Learning Dynamics Based on Social Comparisons*

Juan I. Block† Drew Fudenberg‡ David K. Levine§

First version: June 30, 2016
This version: August 23, 2017

Abstract

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 only observe the aggregate distribution of payoffs and only recall information from the last period, we show that aggregate play comes close to Nash equilibrium behavior for (generic) games, and that pure equilibria are generally more stable than mixed equilibria. When agents observe not only the payoff distribution of other agents but also the actions that led to those payoffs, and can remember this for some time, the length of memory plays a key role. When agents' memory is short, aggregate play may not come close to Nash equilibrium, but it does so if the game satisfies an acyclicity condition. When agents have sufficiently long memory their behavior comes close to Nash equilibrium for generic games. 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: Evolution, social learning, strict equilibria, best response dynamics, equilibrium selection.

*We thank Glenn Ellison, Salvatore Modica, and Larry Samuelson for helpful conversations, and Harry Pei for detailed comments on an earlier draft.

†Faculty of Economics, University of Cambridge. Email: jb2002@cam.ac.uk
‡Department of Economics, MIT. Email: drew.fudenberg@gmail.com
§Department of Economics, EUI and WUSTL. Email: david@dklevine.com

1 Introduction

This paper develops and analyzes two models of learning in games based on social comparisons, where agents have limited information and memory. We examine both a *low information* and a *high information* model. We discuss the low information model first. Here, agents observe the highest utility realized in their own population without observing the corresponding actions. If they are getting close to the highest payoff in their population, then they are content, and continue to play the same action. Otherwise, they become discontent and experiment at random with different actions in hopes of doing better. Memory is limited in the sense that agents do not remember all the things that happened in the past, just whether they are content, and if so, what they did last period.[1] In addition, the behavior we specify implicitly supposes that agents do not try to influence the future play of others; this "strategic myopia" makes the most sense when the population is relatively large.

When people can observe the strategies that worked well for others, it seems natural to mimic those strategies. When they do not observe each others' strategies but can observe their payoffs, people may experiment if they find they are doing less well than others. We are motivated by the fact that, for example, individuals may learn from reading newspapers or watching television which report aggregate data on the economy payoffs (stock index, average wages per industry, and income distribution). This restricted form of information seems plausible in many real world social interactions in which it is difficult for people to obtain detailed information about other people's behavior.[2] It is also of relevance to laboratory experiments on games with large extensive forms, such as indefinitely 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 addition to the random play of discontent agents, our low information model has

[1] In a recent paper, 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. For a discussion of recency effects in decision making experiments, see Erev and Haruvy (2016). Recency effects have also been found in the field, for example, in the credit card market as in Agarwal et al. (2008), in the stock market as in Malmendier and Nagel (2011), or in consumers' choices made from a list as in Feenberg et al. (2017).

[2] Of course in some environments people have access to public records that aggregate information. However, many institutions delete all records after a fixed period of time (due to storage costs or law), and record-keeping devices depreciate.

[3] See for example the repeated prisoner's dilemma experiments surveyed in Dal Bó and Fréchette (2016). In many of these, a substantial minority of participants defects most or all of the time, and receive a much lower overall payoff than subjects 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.

three other components that are also common to our high information model. First, with a probability that we send to 0, content agents tremble and become discontent, which triggers a wide search on the strategy space. Second, agents only reassess their play with a probability that is bounded away from 1. Finally, we assume there is a small number of *committed* agents who play specific actions regardless of what they observe so that no strategy ever vanishes from play. Trembles make the resulting system ergodic,[4] so that all of the states with positive probability in the limit of the ergodic distributions as the probability of trembling goes to zero – that is, the "stochastically stable" states – assign probability one to states where all but the committed agents are getting about the same payoff. Moreover, the presence of the committed agents means that every possible action has positive probability even when there are no trembles, so in generic games these limit states must be approximate Nash equilibria. We provide an equilibrium selection criterion based on the fact 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)). We find that while the radius of a pure equilibrium is generally large, growing linearly with the size of the population, every mixed equilibrium has radius one. We use this to show that in large populations mixed equilibria are significantly less stable than any of the pure equilibria, even pure strategy equilibria that are not stochastically stable, and even when the mixed equilibrium gives the players a higher payoff in line with experimental evidence (see, for example, Van Huyck et al. (1990)).[5] Moreover we show that the same conclusion holds when the noise in the system comes from noisy observation of others' payoffs as opposed to exogenous trembles.

Our high information model explores the effect of allowing the agents to use more information and memory while still basing their decisions mainly on social information. It supposes that agents observe the highest payoff realized in their own population together with the corresponding action, and moreover that they recall the actions that were best responses in the last finite T periods.[6] Since agents generally experiment less when they are more experienced we assume that discontent agents randomize over the set of remembered

[4]This rules out the long-run effect of history or initial conditions epitomized in Schelling's (1960) focal points, which allows us to make predictions based solely on the payoff matrix of the game; we view this as an approximation of social norms or conventions where payoff considerations are the most important forces.

[5]This is the first such result we know of for this sort of process. Fudenberg and Imhof (2006) characterize the relative frequencies of various homogeneous steady states in a family of imitation processes, but the processes they study can in some games spend most of their time near non-Nash states. Levine and Modica (2013) like us examine the relative amount of time spent at different steady states corresponding to Nash equilibria but examine a dynamic based on group conflict rather than driven by learning errors.

[6]In contrast, Young's (1993) adaptive learning rule has one agent revising at a time that observes a sample of size K from the last T periods and chooses among those actions that are best response to the empirical distribution of actions in the sample.

best responses and last period action instead of over all actions. If agents only recall best responses in the last period, 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 we believe that it is more likely that the system would be bogged down in a best response cycle rather than moving to a mixed equilibrium in generic games.[7] However, when there is sufficiently long memory, the high-information learning process leads to stochastic stability of approximate Nash equilibria in generic games, even games that have only mixed Nash equilibria, unlike the best response with inertia dynamic. We highlight the role of memory by showing only Nash equilibria are stochastically stable if agents have a memory at least $k \times l$, 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 (Basu and Weibull (1991)). Because every game is acyclic for k and l at least as large as the action spaces, this means that stochastic stability is guaranteed in any game when memory is sufficiently long. Finally, we show by example that in the high information case with long memory, mixed equilibria can be favored over pure ones.

The main methodological contribution of the paper is to characterize the learning dynamics combining the standard theory of perturbed Markov chains and the method of circuits (see Levine and Modica (2016)), adapting their 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-coradius theorem. Our results also contribute to the long-standing debate about pure versus mixed equilibria, providing a clear connection between what players observe and equilibrium selection. We show that, in large populations, pure equilibria are more stable in environments where agents only know that there is a better response, but mixed equilibria are sometimes more stable in environments where agents have enough information that they know the best response.

Related Literature

In addition to its focus on learning from summary statistics based on social information, this paper contributes 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 belief-based learning models such as 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,

[7] We are not aware of a general characterization for best response with inertia dynamics.

but also concludes there can be stable cycles. The same conclusion applies 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), others), although that literature, unlike the one on stochastic fictitious play, does generically provide a way of selecting between strict equilibria; for example it selects the risk-dominant equilibrium in 2×2 coordination games, as does our social comparison dynamic. In larger coordination games, the two dynamics can make different selections; we discuss this further in Section 7. Young (1993), Hurkens (1995), Young (1998) consider models with one agent in each player role, where the players observe and best respond to subset of the actions taken in the last T periods, and relate the long run outcomes to curb blocks.[8] 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. Babichenko (2013), Pradelski (2015) analyze models of social influence in which agents' payoffs depend on an aggregate statistic and agents observe and best respond either to information about the actions played.[9]

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.[10] The ideas that agents only change their actions if they are "dissatisfied" and/or that they have information about the distribution of payoffs have also been explored in the literature; these papers (for example 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 "non-conventional" agents proposed by Myerson and Weibull (2015) in that committed agents consider a (strict) subset of actions, however, we focus on committed agents with singleton action sets.

A more recent literature has considered learning procedures that involve a substantial amount of randomization when players are "dissatisfied." These papers are oriented at determining when all stochastically stable points are Nash equilibria.[11] By contrast we are

[8] These papers differ in whether sampling is with or without replacement and in how beliefs are related to the sample that is observed.

[9] Pradelski (2015) considers two models. In the adoption model agents best respond to the current state with no sampling error, and there are action trembles that are used to compute the stochastically stable sets. In the usage model agents only sample from the last T periods and best respond to the cumulative state. In Babichenko's (2013) models, agents observe actions and either play a best or a distorted best response.

[10] 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)).

[11] See for example Foster and Young (2003; 2006), Young (2009), Pradelski and Young (2012), Foster and Hart (2015). Additionally papers such as Hart and Mas-Colell (2006), Fudenberg and Levine (2014) study

focused on long-run comparative statics: we compare a range of different learning procedures to characterize which ones lead to the stochastic stability of Nash equilibria in which types of games, and we also determine the relative time spent at different steady states, for example, mixed versus pure. Building on Young (2009), Pradelski and Young (2012) consider a different learning process that can spend almost all the time at efficient action profiles that are not Nash equilibrium, and show that an efficient equilibrium is selected in generic games for which a pure Nash equilibrium exists. In contrast, our procedure selects the risk dominant equilibrium in 2×2 coordination games, whereas their procedure does not; and spends most of the time at Nash equilibria.

