Review of Economic Studies (2000) **67**, 17–45 0034-6527/00/00020017$02.00

Basins of Attraction, Long-Run Stochastic Stability, and the Speed of Step-by-Step Evolution

GLENN ELLISON
Massachusetts Institute of Technology

First version received September 1995; *final version accepted February* 1999 (*Eds.*)

The paper examines the behaviour of "evolutionary" models with ε-noise like those which have been used recently to discuss the evolution of social conventions. The paper is built around two main observations: that the "long run stochastic stability" of a convention is related to the speed with which evolution toward and away from the convention occurs, and that evolution is more rapid (and hence more powerful) when it may proceed via a series of small steps between intermediate steady states. The formal analysis uses two new measures, the radius and modified coradius, to characterize the long run stochastically stable set of an evolutionary model and to bound the speed with which evolutionary change occurs. Though not universally powerful, the result can be used to make many previous analyses more transparent and extends them by providing results on waiting times. A number of applications are also discussed. The selection of the risk dominant equilibrium in 2×2 games is generalized to the selection of $\frac{1}{2}$-dominant equilibria in arbitrary games. Other applications involve two-dimensional local interaction and cycles as long run stochastically stable sets.

1. INTRODUCTION

To explore whether some outcomes might be regarded as more reasonable than others in games with multiple equilibria Foster and Young (1990), Kandori, Mailath and Rob (KMR) (1993) and Young (1993*a*) proposed examining the process by which conventions might become established using "evolutionary" models with persistent randomness. While any outcome can occur in any period in such models, one can explore equilibrium selection by asking whether some outcomes are much more likely than others. Specifically, an outcome is referred to as a "long run equilibrium" or as being "stochastically stable" if the long run probability with which it occurs does not vanish as the amount of noise goes to zero.[1] A striking observation of the aforementioned papers is that this seemingly weak criterion not only rules out unstable mixed equilibria, but also selects a unique outcome (the "risk-dominant" equilibrium) in 2×2 coordination games. Fascinated by the possibility of selecting between multiple strict equilibria, a rapidly growing literature has similarly explored how the development of conventions is influenced by the nature of the social interactions and the way in which players behave, and what similar analyses have to say about bargaining, signalling, and other classes of games with multiple equilibria.[2]

This paper provides a general analysis of the behaviour of evolutionary models with persistent randomness in an attempt to clarify and extend our understanding of this literature. Intuitively, if a social convention tends to persist for a long time after it is established

1. In this paper, I adopt the compromise term "long run stochastically stable" for such states.
2. See, for example, Binmore and Samuelson (1997), Ellison (1993), Kandori and Rob (1993, 1995), Nöldeke and Samuelson (1993, 1997), Robson and Vega-Redondo (1996), and Young (1993*b*).

and is sufficiently attractive in the sense of being likely to emerge relatively soon after play begins in any other state, then in the long run that convention will prevail most of the time. This paper builds an algorithm for identifying long-run stochastically stable sets around this observation by providing measures of the speed with which evolution will occur. The "step-by-step evolution" idea, which turns out to be crucial to the development, is that evolution in a given direction tends to be more rapid (and hence more powerful) when it may proceed via a sequence of smaller steps (rather than requiring sudden large changes). The main theorem of the paper is a sufficient condition for identifying the long run stochastically stable set of a model that (when it applies) also provides an upper bound on the speed with which evolution occurs. Surprisingly, the fairly simple intuitive calculation turns out to be sufficiently powerful as to account for what is happening in most models that have been previously analysed. A number of new applications are also discussed, including a result generalizing the selection of risk-dominant equilibria in 2×2 games to the selection of $\frac{1}{2}$-dominant equilibria in arbitrary games, an analysis of a two dimensional local interaction model and an example illustrating the potential long run stochastic stability of cycles.

Analytically, existing papers on the evolution of conventions tend to be fairly similar in that the main result is usually an identification of the long run stochastically stable set of a model using the tree construction algorithm developed by Foster and Young (1990), KMR, Young (1993*a*) and Kandori and Rob (1995). While this algorithm has been tremendously useful, its implementation can be difficult (particularly when working with complex models) and it has a couple of other drawbacks. First, it is inherently limited in scope to a characterization of the very long run limit. This can be problematic because evolution in these models is at times so slow as to be of limited practical importance.[3] Second, the algorithm is sufficiently complex as to have made the whole literature seem a bit mysterious—a commonly expressed frustration is the feeling that the typical paper writes down a model, says something about trees and after several pages of calculations gives an answer without letting one see how the answer is connected to the assumptions of the model.

To think generally about what is going on in the existing literature, this paper considers a simple (albeit abstract) reduced form framework in which the primitives of the model are a set of possible "states" of the population and a family of transition probabilities on this state space indexed by a noise level ε (as opposed to a set of players, the game they are playing, their behaviour rules, *etc.*). The goal is not to have a universally applicable characterization of long run stochastic stability, but instead to have a characterization that is intuitive and that (when it applies) provides a description of both long run and medium run behaviour.

