, 153]

Sandholm, William H. (2012), "Stochastic imitative game dynamics with committed agents." *Journal of Economic Theory*, 147, 2056–2071. [136, 159]

Tercieux, Olivier (2006), "p-Best response set." *Journal of Economic Theory*, 131, 45–70. [139]

Young, H. Peyton (1993), "The evolution of conventions." *Econometrica*, 61, 57–84. [136, 137, 138, 139, 145, 152, 153, 154, 155, 156, 157]

Young, H. Peyton (1998), *Individual Strategy and Social Structure: An Evolutionary Theory of Institutions*. Princeton University Press. [139, 152, 156, 158]

Young, H. Peyton (2009), "Learning by trial and error." *Games and Economic Behavior*, 65, 626–643. [139, 145]

Co-editor George J. Mailath handled this manuscript.

Manuscript received 1 September, 2016; final version accepted 25 April, 2018; available online 4 May, 2018.