2 Setup

Let $G = ((u^j, A^j)_{j=1,2})$ be a finite 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[12] against each agent of the opposing population. Aggregate play in population j can be represented by the mixed strategy $\alpha^j \in \Delta(A^j)$, and $\alpha^j(a^j)$ can be interpreted as 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. 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 will want to deal with the population fractions playing 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 will also make use of the grids for subsets of the population.

We make the following assumptions about payoffs:

procedures that converge with probability one to Nash equilibrium.

[12]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.

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. Since a unique best response must be strict and there are only finitely many pure actions, we may define $g > 0$ as the smallest difference between the utility of the best response and second best response to any pure strategy.

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 strategy that weakly dominates all others. Throughout the paper, we maintain this and all other numbered assumptions from the point they are stated.

3 Low Information Social Learning

We propose a learning procedure in which agents have no direct information about the behavior of others, but observe only the frequency of utilities in their own population. In addition we assume that agents have only partial ability to keep track of that information over time due to limited memory.

The population game described above is played in every period $t = 0, 1, 2, \ldots$. 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*. An agent's *type* at the start of period t is $\theta_t^j \in \Theta^j \equiv A^j \cup \{0\} \cup \Xi^j$. 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, meaning that the agent randomizes uniformly over all actions.[13] We assume the action choice is held fixed throughout the round robin. A discontent learner randomizes uniformly even if he does not tremble, while a content learner who does not tremble plays $a_{it}^j = \theta_{it}^j$.[14]

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

[14]In place of uniform play we can allow state dependent probability distributions that may have a bias towards certain actions. As long as these probabilities are bounded away from zero independent of ϵ our results are robust.

A learner who trembled is discontent, $\theta^j_{it+1} = 0$. Each non-trembling learner has an independent probability $1 > p > 0$ of being *active* and complementary probability $1 - p$ of being *inactive*. Inactive learners do not change their type so that $\theta^j_{it+1} = \theta^j_{it}$. Given the population play α_t, let $U^j(\alpha_t^{-j})$ denote the finite vector of utilities corresponding to $u^j_i(a^j_{it}, \alpha_t^{-j})$ for each $a^j_{it} \in A^j$, and let $\phi^j(\alpha_t) \in \Delta(U^j(\alpha_t^{-j}))$ be the frequency of utilities of population j.[15] Let $\overline{u}^j(\phi^j(\alpha_t))$ denote the highest time-t utility received in population j.[16] If $u^j_i(a^j_{it}, \alpha_t^{-j}) > \overline{u}^j(\phi^j(\alpha_t)) - \nu$ the active learner becomes or remains content, so $\theta^j_{it+1} = a^j_{it}$. Otherwise he becomes or remains discontent, so $\theta^j_{it+1} = 0$.[17] Note that this social comparison allows the agent to determine whether he is playing a ν-best response, since there is always a committed agent playing a ν-best response. However, agents cannot identify which actions are ν-best responses, as they do 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.[18]

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 social comparison parameter ν, the *tolerance* for getting less than the current highest possible payoff.[19]

We assume that the social comparison parameter is less than the smallest utility difference g between the best response and second best response to any pure strategy.

Assumption 3. $\nu < g$.

This assumption implies there is a unique ν-best response to every pure strategy. Note that this will hold even in the population game with committed agents, provided there are not too of them relative to N. In conjunction with Assumption 2, Assumption 3 implies that there is not an approximately dominant strategy. Formally, there is no player j and action $\hat{a}^j \in A^j$ such that $u^j(\hat{a}^j, \alpha^{-j}) + \nu \geq \max_{a^j \in A^j} u^j(a^j, \alpha^{-j})$ for all $\alpha^{-j} \in \Delta^N(A^{-j})$.

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

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

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

[18] For simplicity we assume that the agent randomizes over all actions, including the one he used the previous period; our theorems still hold under the alternative specification where the agent randomizes over all other actions.

[19] We assume that the learning model parameters are common to all players, and that the actions of the discontent players are drawn from a uniform distribution. As long as all errors and actions have positive probability and the order of magnitude of the error rates is common to all players these assumptions do not change our conclusions.

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 upon ϵ, the tremble probability of each learner.

We start by identifying those states that correspond to approximate Nash equilibria. 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 i from each population j are content and playing a ζ-best response to $\alpha^{-j}(z)$.

Note that a ζ-robust state is automatically ζ'-robust for any $\zeta' > \zeta$. Note that by assumption content agents all use pure actions, but not all of those actions need be 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 as a *mixed* state. Notice that the fact that the learners are playing a ζ-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 ζ-Nash equilibrium, that is, the learners' action profile $\tilde{\alpha}(z)$ is a ζ-Nash equilibrium.

In what follows, we will set the robustness measure ζ to equal the social comparison parameter ν. Intuitively, a mixed state can correspond to a mixed strategy Nash equilibrium, with different learners using different actions. However, the fact that the population is finite means that some mixed equilibria can only be approximated, for example those 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 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. All omitted proofs are presented in the Appendix or the Online Appendix.

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

The next lemma says that if N/M is large then it is also the case that best responses are robust to small changes in opponents' play.

Lemma 2. *There is a η such that if $N/M > \eta$ then if a^j is a strict best response to $a^{-j} \in A^{-j}$ then a^j is a strict best response to all $\alpha^{-j} \in \Delta^N(A^{-j})$ such that $\alpha^{-j}(a^{-j}) > 1 - M/N$. 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. *$N/M \geq \eta$ where η is large enough that Lemmas 1 and 2 hold.*

This assumption and our maintained 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 B4 (Online Appendix B.2), 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 for each z the (ergodic) probability assigned to state z. To characterize the support of the ergodic distribution on states 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', it 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 non-negative number $r(z, z')$ we have $\lim_{\epsilon\to 0} P_\epsilon(z'|z)\epsilon^{-r(z,z')}$ 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 $r(\mathbf{z}) = \sum_{k=0}^{t-1} r(z_k, z_{k+1})$. 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 the Markov process aggregate play. We will show that in case of games with pure strategy equilibria, the states that have 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 agents 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 agents 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 next lemma says that in calculating least resistance paths we may assume that discontent agents remain discontent. We refer to it as the *no cost to staying discontent* principle.

Lemma 5. *For any path $\mathbf{z} = (z_0, z_1, \ldots, z_t)$ starting at any z_0 then 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 agents must switch actions to move from one to the other and then computing the resistance to making those agents 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 agents. 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 $w(z) = 2$.

We then define a *proto* ν-robust state z, which is a state in which all content agents from each population j are playing a ν-best response to $\alpha^{-j}(z)$. We divide these into three types: a *totally discontent* state is one 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 player is content, and 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. *(1) If z is totally discontent there is a zero resistance path to every ν-robust state.*

(2) 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 $w(z) \geq w(\hat{z})$.

(3) 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 transient.

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

Proof. First we establish that if $\epsilon = 0$ a ν-robust state z is absorbing. Start in a ν-robust z_t. Since $\epsilon = 0$ nobody trembles. By assumption all learners are content so they all remain at $\theta^j_{it+1} = a^j_{it}$. The committed agents never change state by assumption. This implies that $z_{t+1} = z_t$ with probability 1.

If states are proto ν-robust there is a zero resistance path to a ν-robust state by Lemma 6 part (2), otherwise, by Lemma 6 part (3), 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 part (2) and (3) 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 part (1). □

Since there are a finite number of states, every state is either recurrent or transient, 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), Young (1993), Young (2009) and others) and mixed equilibria (similar to Foster and Young (2006), Hart and Mas-Colell (2006), Pradelski and Young (2012), 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. We remind the reader of the definition of *stochastically stable* states (Foster and Young (1990)), which are the states z such that $\lim_{\epsilon\to 0} \mu^\epsilon_z > 0$. 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'.[20] 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), specialized to the case where the only recurrent classes when $\epsilon = 0$ are singletons, and there is a single circuit.

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

For completeness we sketch two proofs. First we use the method of Ellison (2000) to show that the stochastically stable states are those with the largest radius. For any target $z = z_t$ define the *modified resistance from* $z' = z_0$ to be $mr(z', z) = \min_{\mathbf{z}=(z', z_1, \ldots, z)} \sum_{k=0}^{t-1} R(z_k, z_{k+1}) - \sum_{k=1}^{t} r_{z_k}$ and the *modified co-radius* as $c_z = \max_{z'} mr(z', z)$. 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 $mr(z', z) = r_{z'}$ and since $r_{z'} < \bar{r} = r_S$ the conclusion follows.

For the sharper result that $\frac{\mu^\epsilon_z}{\mu^\epsilon_{z'}} \sim \epsilon^{r_{z'} - r_z}$, 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 (1984), that $\log(\mu^\epsilon_z / \mu^\epsilon_{z'}) / \log \epsilon$

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

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 section, we characterize the stochastic stability of pure strategy Nash equilibria. We assume that pure strategy Nash equilibria exist, and 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 that 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 such that is not using a best response. Then in finding least resistance paths out of the basin of a 0-robust state z, we will establish that the critical threshold to be considered is the smaller of r_z^1, r_z^2. We will use this to characterize the radius of a 0-robust state z, and show that the minimum resistance to any other 0-robust state z' is the same for every z'.