To make the paper easier to follow the main theorem is presented in two steps. Section 3 contains a simplified version of the theorem based on a mutation counting argument. The radius of the basin of attraction of a limit set (or a union of limit sets) Ω, $R(\Omega)$, is defined as the minimum number of "mutations" (ε-probability events) necessary to escape the basin of attraction of Ω, and the coradius of the basin of attraction of Ω, $CR(\Omega)$, is defined as the maximum over all other states of the minimum number of mutations necessary to reach Ω. The radius is shown to provide a bound on the persistence of a set while the coradius provides a bound on its attractiveness. When $R(\Omega) > CR(\Omega)$ all long run equilibria belong to the set Ω and that the expected wait until Ω is reached

3. The papers of Ellison (1993), Binmore and Samuelson (1997), and Robson and Vega-Redondo (1996) including discussions of convergence rates and their importance.

is at most $O(\varepsilon^{-CR(\Omega)})$. A number of simple examples are used to illustrate the application of the theorem, including one in which the long-run stochastically stable set turns out to be a cycle rather than a steady-state. One corollary which may be of independent interest is that the selection of risk-dominant equilibria in 2×2 games generalizes to the selection of $\frac{1}{2}$-dominant equilibria whenever they exist.

Section 4 presents a strengthened version of the previous theorem which uses a more sophisticated measure of what makes evolution toward a limit set fast. The measure, referred to as the modified coradius and denoted by $CR^*(\Omega)$, formalizes the observation that large evolutionary changes will occur more rapidly if it is possible for the change to be effected via a series of more gradual steps between nearly stable states. A biological analogy may be useful in trying to think about why this should be the case. Think of how a mouse might evolve into a bat. If the process of growing a wing required ten distinct independent genetic mutations and a creature with anything less than a full wing was not viable, we would have to wait a very, very long time until one mouse happened to have all ten mutations simultaneously. If instead a creature with only one mutation was able to survive equally well (or had an advantage, say, because a flap of skin on its arms helped it keep cool), and a second mutation at any subsequent date produced another viable species, and so on, then evolution might take place in a reasonable period of time.

Reflecting the increased speed of step-by-step evolution, the modified coradius measure is computed by subtracting from the coradius a correction term which depends on the number of intermediate steady states along the evolutionary path and the sizes of their basins of attraction. The main theorem establishes that the wait to return to a limit set Ω is at most $O(\varepsilon^{-CR^*(\Omega)})$ (a tighter bound) and that $R(\Omega) > CR^*(\Omega)$ is also a sufficient condition for the long run stochastically stable set to be contained in Ω. While I expect that readers will find it intuitive that some accounting for the speed of step-by-step evolution should allow one to develop a more powerful theorem than that of Section 3, what may be surprising is that the fairly simple theorem of Section 4 appears to be sufficiently powerful so as to allow virtually all of the identifications of long-run stochastically stable sets found in the previous literature to be rederived as corollaries (albeit sometimes with a great deal of work). Evidently, behind most of the mysterious tree constructions in the literature, all that is happening is that the models contain a nearly stable state (or a set of such states) which requires a large number of mutations to escape (and hence is persistent) and to which the system returns relatively quickly after starting at any other point (either because it can be reached with few mutations or by a sequence of smaller steps through intermediate limit sets).

Several examples are used to illustrate the application of the theorem. While it is sometimes necessary to use the modified coradius (as opposed to the simple coradius) even in situations which are as uncomplicated as the KMR model with best response dynamics and a 3×3 game, the technique appears to be most useful when analyzing complex models where the unperturbed dynamics contain a large number of limit sets. The most noteworthy application is a demonstration that $\frac{1}{2}$-dominant equilibria are again selected in a two dimensional local interaction model and that evolution in the model is relatively rapid even though the model lacks the "contagion" dynamics of one dimensional local interaction models.

The final section of the paper is concerned with its relationship to the literature. In discussing the applicability of the theorem I attempt to catalogue both the extent to which the results of previous papers *could* have been derived as applications of the main theorem (in which case the primary benefit is that it provides a convergence rate) and the extent to which the main theorem also allows results on long run behaviour to be derived more

easily. Using the former criterion the theorem is very widely applicable; the latter situation is not as common. An alternate proof of the long run stochastic stability part of the main theorem based on a Freidlin–Wentzell "tree surgery" argument is also given and helps illustrate how the paper builds on the techniques of Young (1993*a*), Kandori and Rob (1995), Ellison (1993), Evans (1993), Samuelson (1994), and others.