Theorem 2. *(1) 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.*

(2) If z and z' are 0-robust states, then $\frac{\mu_z^\epsilon}{\mu_{z'}^\epsilon} \sim \epsilon^{r_{z'} - r_z}$ and in particular those states with 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 $r(\tilde{\mathbf{z}})$ in order 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 for both players j $D^j(\tilde{z}_\tau) < r_z$ we show that $\tilde{z}_\tau$ is in the basin of z. Suppose discontents play the unique best response a^j in each population j, which gives rise to a feasible profile of actions, that they do not tremble, 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 any feasible 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 one. Moving along any such path requires that for all content agents of at least one player j it must be that $\hat{a}^j \neq a^j$, from Assumption 1. Since $D^j(\tilde{z}_\tau) < r_z$ for $j = 1, 2$ all content agents are playing a best response which implies that any transition $(\tilde{z}_\tau, z')$ on the path to $\hat{z}$ we must have that $D^j(z') > D^j(\tilde{z}_\tau)$ for at least one player j. But in this transition at least one content agent who is playing a best response becomes discontent so this transition has resistance at least one.

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 the profile α^j is such that all content agents in $-j$ are active and observe a better response played by a committed agent, so become discontent, while learners in population j are inactive and do not tremble. This transition has zero resistance. The next transition has a profile α^{-j} so that contents in j are active and get a signal about a better response provided by a committed agent, do not tremble, and become discontent while agents in $-j$ do not tremble, are inactive, and continue to be discontent. It follows that this transition has no resistance. By Lemma 6 there is a zero resistance path to any 0-robust state. Hence part (2) follows directly from Corollary 1. □

We have shown that computing the radius is determined by two thresholds, one for each player role, that represent the least number of learners that are able to move all learners to discontentment. In words, part (1) establishes that 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. Since once all learners are discontent the system may transition to any other pure equilibrium with no resistance, this implies that there is a single circuit containing all 0-robust states.

Theorem 2 also shows that for any pair of pure strategy Nash equilibria z and z' respectively, the system spends approximately $\epsilon^{r_{z'}-r_z}$ times as much time at the pure equilibrium z as at the pure equilibrium z'. It follows from the fact that if the probability of leaving z is of order ϵ^{r_z}, then the expected length of time spent at z is ϵ^{-r_z}. We provide a characterization for computing all relative ergodic probabilities of equilibria, and shows that the system spends most of the time at Nash equilibrium that have big radii as they are hard to leave. This characterization is based on the property that random search is relatively as likely to find one equilibrium as another, meaning that once an equilibrium is left there is no differentiation as to which equilibrium the system is likely to move next, what matters is the leaving time. Note that our characterization is simple in that it only requires 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 will show that the system starting at a mixed equilibrium either moves with resistance 1 towards mixed equilibria with smaller supports, or transitions along resistance 1 paths to every equilibrium. On the other hand 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 $\Delta^N(A)$.[21] 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 equilibria. 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, $u_i^j(a_i^j, \alpha^{-j}) \geq u_i^j(\tilde{a}_i^j, \alpha^{-j}) - \nu$ for each a_i^j in the support of $\tilde{\alpha}^j$ and all $\tilde{a}_i^j \in A^j$.

There is an essential difference between the structure of basins of pure ν-robust states in the case $\nu = 0$ and the case $\nu > 0$. Note first 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's best responses. First, additional actions may become ν-best responses to the play of population $-j$ in addition to $\hat{a}^j$, and

[21] We then consider content agents that play ν-best responses for $\nu > 0$.

second $\hat{a}^j$ will eventually no longer be a ν-best response to the play of the opposing population $-j$. In the case $\nu = 0$ the assumption of unique best responses on the grid (Assumption 1) assures that these two changes take place for exactly the same play of $-j$. However, with $\nu > 0$, in general additional ν-best responses arise before $\hat{a}^j$ is no longer a ν-best response. This raises the possibility that play might transition from a pure Nash equilibrium by modifying the play of both players so that both have additional ν-best responses. When $\nu = 0$ this possibility does not exist: Since the point at which one player has a different best response already gets out of the basin, it cannot be the least resistance path for both players to tremble so that both are ready to switch. In the following 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 $-j$ that must deviate for there to be a learner of player j that 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. If for both j $\overline{r}_z^j > \underline{r}_z^1 + \underline{r}_z^2$ then "sidewise" escape where both players tremble will have lower resistance than "direct" escape where only one player trembles. However, as $\nu \to 0$ both $|\overline{r}_z^j - \underline{r}_z^j| \to 0$ so we next find conditions under which there is no sidewise escape.

Lemma 7. *There is a χ and γ with $N/M > \gamma$ and $\nu < \chi$ such that for every pure ν-robust state z, there is at least one j 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 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 has lower resistance.

Assumption 5. *$\nu < \chi$ 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 that 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.*

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 $\overline{\alpha}(z)$ 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 will be multiple mixed approximate equilibria in a neighborhood of mixed equilibrium, 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 either moves to another mixed ν-robust state with weakly smaller support or to a pure ν-robust state.

Lemma 9. *The radius of a mixed ν-robust state z is $r_z = 1$, and there is either a path with resistance 1 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, $\frac{\mu^\epsilon_z}{\mu^\epsilon_{z'}} \sim \epsilon^{r_{z'} - r_z}$, so in particular the stochastically stable states are those with the largest radius.*

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 and z' is a pure ν-robust state, and N is large enough that $r_{z'} > 1$, then $\frac{\mu^\epsilon_z}{\mu^\epsilon_{z'}} \to 0$ 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 ones 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 non-stochastically 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 larger than 1.

6 High Information Social Learning

Our learning procedure thus far has focused on agents that have very limited social information and very limited memory. We now consider "high information" models where agents both observe and remember more.

6.1 The Learning Procedure

We make three changes to the learning procedure. Previously we assumed that in every period t an agent observed $\phi^j(\alpha_t)$, the frequency of utilities corresponding to actions actually played. Now we assume that an agent observes the joint frequency of utilities and actions played in the population game[22]

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

Concerning recall, we now assume that at the beginning of a period agents can recall which actions were ν-best responses during the last finite $T \geq 1$ periods, and that all agents recall the last action they played, not only the content agents.[23]

Formally 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. Subtype 0 indicates the agent is discontent, and subtype 1 indicates the agent is content. Thus, the first part of a type for learners $A^j \times \{0,1\}$ gives the previous action taken a_{it-1}^j for both content and discontent agents. The rules concerning the dynamics of contentment do not change. The final part characterizing 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

[22] 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 $u_i^j(a_{it}^j, \alpha_t^{-j}) = u^j$.

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

same for all learners of player j we refer to this as the *common memory* of player j and the actions a^j for which $T^j_{it}[a^j] < T$ as the players' *common memory set*, which we denote by $\mathbf{A}^j_T$. This simplifies the description of the state since we can use a single memory that is relevant for all agents of a player.[24] The *individual memory set* of agent i of player j is the union of the common memory set and the last action that agent played, that is, $\mathbf{A}^j_i = \mathbf{A}^j_T \cup \{a^j_{it-1}\}$.

The impact of the memory set is only on the behavior of discontent agents. We assume they play uniformly only over their individual memory set $\mathbf{A}^j_i$ rather than over all actions A^j. Even though the behavior of the agents in this model can depend on their memory, we will 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 ones 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 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 they were used based on that period's cross-section information. Our model also differs from Young (1993), Hurkens (1995), Young (1998), 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

Observe that when $T = 1$, discontent agents randomize over the last period action and the current ν-best response. This is similar to the two-population version of the best-response-plus-mutation dynamic in Kandori et al. (1993) (KMR henceforth). The specific version of their model we focus on is called *best-response with inertia:* It assumes that in each period with some 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 KMR show by example it can lead to better behaved and more sensible

[24] 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 only have access to the most recent driving records that are cleared after a certain number of years; and in informal markets information is usually transmitted through word of mouth that naturally fades away.

dynamics in the two population case with results similar to those with one population.[25] For an analysis of this dynamic see Samuelson (1994). We will show that when $T = 1$, $\nu = 0$, the high information social learning and best response with inertia dynamic have the same recurrent classes and same resistance transitions from one recurrent class to another, which in turn implies that the stochastically stable set and (for $\varepsilon > 0$) the ergodic probabilities of the recurrent classes are the same.

In order to make this comparison formally, we must extend the state space to incorporate the current population play. Let $\Phi_t^j \in \Delta^N(\Theta_T^j \cup A^{-j})$ be a vector of population shares of the player j types in period t, which includes the description of play of the opposing population α_{t-1}^{-j} in period $t-1$. Both our dynamic and the best-response with inertia dynamic are Markov processes on this extended state space.

As in Section 5.4 we restrict attention to exact best responses, that is, $\nu = 0$. We remark that Assumption 1 implies that for each population j there is a single action a^j with $T_{it}^j[a^j] = 0$ and all the other actions $\tilde{a}^j \neq a^j$ have $T_{it}^j[\tilde{a}^j] = 1$. All actions that are not best responses to the previous population play have been forgotten. Finally, for compatibility we assume that $\#\Xi^j = 0$ for each population, that is there are no committed agents, but rather that players directly observe which actions are best responses. Since we assume that N/M is large this is a reasonable approximation.

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 ever 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 that 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

[25] 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.

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.

It is convenient to define a *block* to be any set $W = W^1 \times W^2$ with non-empty subsets of actions $W^j \subseteq A^j$ for $j = 1, 2$ and the associated *block game* G^W is the original game restricting payoffs and actions to the block W. A block W is *curb* ("closed under rational behavior") 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 (see Basu and Weibull (1991)). That is, a set of action profiles is curb if it contains all best responses to itself. 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.

Definition 2. A game is $k \times l$ *acyclic* if for every action profile 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:[26]

	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

[26]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.

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.

From Theorem 4 and Lemma B5 in Online Appendix B.4 (see also Samuelson (1994)), we know that only Nash equilibria are stochastically stable for acyclic games and $T = 1$.[27] 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 where best response paths need to end up in a curb block. In particular, 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 agents' behavior approaches equilibria regardless of the payoff structure of the game. The next result shows that, unlike best response with inertia, high information social learning without trembling converges with probability one to a Nash equilibrium for generic two player games, if memory is sufficiently long.

Theorem 5. *If the game G is $k \times l$ acyclic then, with memory $T \geq k \times l$ 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 agents 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 agents that 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 to z_{t+3}, we assume agents in population $-j$ hold their play fixed and are inactive, whereas one committed agents in population j plays the ν-best response a^j to α_{t+2}^{-j}, and agents of player j are all active and those not playing a ν-best response become discontent. Consider the transition to state z_{t+4} in which agents 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 best response path we use only best responses to

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

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 agents 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 which belongs to the common memory set $\mathbf{A}_T^j$) are inactive. All agents 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 population play of learners corresponds to the $k \times l$ curb block.

Start at z_t where 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 $T = k \times l$. If in each population j all content agents are playing a ν-best response, and for each j the common memory set $\mathbf{A}_T^j$ only contains 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 agents being active and becoming or staying content, hence reaching a ν-robust state. Otherwise, there exists at least one agent in one of the populations that is not playing a ν-best response to any feasible $\alpha^{-j}(z_t)$. Consider the transition where all agents play the same previous action and in one population j those agents that 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 agents in population $-j$ that 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, all learners are active and become content; therefore reaching the corresponding ν-robust state. □

As we have seen, only pure ν-robust states have radii that increase linearly with population size N. 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 agents 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 contains 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. Similar to Young (1993), Hurkens (1995) considers a learning model where a single player in each population is randomly selected to play the game and draws a sample of K observations with replacement from the set of the last T actions of the opponent. He shows that when T is large ergodic sets correspond to minimal curb blocks, and that if the game has a unique equilibrium then for large enough histories the probability that players are playing the equilibrium tends to one. Building on this model, Young (1998) develops a learning procedure where agents draw a sample without replacement from the last T observations with the possibility of trembles. He finds that absorbing states correspond to minimal curb blocks and in the limit as ϵ vanishes stochastically stable minimal curb blocks are those with minimal stochastic potential.

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, $\#\Xi^j = 0$ for $j = 1, 2$, and agents are able to directly observe the best responses, the computation of the radius and co-radius with high information model is exactly the same as for best response with inertia.

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 (acyclic (Young (1993))), so from Theorem 4 and Lemma B5 the limit invariant distribution for the high information with $T = 1$ dynamic 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 as they are under

best response with inertia. Note that in either dynamic (B,B), (C,C) and (D,D) 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,B), so that is the radius of (A,A). On the other hand to escape from (B,B), (C,C) or (D,D) requires only about $N/11$ of one population to tremble, from (B,B) to (C,C), from (C,C) to (D,D) and from (D,D) to (B,B), so those are the radii of (B,B), (C,C) and (D,D). Hence with low information dynamic (A,A) is stochastically stable according to Theorem 2 as it has the largest radius among pure strategy equilibria. To analyze the best response with inertia dynamic, define S to be the union of the three equilibria (B,B), (C,C), (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), (D,D) one of those strategies must earn at least $(1/2)(6 + 1/3)$ while playing (A,A) yields no more than $5/2$. On the other hand, the co-radius of S is about $N/3$ since (A,A) is the only pure Nash equilibrium outside of S and it takes at least that amount to escape from (A,A). Hence by Ellison's theorem the radius of S is bigger than the co-radius so S contains all stochastically stable states.

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,C), agents' behavior adjusts immediately because they can see that choosing (C,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.[28]

As we have seen, the stochastically stable set consists of the three points (B,B), (C,C), (D,D) and any one of them could become uniquely stochastically stable with a small payoff perturbation in the high information dynamic when $\nu = 0$ and $T = 1$. In Online Appendix B.5 we use the results of Levine and Modica (2016) to show that this is still true when $T > 16$ and $\nu > 0$.

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

[28]The low information dynamic can also predict a different equilibrium than best response with inertia even when the KMR dynamic with inertia has a singleton stochastically stable set. Suppose that a player obtains $\kappa > 0$ instead of 0 when choosing (B,B) against (C,C). To escape from (B,B) now about $N/(11-\kappa)$ of one population needs to mutate so this is the radius of (B,B). Our dynamic selects (A,A) as it continues to have the largest radius among pure strategy equilibria. The set S equal to the union of (B,B), (C,C), (D,D) still contains all stochastically stable states. Let S' be the union of (A,A), (B,B). The radius of S' is about $N/(11-\kappa)$ of one population since escaping from S' requires this agents to move to (C,C) or (D,D); and the co-radius is about $N/11$. Because the radius of S' is larger than its co-radius the stochastically stable states are in S'. Combining this with the fact that they also lie in S shows that the unique stable state is (B,B) although its radius is smaller than the radius of (A,A).

	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 $((\frac{3}{5}\text{H}, \frac{2}{5}\text{T}), (\frac{2}{5}\text{H}, \frac{3}{5}\text{T}))$ and $((\frac{3}{19}\text{H}, \frac{2}{19}\text{T}, \frac{14}{19}\text{P}), (\frac{2}{19}\text{H}, \frac{3}{19}\text{T}, \frac{14}{19}\text{P}))$. 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 actions profiles for ν-robust states consists of the state in which learners play (P,P), along with the sets of mixed approximate equilibrium profiles $\mathcal{B}$ and $\mathcal{C}$.[29]

Lemma 9 shows that ν-robust states which 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 one 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 $\{\text{H}, \text{T}\} \times \{\text{H}, \text{T}\}$ by HT. Here we can easily show from the radius co-radius argument that the block HT contains the stochastically stable set.[30] Within this set play follows a deterministic best response cycle, so that each outcome of the block game G_2^{HT} will have equal weight in the limit invariant distribution. Since the limit invariant distribution is continuous in λ, this means that the agents' time average payoff for small λ is approximately 15/4, which is 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)).[31]

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

[30] To see this, the radius of the block HT is at least $2N/3$ because if 2/3 of one population is playing in 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 block HT is about $N/5$ since at least 1/5 of one population has to mutate to escape from (P,P) and is the only pure Nash equilibrium outside of block HT. Since the radius of block HT is larger than its co-radius the stochastically stable states are contained in block HT.

[31] To the best of our knowledge, there is no general characterization of stochastically stability of mixed

The high information model with large $T > 9$ exhibits very different behavior than the low memory case. First, the block HT still contains the stochastically stable set by radius co-radius argument. Since the block HT also contains only the equilibrium $((\frac{3}{5}\text{H}, \frac{2}{5}\text{T}), (\frac{2}{5}\text{H}, \frac{3}{5}\text{T}))$, the stochastically stable set is a subset of ν-robust states in a neighborhood of $((\frac{3}{5}\text{H}, \frac{2}{5}\text{T}), (\frac{2}{5}\text{H}, \frac{3}{5}\text{T}))$. Importantly, the behavior in the block HT does not exhibit deterministic best response cycles and agents do not receive less than the minmax payoff in the long-run.

8 Discussion and Extensions

8.1 Noisy Information

In the analysis so far, there is a fixed and small (relatively to N) number of committed agents, and agents play all their opponents in round-robin tournaments so there is no sampling error in agents' observations about whether they are playing a ν-best response. In practice, however, there could be noise about what learners observe either because of sampling or because utility is a random function of the actions that are played in matches. Our results would change substantially if this noise is held fixed while the probability ϵ trembles goes to 0, because then the noise would be the only driving force, but it seems natural to allow learners to average over matches within a period to push the noise down.

The population game continues to be played in every period $t = 0, 1, 2, \ldots$. As in the low-information model, let Ξ^j be a fixed set of committed agents, with at least one agent committed to each action, and we refer to the other agents as learners. Types $\theta_t^j \in \Theta^j \equiv A^j \cup \{0\} \cup \Xi^j$ determine the play of agents: committed types play the action they are committed to, content learners play the action they are content with, and discontent learners play uniformly. Each agent holds their chosen action fixed as they play the entire opposing population round robin.