2. MODEL AND DEFINITIONS

2.1. *The model*

The typical paper on the evolution of conventions writes down an explicit model in which a large finite population of players are randomly matched to play some game G, makes some assumptions about what the players observe and how they usually react to their observations, and then adds some source of noise like having each player tremble with some probability in each period. There is, however, a great deal of variation in the precise specification of each of these elements. To obtain a framework which can encompass all of these variants, what I do here is to abstract away from the population game story and focus instead on the dynamic model which is derived from it.

Definition 1. A *model of evolution with noise* is a triple $(Z, P, P(\varepsilon))$ consisting of:

1. A finite set Z referred to as the *state space* of the model;
2. A Markov transition matrix P on Z;
3. A family of Markov transition matrices $P(\varepsilon)$ on Z indexed by a parameter $\varepsilon \in [0, \bar{\varepsilon})$ such that;

   (i) $P(\varepsilon)$ is ergodic for each $\varepsilon > 0$;
   (ii) $P(\varepsilon)$ is continuous in ε with $P(0) = P$;
   (iii) there exists a (possibly asymmetric) cost function $c: Z \times Z \to \mathbb{R}^+ \cup \{\infty\}$ such that for all pairs of states $z, z' \in Z$, $\lim_{\varepsilon \to 0} P_{zz'}(\varepsilon)/\varepsilon^{c(z,z')}$ exists and is strictly positive if $c(z, z') < \infty$ (with $P_{zz'}(\varepsilon) = 0$ for sufficiently small ε if $c(z, z') = \infty$).

The state space Z should be thought of as corresponding to a set of possible descriptors of how the population has been playing some game. Depending on how large a state space one chooses to use, this description may contain any or all of how many players used each strategy in the previous period, how each individual played, how they played in previous periods, who was matched with whom, what each player observed, *etc*. To simplify the analysis one generally wishes to make the set Z as small as possible, with the one crucial constraint being that the dynamics are assumed to be Markov on the state space, so that the state must contain every aspect of the history which affects the transition probabilities. For example, if the players in the model are assumed to play a best response to what they have observed in the previous k periods, one would choose the state space to be the previous k observations of each of the players.

The Markov transition matrix P should be thought of as capturing the effects of whatever boundedly rational rule the players are assumed to follow "most of the time" in choosing their strategies. I will refer to (Z, P) as the *base* or *unperturbed* Markov process of the model. P can be represented as a nonnegative matrix with $P_{zz'}$ giving the probability that state z is followed by state z'. For example, a common specification is the "best response dynamic" under which all players are assumed in each period to play a best response to the distribution of strategies in the population in the previous period. This can be represented by assuming that $P_{zz'} = 1$ for some state z' which is reached when each

player switches to the strategy which is his best response to the distribution of play in state z and that $P_{zz''} = 0$ for all other states z''. A variety of rules in which players use more information or which are stochastic because players adjust their play only with some probability or rely on information whose acquisition is affected by the realization of the random matching are also easily incorporated. For example, it might be assumed that each player only considers changing his strategy with probability α and that in this event he plays a best response to a weighted average of his ten most recent observations. The one important restriction inherent in the specification of Z and P is that Z must be finite and P Markov. This rules out models with a continuum of players or a continuum of actions, and many models in which players' decisions can be affected by events in the arbitrarily distant past.

The third main element of the model is the specification of some type of small random perturbations. The most common specification one sees in the literature is the "independent random trembles" model in which each player in the population (independently of the other players) follows his behaviour rule with probability $1-\varepsilon$ and with probability ε chooses a strategy at random from a uniform distribution over the set of available actions. For example, if the behaviour rules are deterministic and if a state $z \in Z$ is an N-tuple giving the play of each of the N players in the previous period and the game has m pure strategies, then the transition probabilities implied by independent random trembles take the form

$$P_{zz'}(\varepsilon) = \left(\frac{\varepsilon}{m}\right)^{c(z,z')} \left(1 - \frac{m-1}{m}\varepsilon\right)^{N-c(z,z')},$$

with $c(z, z')$ being the number of players who take different actions in state z' and in the state which would have followed z had all players followed their behaviour rules.

The function $c(z, z')$ defined implicitly in part 3(iii) of the model definition is commonly referred to as the cost function of the system, with cost of a transition reflecting how unlikely it is when ε is small. The characterizations of long run stochastically stable sets in this paper depend on the specification of the random perturbations only through the cost function, so I will tend to think of the cost function (rather than $P(\varepsilon)$) as the third primitive of the model. Its importance derives from the behaviour of these models being driven by the relative likelihood of the various unlikely events when ε is small. One should not think of the need for a cost function to exist as restricting the types of allowable perturbations. Rather the specification of the perturbations is intended to be sufficiently flexible so as to allow any stationary specification of noise in which one might be interested to be represented.[4] For example, it can easily accommodate state-dependent mutation rates with unbounded likelihood ratios (as in Bergin and Lipman (1996)), correlated mutations which occur among groups of players, or the impossibility of some mutations (in which case the cost is set to infinity).