We now provide a simple alternative specification to the original low information model which we will then use to study the effect of noisy observations. Previously we assumed that each learner has probability p of being active and observing the entire frequency of payoffs received by other agents in the same player role. We now assume instead that with probability p one learner of each player is chosen to be active, and that this active learner is matched with one randomly chosen comparison agent from the same population. Previously we assume that each learner had independent probability ϵ of trembling, playing

approximate equilibria under the best response with inertia dynamic. Still, we believe it is more typical for the system to be trapped in a best response cycle than to move to a mixed approximate equilibrium, as in our example.

uniformly and becoming or staying discontent prior to taking an action. We simplify by instead assuming that after being matched with the comparison agent the active learner has independent probability ϵ of becoming or staying discontent. Otherwise, with probability $1-\epsilon$ the active learner changes type (or not) based on the social comparison: if the comparison agent has higher utility the active learner is discontent, and if the comparison agent has no higher utility the active learner is content with the current action.

The revised model allows the possibility that a learner who is not playing a best response becomes content without resistance after being matched with a comparison agent who happens to be 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 *no cost to staying 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 and so the stochastically stable set does not change.[32]

We now modify the above model to allow noisy observation of the realized payoffs. First, we eliminate the exogenous probability of becoming discontent (trembling in the original model) of the active learner. Next, we replace the single round of round robin play with K rounds of round robin play against the opposing population, still holding fixed the actions of all agents. Now we introduce noise by assuming that in round τ active learner i of player j with comparison agent k observes their own utility $u^j(a^j_{it}, \alpha^{-j}_t)$ and a noisy signal $u^j(a^j_{kt}, \alpha^{-j}_t) + \eta^j_{\tau t}$ of the utility of the comparison agent. We assume that the random shocks $\eta^j_{\tau t}$ are iid with zero mean, have support on the entire real line and have a moment generating function, and that this function is twice continuously differentiable. Moment generating functions are always log concave; we strengthen this slightly by assuming that the second derivative of the log moment generating function is strictly negative. We also assume that both populations face the same distribution of payoff shocks.

We now replace the assumption that contentment is determined by own utility with that of a comparison agent with the assumption that it is determined by comparing the average own signal over the K rounds with the average comparison signal. That is, letting

[32] Although only one learner can be active at a time in the new model, the number of active learners played no role in the analysis. 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: the resistance of the ratios of ergodic probabilities described in Theorem 2 as well as the waiting times described in Levine and Modica (2016).

$\tau = 1, 2, \ldots K$ denote the rounds, the active learner is discontent if

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

and is content otherwise. This can be written as

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

that is, if the sampling error is large relative to the utility difference.

Our interest is in the case where K is large so the the probability of "accidental" discontentment (or "trembling") is small. To this end, take $K = -\log\epsilon$. The key fact is that resistances in the sampling error model differs from those in the simplified model only in that when positive instead of being a fixed exogenous constant they are now an endogenously determined constant. Since which events have zero resistance and which have positive resistance have not changed this preserves the basic qualitative features of the simplified model, and in particular the *no cost to staying discontent* principle still applies. Therefore all equilibria still lie in a single circuit and their relative ergodic resistances are still computed by differences in the radii between the equilibria. Thus the conclusion of Theorem 2 still applies.

Of course to give the theorem content we need to know what the radii are: To compute them, we must determine for each given configuration the least resistance to a learner who is playing a best response becoming discontent. Let $\mathcal{L}[x]$ denote the logarithm of the moment generating function of $\eta_{\tau t}^j$.[33] Assume that $u^j(a_{it}^j, \alpha_t^{-j}) + \nu > u^j(a_{kt}^j, \alpha_t^{-j})$ and let $r(a_i^j, a_k^j, \alpha^{-j})$ denote the resistance of an active learner playing a_i^j for whom the comparison agent is playing a_k^j both against α^{-j} to becoming discontent. The large deviations theorem from probability theory (see Theorem I.4 of Den Hollander (2008)) shows that

$$r(a_i^j, a_k^j, \alpha^{-j}) = \min_x\left[\mathcal{L}[x] - \left(\nu + u^j(a_i^j, \alpha^{-j}) - u^j(a_k^j, \alpha^{-j})\right)x\right].$$

This resistance is minimized over comparison agents' actions when $u^j(a_i^j, \alpha^{-j}) - u^j(a_k^j, \alpha^{-j}) = 0$ so that in fact the least resistance to a learner becoming discontent when playing a best response is $\min_x\left[\mathcal{L}[x] - \nu x\right]$, a positive constant.[34] In the original model this was 1 but the exact value of the constant does not matter for computing the stochastically stable set, what is important is that a constant probability of becoming discontent gives the same

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

[34] 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.

quantitative result as the original model, that is, the radius is computed by counting number of deviations to reach the edge of the basin for each population then taking the smaller of the two numbers.

Notice that we have ruled out error in observing own payoff, for example because instead of playing round robin opponents are sampled at random, or because the stage game has noisy outcomes. The key point is that while the noise in observation of the comparison payoff may be reasonably taken to be independent of the actions the same is not true of sampling error or noisy outcomes. In this case the logarithm of the moment generating function of the errors would be a function of actions $\mathcal{L}(a_{it}^j, a_{it}^k, \alpha_t^{-j})[x]$ which in turn would lead to the least resistance to a learner becoming discontent $\min_x \left[\mathcal{L}(a_{it}^j, a_{it}^k, \alpha_t^{-j})[x] - \nu x\right]$ would also depend on the actions. The result that stochastic stability is determined entirely by the radius would be unchanged, but the radii of different equilibria would change.[35]

8.2 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. 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.[36]

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 $\pi^{-j}(z)$, 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 $\pi^{-j}(z)$.[37] 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

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

The reason for this is simply that z_t is at a ν-robust state except for a fraction of the time

[35] 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).

[36] 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.

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

$(1-\mu_S^\epsilon)$, 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 his 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, and showed that behavior comes close 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.

We considered a low information social learning model in which agents observe aggregate information about how well others are doing, but not how they obtain those payoffs, so agents are not able to directly imitate successful actions. Here we assume that agents use their limited memory to keep track of their own actions that recently did well and a "search state" that indicates that there might be better actions to experiment with. In principle agents might do better by using more memory, for instance, building a picture of the payoff matrix by remembering past play. Nonetheless this is likely to be cognitively and computationally costly, and it will work well only if the environment is stationary. We demonstrated that pure strategy equilibria should be expected to be seen a larger fraction of the time than mixed strategy equilibria when people cannot easily see what actions did well. By way of examples, we compared 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 one 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.

References

Agarwal, S., Driscoll, J. C., Gabaix, X., and Laibson, D. (2008). Learning in the credit card market. Technical report, National Bureau of Economic Research.

Babichenko, Y. (2013). Best-reply dynamic in large aggregative games. Working paper.

Basu, K. and Weibull, J. W. (1991). Strategy subsets closed under rational behavior. *Economics Letters*, 36(2):141–146.

Benaïm, M. and Hirsch, M. W. (1999). Mixed equilibria and dynamical systems arising from fictitious play in perturbed games. *Games and Economic Behavior*, 29(1):36–72.

Binmore, K. and Samuelson, L. (1997). Muddling through: Noisy equilibrium selection. *Journal of Economic Theory*, 74(2):235–265.

Björnerstedt, J. and Weibull, J. (1996). Nash equilibrium and evolution by imitation. In K. Arrow, E. Colombatto, M. P. and Schmidt, C., editors, *The Rational Foundation of Economic Behavior*, pages 155–71. London: MacMillan.

Bott, R. and Mayberry, J. (1954). *Matrices and trees*. John Wiley and Sons, Inc., New York.

Dal Bó, P. and Fréchette, G. R. (2016). On the determinants of cooperation in infinitely repeated games: A survey. *Journal of Economic Literature*. Forthcoming.

Den Hollander, F. (2008). *Large deviations*, volume 14. American Mathematical Society.

Ellison, G. (2000). Basins of attraction, long-run stochastic stability, and the speed of step-by-step evolution. *Review of Economic Studies*, 67(1):17–45.

Ellison, G., Fudenberg, D., and Imhof, L. A. (2009). Random matching in adaptive dynamics. *Games and Economic Behavior*, 66(1):98–114.

Erev, I. and Haruvy, E. (2016). Learning and the economics of small decisions. In Kagel, J. H. and Roth, A. E., editors, *The Handbook of Experimental Economics, Volume 2*, pages 638–702. Princeton University Press.

Feenberg, D., Ganguli, I., Gaulé, P., and Gruber, J. (2017). It's good to be first: Order bias in reading and citing nber working papers. *The Review of Economics and Statistics*, 99(1):32–39.

Foster, D. P. and Hart, S. (2015). Smooth calibration, leaky forecasts, finite recall, and nash dynamics. Working Paper.

Foster, D. P. and Young, H. P. (1990). Stochastic evolutionary game dynamics. *Theoretical Population Biology*, 38(2):219–232.

Foster, D. P. and Young, H. P. (2003). Learning, hypothesis testing, and nash equilibrium. *Games and Economic Behavior*, 45(1):73–96.

Foster, D. P. and Young, H. P. (2006). Regret testing: learning to play nash equilibrium without knowing you have an opponent. *Theoretical Economics*, 1(3):341–367.

Freidlin, M. and Wentzell, A. (1984). *Random Perturbations of Dynamical Systems*. Springer.

Fudenberg, D. and Imhof, L. A. (2006). Imitation processes with small mutations. *Journal of Economic Theory*, 131(1):251–262.

Fudenberg, D. and Kreps, D. M. (1993). Learning mixed equilibria. *Games and Economic Behavior*, 5(3):320–367.

Fudenberg, D. and Levine, D. K. (1993). Self-confirming equilibrium. *Econometrica*, 61(3):523–545.

Fudenberg, D. and Levine, D. K. (1995). Consistency and cautious fictitious play. *Journal of Economic Dynamics and Control*, 19(5):1065–1089.

Fudenberg, D. and Levine, D. K. (1998). *The theory of learning in games*. MIT Press.

Fudenberg, D. and Levine, D. K. (2014). Recency, consistent learning, and nash equilibrium. *Proceedings of the National Academy of Sciences*, 111(3):10826–10829.

Fudenberg, D. and Peysakhovich, A. (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*, (16):971–986.

Hannan, J. (1957). Approximation to bayes risk in repeated play. In Dresher, M., Tucker, A. W., and Wolfe, P., editors, *Contributions to the Theory of Games Volume III*, pages 97–139. Princeton University Press.

Hart, S. and Mas-Colell, A. (2000). A simple adaptive procedure leading to correlated equilibrium. *Econometrica*, 68(5):1127–1150.

Hart, S. and Mas-Colell, A. (2006). Stochastic uncoupled dynamics and nash equilibrium. *Games and Economic Behavior*, 57(2):286–303.

Hofbauer, J. and Sandholm, W. H. (2002). On the global convergence of stochastic fictitious play. *Econometrica*, 70(6):2265–2294.

Hurkens, S. (1995). Learning by forgetful players. *Games and Economic Behavior*, 11(2):304–329.

Kandori, M., Mailath, G. J., and Rob, R. (1993). Learning, mutation, and long run equilibria in games. *Econometrica*, 61(1):29–56.

Levine, D. K. and Modica, S. (2013). Conflict, evolution, hegemony, and the power of the state. Working paper.

Levine, D. K. and Modica, S. (2016). Dynamics in stochastic evolutionary models. *Theoretical Economics*, 11(1):89–131.

Malmendier, U. and Nagel, S. (2011). Depression babies: Do macroeconomic experiences affect risk taking? *Quarterly Journal of Economics*, 126(1):373–416.

Myerson, R. and Weibull, J. (2015). Tenable strategy blocks and settled equilibria. *Econometrica*, 83(3):943–976.

Nöldeke, G. and Samuelson, L. (1993). An evolutionary analysis of backward and forward induction. *Games and Economic Behavior*, 5(3):425–454.

Oyama, D., Sandholm, W. H., and Tercieux, O. (2015). Sampling best response dynamics and deterministic equilibrium selection. *Theoretical Economics*, 10(1):243–281.

Pradelski, B. S. (2015). The dynamics of social influence. Working paper.

Pradelski, B. S. R. and Young, H. P. (2012). Learning efficient nash equilibria in distributed systems. *Games and Economic Behavior*, 75(2):882–897.

Samuelson, L. (1994). Stochastic stability in games with alternative best replies. *Journal of Economic Theory*, 64(1):35–65.

Schelling, T. C. (1960). *The strategy of conflict*. Harvard university press.

Van Huyck, J. B., Battalio, R. C., and Beil, R. O. (1990). Tacit coordination games, strategic uncertainty, and coordination failure. *American Economic Review*, 80(1):234–248.

Young, H. P. (1993). The evolution of conventions. *Econometrica*, 61(1):57–84.

Young, H. P. (1998). *Individual strategy and social structure: An evolutionary theory of institutions*. Princeton University Press.

Young, H. P. (2009). Learning by trial and error. *Games and Economic Behavior*, 65(2):626–643.

A Appendix

A.1 Description of the Aggregate State Process

The Markov process we are interested in describes the evolution of the states z describing the population shares of the different types. This aggregate-level process is generated by a micro-level process that describes the evolution of the agent-states describing the types of individual agents. Define the (finite) *agent state* $x = (x^1, x^2)$ to be an assignment of types to agents $x^j \in (\Theta^j)^N$. 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 $P_\epsilon(z_{t+1}|z_t) \equiv \sum_{x_{t+1} \in X(z_{t+1})} P_\epsilon(x_{t+1}|x_t)$.[38] Let $D^j(x_t)$ be the number of discontent agents of population j in x_t, and let $\mathcal{C}(x_t)$ be the set of content agents in x_t. Let $\mathcal{T}^j$ denote the trembling learners of player j and let $\mathcal{N}^j$ be

[38]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.

the non-trembling learners. Let $\mathcal{R}^j \subseteq \mathcal{N}^j$ be the active learners. Denote an assignment of actions to all agents by $\sigma^j \in (A^j)^N$.

Lemma A1. *The aggregate transition probabilities are given by*

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

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 non-trembling 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 non-tremblers 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 non-tremblers $i \in \mathcal{N}^j$ and each subset $\mathcal{R}^j \subseteq \mathcal{N}^j$ of active non-tremblers, there is probability $p^{\#\mathcal{R}^j}(1-p)^{\#\mathcal{N}^j-\#\mathcal{R}^j}$ that exactly this subset of agents is active and updates its type according to this period's highest payoff.

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 $\overline{\alpha}^j(x_t, \mathcal{T}^j) \in \Delta^{\#\Xi^j+\#(\mathcal{N}^j\setminus\mathcal{C}(x_t))}(A^j)$ be the strategy profile corresponding to the play of the committed and non-trembling content types in x_t.[39] A strategy profile $\alpha^j \in \Delta^N(A^j)$ in x_t is *feasible with respect to* $\mathcal{T}^j$ if $N\alpha^j = (\#\Xi^j + \#(\mathcal{N}^j \setminus \mathcal{C}(x_t)))\overline{\alpha}^j(x_t, \mathcal{T}^j) + (D^j(x_t) + \#(\mathcal{T}^j \cap \mathcal{C}(x_t)))\tilde{\alpha}^j$ for some strategy profile $\tilde{\alpha}^j \in \Delta^{D^j(x_t)+\#(\mathcal{T}^j\cap\mathcal{C}(x_t))}(A^j)$.[40] In particular, let $\overline{\alpha}^j(z_t) \equiv \overline{\alpha}^j(x_t, \emptyset)$ be the strategy profile corresponding to the aggregate play of contents and committed agents in state z_t which is well-defined since $\overline{\alpha}^j(x_t, \emptyset)$ 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 $\mathcal{T}, \sigma, \mathcal{R}$ 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 non-trembling content agents are playing the action with which they are content and all other learners are playing uniformly; this implies that $\Gamma^j(x_t, \mathcal{T}^j)[\sigma^j] =$

[39] Where the non-trembling content agents play the action corresponding to their type and the committed types play their committed action.

[40] That is, if it is consistent with the play of the non-trembling content and committed types.

$(1/\#A^j)^{D^j(x_t)+\#(\mathcal{T}^j\cap\mathcal{C}(x_t))}$. 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 $P_\epsilon(x_{t+1}|x_t)=\sum_{\mathcal{T},\sigma,\mathcal{R}}P(\mathcal{T},\sigma,\mathcal{R},x_{t+1}|x_t)$. □

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

Lemma A2. *If σ^j is feasible with respect to $\mathcal{T}^j$ and some content agent $i\in\mathcal{R}^j$ is playing an a^j_{it} which 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 agent 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 in Section 5.1

Since $P_\epsilon(z'|z)$ is defined as a sum, and the terms in the sum are $P(\mathcal{T},\sigma,\mathcal{R},x_{t+1}|x_t)$, it is sufficient when analyzing resistance to look for a target $x_{t+1}\in X(z_{t+1})$ and realizations $\mathcal{T},\sigma,\mathcal{R}$ 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 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.

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 $\hat{z}$.*

Proof. 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 $N\overline{\alpha}^j(\hat{z})=(N-D^j(z))\overline{\alpha}^j(z)+D^j(z)\tilde{\alpha}^j$ 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 0. □

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

Lemma A3. *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 agent trembles and every content agent, including those who tremble, plays the action with which they are content.*

Proof. First observe that we can replace the discontent agents who tremble with discontent agents 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 agent ever trembles. To show that we can have every content agent 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 $\tilde{\sigma}_\tau$ such that all content agents play the action with which they are content, $\tilde{\sigma}_t^j$ is consistent with $\overline{\alpha}^j(x_\tau, \emptyset)$, and all agents 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 $\tilde{\mathcal{T}}^j_{2\tau+1} = \emptyset$, $\tilde{\sigma}_{2\tau+1} = \sigma_\tau$ 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 $r(z_\tau, \tilde{z}_{2\tau+1}) + r(\tilde{z}_{2\tau+1}, z_{\tau+1}) \leq r(z_\tau, z_{\tau+1})$. □

Lemma 5. *For any path* $\mathbf{z} = (z_0, z_1, \ldots, z_t)$ *starting at any* z_0 *then 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* $1 \leq \tau \leq t$*.*

Proof. 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 A3 that in $\mathbf{z}$ the least resistance transitions have agent transitions in which no discontent trembles and every content plays the action with which they are content. 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 agents 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 $N - D^j(\tilde{z}_\tau) - \#\Xi^j$ agents 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 $\mathcal{T}_\tau, \sigma_\tau, \mathcal{R}_\tau$, and that $r(z_\tau, z_{\tau+1}) = \#\mathcal{T}_\tau^1 + \#\mathcal{T}_\tau^2$ 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 $N\alpha^j = (N - D^j(z))\overline{\alpha}^j(z) + D^j(z)\tilde{\alpha}^j$ 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 agents who are content in $\tilde{x}_\tau$, that is, discontent agents are inactive, but content agents are active if and only if the corresponding agent did 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 $r(\tilde{z}_\tau, \tilde{z}_{\tau+1}) = \#\tilde{\mathcal{T}}_\tau^1 + \#\tilde{\mathcal{T}}_\tau^2 \leq \#\mathcal{T}_\tau^1 + \#\mathcal{T}_\tau^2 = r(z_\tau, z_{\tau+1})$ since $\mathcal{T}_\tau$ applies to every agent to which $\tilde{\mathcal{T}}_\tau$ applied. By construction no agent 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 agent who is content in $\tilde{x}_{\tau+1}$ has the same type as the corresponding agent in $x_{\tau+1}$ so indeed $\tilde{z}_{\tau+1} \succeq z_\tau$. □

Lemma 6. *(1) If z is totally discontent there is a zero resistance path to every ν-robust state.*

(2) 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 $w(z) \geq w(\hat{z})$.