2.2. *Descriptions of the behaviour in the medium and long run*

For any fixed $\varepsilon > 0$, the Markov process corresponding to the model with ε-noise has a unique invariant distribution μ^ε (given by $\mu^\varepsilon = \lim_{t\to\infty} v_0 P^t(\varepsilon)$) which can be thought of as

4. The one thing which I am intentionally ruling out with the assumption is oscillating functions like $P_{zz'}(\varepsilon) = \varepsilon(1 + \sin(1/\varepsilon))$ which can prevent the limit distribution from being well defined. While $c(z, z')$ need not be an integer the specification does end up ruling out some innocuous functional forms for the transition probabilities like $\varepsilon \log(1/\varepsilon)$ or $e^{-1/\varepsilon}$. This is a tradeoff which has been made to simplify notation, and if one ever had a reason to analyse a model with such transition probabilities it would not be difficult to extend the main theorem to do so.

giving the probability of observing each of the states after evolution has been going on for a long, long time. The view of the literature seems to be that a small but significant amount of noise is present in the real world, and thus a characterization of μ^ε would be the ideal description of the long run consequences of evolution. This distribution is, however, difficult to compute, and hence it has become standard to focus instead on the *limit distribution* μ^* defined by $\mu^* = \lim_{\varepsilon \to 0} \mu^\varepsilon$.[5] The motivation for this is that μ^* is easier to compute and provides an approximation to μ^ε when ε is small. Also following the literature we will say that a state z is *long-run stochastically stable* (or refer to it as being an element of the long-run stochastically stable set) if $\mu^*(z) > 0$.

The most basic result on what long-run stochastically stable states look like (found in Young (1993*a*)) is that the long-run stochastically stable set of the model will be contained in the recurrent classes of (Z, P).[6] The recurrent classes are often referred to as *limit sets*, because they indicate the sets of population configurations which can persist in the long run absent mutations. The prototypical example is the singleton set consisting of the state $\vec{A}$ in which all players play A in a model where G has (A, A) as a symmetric strict Nash equilibrium, and where players do not switch away from a best response unless a mutation occurs. Limit sets, however, need not involve all players coordinating on an equilibrium. Instead they may be steady state population configurations in which different players use different strategies, cycles, or collections of states between which the system shifts randomly.

While it is common in the literature to examine only the long run behaviour of models, this is not because the medium run is not important. In some models play reaches the long run stochastically stable set very quickly, while in others what happens in the "long run" may be very different from what is observed in the first trillion periods.[7] That many papers discuss only the long run is attributable at least in part to this being what the standard tree construction algorithm allows one to do.

In this paper, I write $W(x, Y, \varepsilon)$ for the expected wait until a state belonging to the set Y is first reached given that play in the ε-perturbed model beings in state x, and treat the question of how quickly a system converges to its long-run stochastically stable set Ω by characterizing the behaviour of $\max_{x \in Z} W(x, \Omega, \varepsilon)$ when ε is small. If this maximum wait is small, convergence is fast and Ω can be regarded as a good prediction for what we might expect to see in the medium run. If the maximum wait is large, one should be cautious in drawing conclusions from an analysis of long run stochastic stability. All these waiting times will of course tend to infinity as ε goes to zero; whether convergence will be regarded as fast or slow depends on how quickly the waiting times increase as ε goes to zero.

3. A SIMPLE RADIUS-CORADIUS THEOREM

This section defines two new concepts, the radius and coradius of the basin of attraction of a limit set, and uses them to provide a simple characterization of the long-run and medium-run behaviour of some evolutionary models with noise. The theorem formalizes

5. The assumptions on the nature of the perturbations we have made are sufficient to ensure that the limit does in fact exist. To see this note that for any two states x and y Lemma 3.1 of Chapter 6 of Freidlin and Wentzell (1984) expresses the quantity $\mu_x^\varepsilon / \mu_y^\varepsilon$ as a ratio of polynomials in the transition probabilities. Because the transition probabilities themselves are asymptotically proportional to powers of ε, each of these ratios will have a limit of $\varepsilon \to 0$.

6. Recall that $\Omega \subset Z$ is a recurrent class of (Z, P) if $\forall w \in \Omega$, Prob $\{z_{t+1} \in \Omega | z_t = w\} = 1$, and if for all w, $w' \in \Omega$ there exists $s > 0$ such that Prob $\{z_{t+s} = w' | z_t = w\} > 0$.

7. See Ellison (1993).

the intuition that if a model has a collection of states Ω that is very persistent once reached, and that is reached relatively quickly after play begins in any other state, then in the long run we will observe states in Ω most of the time. Developing this observation into a technique for identifying long-run stochastically stable sets requires developing measures of the persistence of sets and of the tendency of sets to be reentered which are reasonably easy to compute. The measures used in this section are mutation counts which reflect the size of the basin of attraction of a limit set.

3.1. *The radius and coradius (and other definitions)*

The measure of persistence which appears in both versions of the main theorem of this paper is what I call the radius of a basin of attraction. Suppoze (Z, P) is the base Markov process of a model of evolution with noise, and let Ω be a union of one or more of its limit sets. The *basin of attraction* of Ω is the set of initial states from which the unperturbed Markov process converges to Ω with probability one, *i.e.*

$$D(\Omega) = \{z \in Z \mid \text{Prob}\{\exists T \text{ s.t. } z_t \in \Omega \ \forall t > T \mid z_0 = z\} = 1\}.$$

The radius of the basin of attraction of Ω is simply the number of "mutations" necessary to leave $D(\Omega)$ when play begins in Ω. Define a *path* from a set X to a set Y to be a finite sequence of distinct states $(z_1, z_2, \ldots, z_T)$ with $z_1 \in X$, $z_t \notin Y$ for $2 \leq t \leq T-1$, and $z_T \in Y$. Write $S(X, Y)$ for the set of all paths from X to Y. The definition of cost can be extended to paths by setting $c(z_1, z_2, \ldots, z_T) \in \sum_{t=1}^{T-1} c(z_t, z_{t+1})$. Define the *radius* of the basin of attraction of Ω, $R(\Omega)$, to be the minimum cost of any path from Ω out of $D(\Omega)$, *i.e.*

$$R(\Omega) = \min_{(z_1, \ldots, z_T) \in S(\Omega, Z - D(\Omega))} c(z_1, z_2, \ldots, z_T).$$

Note that if we define set-to-set cost functions by

$$C(X, Y) = \min_{(z_1, \ldots, z_T) \in S(X, Y)} c(z_1, z_2, \ldots, z_T),$$

then $R(\Omega) = C(\Omega, Z - D(\Omega))$.

The calculation of the radius of Ω does not require that one explore the full dynamic system; it typically suffices to examine the dynamics in a neighbourhood of Ω. In practice, the least costly path from Ω out of $D(\Omega)$ is often a direct path (w, z_2). In each of the examples presented in this paper, the cost function $c(x, y)$ takes the form $c(x, y) = \min_{\{z \mid P_{xz}(0) > 0\}} d(z, y)$, with d satisfying the triangle inequality.[8] In this case, a simple method for proving that $R(\Omega) = k$ is to exhibit states $w \in \Omega$, and $z_2 \notin D(\Omega)$ with $c(w, z_2) = k$ and to show both that $\min_{w \in \Omega} c(w, z) < k \Rightarrow z \in D(\Omega)$, and also that $\min_{w \in \Omega} c(w, z) < k \Rightarrow \min_{w \in \Omega} c(w, z') \leq \min_{w \in \Omega} c(w, z)$ for all z' with $P_{zz'}(0) > 0$.[9]

The second property of a union of limit sets Ω which will play a critical role in both theorems is the length of time necessary to reach the basin of attraction of Ω from any

8. The cost functions have this form because the dynamics consist of first applying the unperturbed dynamics and then adding "mutations" (under which a transition from a state z to a state y has a probability which is asymptotically proportional to $\varepsilon^{d(z,y)}$) to the state which results.

9. To see that this is sufficient, one shows that the least costly path is a direct path by showing inductively that there exist $w_1, w_2, \ldots, w_{T-1} \in \Omega$ such that $c(w, z_1, z_2, \ldots, z_T) \geq c(w_1, z_2, \ldots, z_T) \geq \cdots \geq c(w_{T-1}, z_T)$. The inductive step comes from noting that for $z_1' \in \arg\min_{\{z \mid P_{z_1 z}(0) > 0\}} d(z, z_2)$, $w_1 \in \arg\min_{w' \in \Omega} c(w', z_1')$, and $w_1^* \in \arg\min_{\{w' \mid P_{w_1 w'}(0) > 0\}} d(w', z_1')$,

$$c(w, z_1, z_2) = c(w, z_1) + c(z_1, z_2) = c(w, z_1) + d(z_1', z_2) \geq c(w, z_1) + d(w_1^*, z_2) - d(w_1^*, z_1')$$
$$= c(w, z_1) + d(w_1^*, z_2) - c(w_1, z_1') \geq d(w_1^*, z_2) \geq c(w_1, z_2).$$

other state. A simple way to put a (not particularly tight) bound on this waiting time is to count the "number of mutations" which are required to each Ω. Formally, the *coradius* of the basin of attraction of Ω, $CR(\Omega)$, is defined by

$$CR(\Omega) = \max_{x \notin \Omega} \min_{(z_1, \ldots, z_T) \in S(x, \Omega)} c(z_1, z_2, \ldots, z_T).$$

It simplifies the calculation of the coradius to keep in mind that the maximum in the formula above is always achieved at a state which belongs to a limit set. Also, because it is always possible to get from any element of $D(\Omega)$ to Ω at zero cost, the coradius is equivalently defined by

$$CR(\Omega) = \max_{x \notin D(\Omega)} \min_{(z_1, \ldots, z_T) \in S(x, D(\Omega))} c(z_1, z_2, \ldots, z_T).$$

3.2. *The theorem*

The first theorem of this paper is a sufficient condition for identifying the long-run stochastically stable set of a model. When the condition is satisfied we have also an upper bound on the rate at which convergence to that set occurs.