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

Proof. 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 $\alpha_t^j \in \mathcal{A}^j(\hat{z})$. 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 (1).

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 agents play the same action as in the last period, 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 a 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, 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, discontent agents 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 (2).

Finally, to show part (3) 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 agent 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 agents 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 in Section 5.5

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 a ν-robust state there is a path from z to $\overline{z}$ with resistance equal to r_z.*

Proof. 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 $\overline{r}^j \leq \underline{r}_z^1 + \underline{r}_z^2$ 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 agents are playing a ν-best response while discontent learners need not be. Consider the transition in which nobody trembles, discontents play a^{-j} and are active. This transition has no resistance. If discontent agents in population j do not play a ν-best response, are active and nobody 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 agent that is playing a ν-best response to become discontent so this transition has resistance at least one by Lemma A2. 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 agents 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 no agents tremble. This transition has no resistance. In the following state, suppose that all the discontent agents in j induce a feasible action so that content agents in $-j$ are not playing a ν-best response. Then discontent agents in j and $-j$ are inactive, content agents in $-j$ are active to the fact that they are not playing ν-best response and there are no trembles. This zero resistance transition results in a state where all agents are discontent. By Lemma 6 part (1) there is a zero resistance path to any ν-robust state. □

Lemma 9. *If a ν-robust state z has $w(z) > 2$, its radius is $r_z = 1$, and there is either a path with resistance 1 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 $h(\tilde{z}) > h(z)$.*

Proof. 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 A2 at least one content learner that 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 $\text{supp}(\overline{\alpha}^j(z))$. 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 $\tilde{a}^j \in \text{supp}(\overline{\alpha}^j(z''))$), 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, applying repeatedly Lemma 6, we construct a zero resistance path $\mathbf{z}'$ from $z'_0 = \tilde{z}$ to other 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

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 agents play entirely within the curb block is at least κN.*

Proof. Let z be any ν-robust state such that the support of $\alpha \in \mathcal{A}(z)$ is a curb block, and denote that block by W. Let $\hat{z}$ be any ν-robust state such that the support of $\hat{\alpha} \in \mathcal{A}(\hat{z})$ intersects $A \setminus W$. Define κ_z^j to be the least fraction of learners from population $-j$ that play $a^{-j} \in A^{-j} \setminus W^{-j}$ such that any ν-best response played by the agents from population j lies in $A^j \setminus W^j$. Let $\kappa_z = \min\{\kappa_z^1, \kappa_z^2\}$. Any z' such that for either population $D^j(z') < \kappa N$ belongs to the basin of z since the system returns to z with probability 1. This is because $\mathbf{A}_T^j \subseteq W^j$ for both j which in turn implies that discontent agents choose a ν-best response, and when active become content, and $\text{supp}(\alpha) = W$. If $D^j(z') \geq \kappa N$ for at least one population j, then committed agents in population $-j$ may reveal a ν-better response $\hat{a}^{-j}$ in the support

of $\hat{\alpha}^{-j}$ so that $\mathbf{A}_T^{-j}$ is not contained in W^{-j} and all agents in population $-j$ that are active become discontent. Next all discontent agents in population $-j$ play $\hat{a}^{-j} \notin W^{-j}$ with positive probability and a committed agent in population j may play a ν-better response $\hat{a}^j$ in the support of $\hat{\alpha}^j$ so that all agents with positive probability are active. Then, discontent agents in population j play $\hat{a}^j$ in $\mathbf{A}_T^j$ with positive probability, reaching the state $\hat{z}$. □

B Online Appendix

B.1 Proofs in Section 4

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

Proof. Since the game is finite it has a mixed strategy Nash equilibrium, and for any $\nu > 0$ and any such Nash equilibrium $\hat{\alpha} \in \Delta(A)$, there is an open neighborhood $\mathscr{U}$ of $\hat{\alpha}$ in which every element is a $\nu/2$ equilibrium. For N sufficiently large there is a grid point $\alpha \in \Delta^N(A)$ in $\mathscr{U}$, and consequently for large enough N/M if the learners are content with this grid point it is ν-robust. We may choose N/M large enough that the behavior of the committed agents does not move the grid point outside of $\mathscr{U}$. □

Lemma 2. *There is a η such that if $N/M > \eta$ then if a^j is a strict best response to a pure strategy $a^{-j} \in A^{-j}$ then a^j is a strict best response to all $\alpha^{-j} \in \Delta^N(A^{-j})$ such that $\alpha^{-j}(a^{-j}) > 1 - M/N$. 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.*

Proof. The hypothesis $\nu < g$ implies that ν-best responses are strict best responses,[41] and for each pure opponent's action a^{-j} for which some a^j is the (unique) strict best response, there is a $\gamma \geq 0$ such that a^j is also a best response to any mixed strategy $\alpha^{-j} \in \Delta(A^{-j})$ such that $\alpha^{-j}(a^{-j}) \geq 1 - \gamma$. Because A^{-j} is finite, there is a $\overline{\gamma}$ such that for all $\gamma \in (0, \overline{\gamma})$ the previous conclusion holds for all such best responses a^j, which proves the statement. □

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 a 0-robust state.*

Proof. If z is 0-robust all learners are content and are playing a best response to the unique $\alpha^{-j}(z) \in \mathcal{A}^j$. By Assumption 1, content learners in each population j must be playing the same best response $\hat{a}^j$ and so z is pure. This implies that at the 0-robust state $\alpha^j(\hat{a}^j) >$

[41] Note that this is true even for $\nu = 0$.

$1 - M/N$ for each j , so $\hat{a}^j$ is a strict best response to $\hat{a}^{-j}$ and and $(\hat{a}^1, \hat{a}^2)$ is a pure strategy Nash equilibrium.

Conversely, suppose that $\hat{a}$ is a pure equilibrium, and that all learners in each population j are playing $\hat{a}^j$ and are content. Since $\hat{a}$ is strict, by Lemma 2, there is a N/M sufficiently large such that for each j the action $\hat{a}^j$ is a strict best response to any $\alpha^{-j}(\hat{a}^{-j}) > 1 - N/M$, and for such N/M there is a 0-robust state for the learners to play $\hat{a}$. □

B.2 Auxiliary Result of Section 4

The following result was noted in Section 4.

Lemma B4. *When $\epsilon > 0$ the Markov process P_ϵ generated by the low-information model is irreducible and aperiodic.*

Proof. Pick any state $\hat{z}$ where $D^j(\hat{z}) = N - \#\Xi^j$ for each population j. Start with any state z_t and take any agent state $x_t \in X(z_t)$. There is probability $\epsilon^{\#\mathcal{T}^j}$ that all learners tremble, and $\#\mathcal{T}^j = N - \#\Xi^j$, so $D^j(z_{t+1}) = N - \#\Xi^j$ for $j = 1, 2$. Take $\alpha^j_{t+1} \in \mathcal{A}^j(\hat{z})$ and choose $\hat{x}_{t+1} \in X(\hat{z})$ with an action assignment $\hat{\sigma}^j$ consistent with α^j_{t+1}. Starting at $\hat{x}_{t+1}$ there is probability $(1/\#A^j)^{2N - \#\Xi^1 - \#\Xi^2}$ that all agents play $\hat{\sigma}^j$. There is probability $(1-p)^{2N-\#\Xi^1-\#\Xi^2}$ that all agents are inactive so they all stay discontent, hence entering $\hat{z}$.