Theorem 1. *Let $(Z, P, P(\varepsilon))$ be a model of evolution with noise, and suppose that for some set Ω which is a union of limit sets $R(\Omega) > CR(\Omega)$. Then:*

(a) *the long-run stochastically stable set of the model is contained in Ω;*
(b) *for any $y \notin \Omega$, $W(y, \Omega, \varepsilon) = O(\varepsilon^{-CR(\Omega)})$ as $\varepsilon \to 0$.*

In the theorem and elsewhere in this paper I write "$f(x) = O(g(x))$ as $x \to 0$" as shorthand for "there exists C, $\bar{x} > 0$ such that $|f(x)| < C|g(x)|$ for all $x \in (0, \bar{x})$." Theorem 1 is a special case of Theorem 2, and therefore a formal proof is omitted. Informally it is easy to see why the result holds. Every time Ω is reached, the system will not leave $D(\Omega)$ until $R(\Omega)$ mutations occur in a relatively short period of time (short enough so that the unperturbed dynamics have not brought us back to Ω.) The probability that this occurs is proportional to $\varepsilon^{R(\Omega)}$ and hence the number of periods spent in $D(\Omega)$ before leaving turns out to be bounded below by $k_1 \varepsilon^{-R(\Omega)}$ for some constant k_1. On the other hand, starting at any state outside Ω we can find a T such that the probability of returning to Ω in T periods is at least of order $\varepsilon^{CR(\Omega)}$. Hence, the expected number of periods spent outside Ω is bounded above by $k_2 \varepsilon^{-CR(\Omega)}$. One can think of the evolution of the system as involving blocks of time spent in $D(\Omega)$ after a state in Ω occurs, alternating with blocks of time mostly spent outside $D(\Omega)$ until Ω recurs. The ratio of time spent in the latter blocks to time spent in the former is bounded above by $(k_2/k_1)\varepsilon^{R(\Omega) - CR(\Omega)}$ which goes to zero as ε goes to zero, from which we conclude that most of the time is spent in $D(\Omega)$. That most of the time is spent in Ω follows from the fact that long run stochastically stable set is always contained in the union of a model's limit sets.

3.3. *Examples*

I now present a couple of easy applications of the theorem intended mostly to illustrate the use of the radius–coradius theorem and to provide some geometric intuition applicable to uniform matching models where noise takes the form of independent random trembles. The game on the left below is that which Young (1993*a*) first used to show that risk

dominant equilibria need not be selected in 3×3 games in his model ((B, B) is risk dominant but playing A is the long-run stochastically stable outcome). The game on the right is intended as a cautionary note in light of the temptation to view the long-run stochastic stability as a criterion for equilibrium selection. Such an interpretation has been encouraged by Young's (1993*a*) proof that in weakly acyclic games all long-run stochastically stable states involve players playing a Nash equilibrium and by the reasonableness of the selection in a variety of settings.[10] However, the example points out that beyond the class of weakly acyclic games the selected outcome may instead be a cycle in which players play actions which do not occur in any equilibrium. In the example, (A, A) is the unique Nash equilibrium of the game, but the long-run stochastically stable set turns out to be the cycle in which the players alternate between all playing B and all playing C.

Example 1. *Consider a model in which N players are uniformly randomly matched to play the game G at* $t = 0, 1, 2, \ldots$. *Let the base Markov process be that determined by the best-reply dynamics, i.e. with player i in period t playing a best response to the distribution of strategies in the population in period* $t - 1$. *Let the perturbed dynamics be those generated by assuming that rather than always following the best reply dynamics each player independently in each period trembles with probability* ε, *in which case he picks a strategy at random from a uniform distribution. Write* $\vec{s}$ *for the state where all players play s.*

If G is the game shown on the left below, then for N sufficiently large $\mu^*(\vec{A}) = 1$. *If G is the game shown on the right below, then for N sufficiently large* $\mu^*(\vec{B}) = \mu^*(\vec{C}) = \frac{1}{2}$.

	A	*B*	*C*
A	7, 7	5, 5	5, 0
B	5, 5	8, 8	0, 0
C	0, 5	0, 0	6, 6

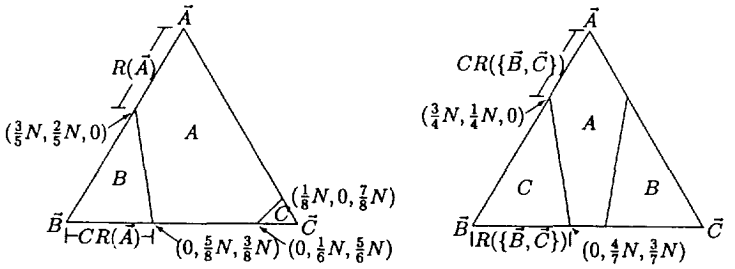
	A	*B*	*C*
A	1, 1	0, 0	0, 0
B	0, 0	-4, -4	3, 3
C	0, 0	3, 3	-4, -4

Proof. The diagrams below show the best-response regions corresponding to each of the games in $(\sigma_A, \sigma_B, \sigma_C)$-space. The deterministic dynamics in each case consist of immediate jumps from each point to the vertex of the triangle corresponding to everyone playing the best response. From the figure on the left, it is fairly obvious that in the first game $R(\vec{A})$, the minimum distance between A and any point not in $D(\vec{A})$ (here $D(\vec{A})$ is approximately the region where A is the best response), is about $\frac{2}{5}N$. The maximum distance between any state and $D(\vec{A})$ (here the distance from $\vec{B}$ to $D(\vec{A})$) is approximately $\frac{3}{8}N$. For this reason, $R(\vec{A}) > CR(\vec{A})$ for N large and the first result follows from Theorem 1. Formalizing this argument requires only a straightforward but tedious accounting for integer problems (and if desired for players not including their own play in the distribution to which they play a best response).

The result for the game on the right is also straightforward given the structure of the best response regions once one gets used to looking at the radius and coradius of limit sets which are not singletons. The deterministic cycle where the population goes from

10. Among the most appealing equilibrium selection results are Young's (1993*b*) demonstration that the Nash bargaining solution may be selected in a bargaining model and Nöldeke and Samuelson's (1997) demonstration that the Riley outcome is selected in some signalling models.

everyone playing B in one period to everyone playing C in the next and back is a limit set. The radius of the basin of this limit set, $R(\{\vec{B}, \vec{C}\})$ is approximately $\frac{3}{7}N$. The coradius $CR(\{\vec{B}, \vec{C}\})$ is approximately $\frac{1}{4}N$. For N large Theorem 1 gives $\mu^*(\{\vec{B}, \vec{C}\}) = 1$, and the result follows by symmetry. ||



The bound on the convergence rate in these games provided by the theorem is that the expected waits until reaching the long-run stochastically stable set is at most of order $\varepsilon^{-(3/8)N}$ and $\varepsilon^{-(1/4)N}$. There is nothing particularly interesting about this result, other than that as is typical in models with uniform matching the convergence time is rapidly increasing in the population size.

In the game on the right, one might be tempted to view the selection of the cycle as a curiosity attributable to an unreasonable specification of the deterministic dynamics. If I had instead chosen an unperturbed dynamic in which play changed more gradually or where players had a longer memory, for example, the cycle would not be a limit set. While the example may not be robust, the problem is fundamental. A variety of learning and evolutionary models have cycles as limit sets. The limit set which is selected by models with noise is determined by the sizes of and relations between the basins of attraction, and there is no reason to think that limit sets corresponding to equilibria will have the largest basins of attraction.

3.4. *An application: generalizing risk dominance to $\frac{1}{2}$-dominance*

In 2×2 games, the models of KMR and Young (1993*a*) provide an elegant and robust characterization of the long run impact of mutations on the development of social conventions—societies are led to adopt the risk-dominant equilibrium. In this section the radius–coradius theorem is used to derive a generalization of this result.

To derive the result it is necessary to specify a KMR-style model in a bit more detail. Suppose G is a symmetric $m \times m$ game with S the set of pure strategies. Suppose that N players are uniformly randomly paired to play G in periods $t = 1, 2, \ldots$. Let $z_t \in Z$ be the period t state which will correspond to the period t action profile $(s_{1t}, s_{2t}, \ldots, s_{Nt})$. For P some Markov transition matrix on Z reflecting a set of behavioural rules, let the perturbed process $P(\varepsilon)$ be that which results when each player independently follows his behaviour rule with probability $1 - \varepsilon$ and chooses a strategy at random with probability ε (with all strategies having positive probability.)

In this model, let σ_{it} be the mixed strategy in which the probability of action s being played is $1/(N-1)$ times the number of players $j \neq i$ with $s_{jt} = s$. Let $BR(z_t)$ denote the set of strategies which are a best response to σ_{it} for some i. KMR call the unperturbed process (Z, P) *Darwinian* if whenever $BR(z)$ is a singleton, $P_{zz'} > 0$ only if z' has more players playing strategy $BR(z)$ than does z (or z' has all players playing $BR(z)$). The definition is

motivated by the supposition that if player i decides to switch strategies after period t, a reasonable thing for him to do would be to play a best response to the strategies used by the other players in the most recent period.