Next we observe that once at $\hat{z}$ there is positive probability of staying there for any finite length of time. That is, starting at an agent state $\hat{x} \in X(\hat{z})$ consistent with $\hat{z}$ there is positive probability that no agent trembles and is active so that learners will all remain with their contentment and action. Since starting at any state there is a positive probability of reaching a single state $\hat{z}$ where the system may rest for any length of time with positive probability implies that the system is irreducible and aperiodic. □

B.3 Proof of Lemma 7

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

Proof. For each pure Nash equilibrium $\hat{a} = (\hat{a}^j, \hat{a}^{-j})$ of the game G define $\underline{\rho}^j_{\hat{a}}(\nu)$ for player j to be the maximum probability $\alpha^{-j}(\hat{a}^{-j})$ such that $\hat{a}^j$ is not the *only* ν-best response to $\hat{a}^{-j}$. Analogously, let $\rho^j_{\hat{a}}(\nu)$ for player j be the supremum probability $\alpha^{-j}(\hat{a}^{-j})$ such that $\hat{a}^j$ is not a ν-best response to $\hat{a}^{-j}$. From Assumption 1 $\underline{\rho}^j_{\hat{a}}(0) = \rho^j_{\hat{a}}(0)$ for each player j , and by Assumption 2 $\rho^j_{\hat{a}}(0) > 0$ for each j. By definition of equilibrium $\underline{\rho}^j_{\hat{a}}(0) < 1$ for both j. Then, since $\underline{\rho}^j_{\hat{a}}(0) = \rho^j_{\hat{a}}(0)$ it follows that for each player j we have $(1 - \rho^j_{\hat{a}}(0)) < (1 - \underline{\rho}^1_{\hat{a}}(0)) +$

$(1 - \underline{\rho}^2_{\hat{a}}(0))$ and $\underline{\rho}^j_{\hat{a}}(0), \rho^j_{\hat{a}}(0) < 1$. Notice that $\underline{\rho}^j_{\hat{a}}(\nu)$ is continuous at $\nu = 0$ by Assumption 3, and that $\rho^j_{\hat{a}}(\nu)$ is continuous at $\nu = 0$ by Assumption 2 and Assumption 3. Hence for sufficiently small $\nu > 0$ for each player j we still have $(1 - \rho^j_{\hat{a}}(\nu)) < (1 - \underline{\rho}^1_{\hat{a}}(\nu)) + (1 - \underline{\rho}^2_{\hat{a}}(\nu))$ and $\underline{\rho}^j_{\hat{a}}(\nu), \rho^j_{\hat{a}}(\nu) < 1$. Since there are finitely many pure equilibria, we may choose $\overline{\nu}$ so that these conditions are satisfied at all such equilibria for all $\nu \leq \overline{\nu}$.

Take any $\nu \leq \overline{\nu}$. Since $(1 - \rho^j_{\hat{a}}(\nu)) < (1 - \underline{\rho}^1_{\hat{a}}(\nu)) + (1 - \underline{\rho}^2_{\hat{a}}(\nu))$ and $\underline{\rho}^j_{\hat{a}}(\nu), \rho^j_{\hat{a}}(\nu) < 1$ it must be that for sufficiently large $N - M$ we have $(N - M)(1 - \rho^j_{\hat{a}}(\nu)) + 3 < (N - M)[(1 - \underline{\rho}^1_{\hat{a}}(\nu)) + (1 - \underline{\rho}^2_{\hat{a}}(\nu))]$. Denote by $\lceil x \rceil$ (resp. $\lfloor x \rfloor$) the smallest (resp. the largest) integer greater than or equal to x (resp. not larger than x) so that $\overline{r}^j_z = \lceil (N - M)(1 - \rho^j_{\hat{a}}(\nu)) \rceil$ and $\underline{r}^j_z = \lfloor (N - M)(1 - \underline{\rho}^j_{\hat{a}}(\nu)) \rfloor$. Since there are finitely many equilibria there is therefore a constant Γ such that for $N - M \geq \Gamma$ we have $\overline{r}^j_z \leq \underline{r}^1_z + \underline{r}^2_z$. Since $M \geq 1$ there is a γ such that for $N/M \geq \gamma$ we have $N - M \geq \Gamma$. Since $\rho^j_{\hat{a}}(\nu) > 0$, a similar argument establishes that $\underline{r}^j_z \geq 1$ for both j. □

B.4 Absorbing States with Stochastic Best Response with Inertia Dynamic

We next provide a proof that in acyclic games with a unique best response to each pure action of the opponent, the limit invariant distribution for the best response plus mutation dynamic with inertia contains only singleton pure Nash equilibria.

Lemma B5. *Every state that does not correspond to a pure strategy Nash equilibrium is transient under best response with inertia dynamic.*

Proof. Fix a time t and suppose that the state does not correspond to a pure strategy equilibrium. There is positive probability that this period all agents of one player, say j, do not adjust their play while all agents of the other player $-j$ play the best response to the date-t state, and that at date $t + 1$ all agents of j play the best response to the date $t + 1$ state while all agents of player $-j$ hold their actions fixed. Thus there is positive probability that play in each population corresponds to a pure strategy from period $t + 2$ on. Because the game is finite and acyclic, the best response path from this state converges to a pure strategy Nash equilibrium in a number of steps no greater than $J \equiv \#A^1 \times \#A^2$. There is positive probability that the populations will take turns adjusting, all of the $-j$ agents adjusting in periods $t, t+2, t+4, \ldots$, and all of the j agents adjusting at $t+1, t+3, t+5, \ldots$, so this equilibrium has probability bounded away from 0 of being reached in $2 + J$ steps, showing the initial time t state is transient. □

B.5 Analysis of Example 1 (Continued)

We show that when $T > 16$ and $\nu > 0$ in the high information dynamic the stochastically stable set consists exactly of the three equilibria (B,B), (C,C) and (D,D).

The block game G_1^{BCD} has seven Nash equilibria: the pure equilibria (B,B), (C,C) and (D,D), the binary mixed equilibria $((\frac{10}{11}\text{C}, \frac{1}{11}\text{D}),(\frac{10}{11}\text{C}, \frac{1}{11}\text{D}))$, $((\frac{10}{11}\text{B}, \frac{1}{11}\text{D}),(\frac{10}{11}\text{B}, \frac{1}{11}\text{D}))$, $((\frac{10}{11}\text{B}, \frac{1}{11}\text{C}),(\frac{10}{11}\text{B}, \frac{1}{11}\text{C}))$; and a mixed equilibrium in which players randomize uniformly across B, C and D.[42] Since all ν-robust states of this dynamic do not belong to the same circuit, we have to analyze circuits of circuits, but first we must establish what the structure of the circuits is.[43]

First, the three pure ν-robust states corresponding to the equilibria (B,B), (C,C) and (D,D) form a circuit since we can move from one of these equilibria to the next with resistance equal to the common radius of these equilibria, which is about $N/11$.

The mixed ν-robust states corresponding to a binary mixed equilibrium have a simple structure. Consider a binary mixed equilibrium. As weight shifts from one of the two actions for one of the players to the other until we reach an extremal point at which a further shift causes the other player no longer to be playing a ν-best response for both of his actions. The structure of these equilibria is that of a square: for each player there is a sequence of consecutive grid points between the two actions for which the opponent's two actions are a ν-best response. The complete collection of mixed ν-robust states corresponding to the binary mixed equilibrium is then the Cartesian product of these two sets. Each of these collections form a circuit, but these collections are also in a common circuit with the pure equilibria that we call the "pure/binary" circuit.[44]

The structure of the mixed ν-robust states corresponding to the mixed equilibrium over B, C and D is more complicated, since shifts are no longer one-dimensional for each player. However, the least resistance from a ν-robust state in the pure/binary circuit to some ν-robust state corresponding to the completely mixed equilibrium is about $N/2$.[45] Since this is greater than $N/11$ none of the ν-robust states corresponding to the completely mixed equilibrium are in the pure/binary circuit. Moreover, transitions from these mixed ν-robust

[42]Notice that when analyzing ν-robust states there is a subset of ν-robust states in a neighborhood of each mixed equilibrium.

[43]Recall that a circuit is a set of ν-robust states such that for any pair of states z, z' there exists a least resistance chain from z to z'.

[44]Because they can be reached from the corresponding pure equilibria with resistance equal to about $N/11$, while within each collection corresponding to a binary mixed equilibrium there is always a ν-robust state from which we can move to either of the two pure equilibria in the support of the mixed equilibrium with resistance one.

[45]As half of one population may play the remaining action to make it a ν-best response and be in the memory set.

states to the pure/binary circuit all have resistance 1.

Finally, (A,A) lies also in a separate circuit. This is because the least resistance from a ν-robust state in the pure/binary circuit to (A,A) is about $N/2$.[46] Being greater than $N/11$ implies that (A,A) does not belong to the pure/binary circuit. We can move from (A,A) to any ν-robust state in the pure/binary circuit with resistance $N/3$.

We next need to compute the *modified resistance* of going from one circuit to the next circuit, which is the least resistance from one circuit to the next circuit minus the least resistance path out of the circuit. We can then define *circuits of circuits*, which are collections of circuits such that for any pair of circuits in the collection we have a route from one to the other such that at each step the modified resistance of moving from one circuit to the next is the least resistance of moving from the one circuit to any other.

Although the structure of ν-robust states corresponding to the mixed equilibrium over B,C, D involves several circuits, note that transitions from the pure/binary circuit to any circuit containing such ν-robust states have a modified resistance of $N/2 - N/11$.[47] Moving on the other direction requires a modified resistance of no more than 1. Hence from Theorem 10 of Levine and Modica (2016) we know that the stochastically stable set belongs to the pure/binary circuit, and within that circuit we look for the largest radii: the three pure equilibria.

[46]Since if 1/2 of one population is playing in the block BCD one of those strategies must earn at least 19/6 while playing (A,A) yields no more than 5/2.

[47]Transitions from a pure ν-robust state to any ν-robust state corresponding to the mixed equilibrium over B,C and D have resistance of about $N/2$ while the radius of such a pure ν-robust state is about $N/11$. Moving on the other direction requires a modified resistance of no more than 1.