The selection of risk-dominant equilibria in symmetric 2×2 games in the KMR model is quite robust in that it holds for all Darwinian dynamics, even those which are rigged so that, absent mutations, evolution toward one strategy is much faster than evolution in the reverse direction.[11] Given that Harsanyi and Selten's definition of risk dominance appeared in a book titled *A General Theory of Equilibrium Selection in Games*, the obvious first direction to explore in looking for generalizations was to see if the Harsanyi–Selten definition corresponded to what was selected in other games. As was noted in connection with Example 1, Young (1993*a*) immediately showed that it did not. Subsequent work revealed that no robust generalization is possible: KMR showed that what is selected in an asymmetric 2×2 game in a two population model depends on which Darwinian dynamic one looks at, and Ellison (1993) showed that in 3×3 games what is selected may differ between uniform and local matching models.[12]

While robust results cannot be obtained generally, the main result of this section provides a robust generalization for a restricted class of games. In a symmetric 2×2 game with pure strategies A and B, the Nash equilibrium (A, A) is said to be risk dominant if A is the best response to $\frac{1}{2}A + \frac{1}{2}B$. Harsanyi and Selten's (1988) generalization to symmetric $N \times N$ games is to call the Nash equilibrium (A, A) risk dominant if A is a better response than s to $\frac{1}{2}A + \frac{1}{2}s$ for all pure strategies s such that (s, s) is also a Nash equilibrium. Morris, Rob and Shin (1995) call a symmetric equilibrium (A, A) p-dominant if A is a strict best response against any mixed strategy placing probability at least p on A. Note that $\frac{1}{2}$-dominance is a refinement of risk dominance, and that $\frac{1}{2}$-dominant equilibria will only exist in some games. The following Corollary contains the main result of this section—that the selection of risk-dominant equilibria in 2×2 games in the KMR model generalizes to the selection of $\frac{1}{2}$-dominant equilibria whenever they exist.[13]

Corollary 1. *In the* KMR *style model described above with independent random mutations, suppose* (A, A) *is a* $\frac{1}{2}$*-dominant equilibrium of* G*. Then for* N *sufficiently large the limit distribution* μ^* *corresponding to any Darwinian dynamic has* $\mu^*(\vec{A}) = 1$.

Proof. The proof is quite easy (with its taking more than three lines here being due to its being presented in excessive detail). Since A is a strict best response to any distribution placing probability at least $\frac{1}{2}$ on A there exists a $q < \frac{1}{2}$ such that A is also a strict best response to any distribution placing probability at least q on A. The first step of the proof is to show that $R(\vec{A}) > (1-q)(N-1)$. To see this, note that whenever $z \neq \vec{A}$ has at least $q(N-1)+1$ players playing A, each player sees at least $q(N-1)$ of the $N-1$ others playing A. Thus, each of them has A as a unique best response and the Darwinian property implies that any state which can be reached at cost zero has strictly more players playing A (*i.e.* we have shown $c(\vec{A}, z') < c(\vec{A}, z)$ for all z, z' with $c(\vec{A}, z) < q(N-1)+1$ and $P_{zz'}(0) > 0$). Iterating the argument above shows that any such z is in $D(\vec{A})$. From the argument following the definition of the radius, these conditions are sufficient to show $R(\vec{A}) > N - (q(N-1)+1)$.

11. Ellison (1993) further demonstrates the robustness of this selection by showing that risk-dominant equilibria are also selected in a local interaction model involving k-nearest neighbour matching on a circle with best-reply dynamics.

12. See Hahn (1995) for a further discussion of the two population model and Ellison (1995) for a one population 3×3 example where the selection is not independent of the choice of the Darwinian dynamic.

13. We will later see that this selection is also robust to local vs. global interaction.

Next, observe that $CR(\vec{A}) \leq q(N-1)+1$, because any state in which at least $q(N-1)+1$ players play A has $\sigma_i(A) > q$ for all i, and hence as above is in $D(\vec{A})$. The direct path from any state x to a state in which the first $q(N-1)+1$ players play A while the rest play as in some state which has positive probability in the unperturbed dynamics is a path from x to $D(\vec{A})$ of cost at most $q(N-1)+1$.

Taking N sufficiently large so that $(1-2q)(N-1) \geq 1$, $R(\vec{A}) > CR(\vec{A})$ and Theorem 1 applies. ||

I hope that the fact that the proof of the Corollary is fairly trivial will be taken as evidence that the main theorem of the paper can facilitate the development of general results. Because I realize that one might be tempted to conclude instead that the result itself is trivial, I would like to note that both Kandori and Rob's (1995) result on pure coordination games and their (1993) result on risk-dominance in games with the total bandwagon property are immediate corollaries of Corollary 1. In a game where there is a different payoff to coordinating on each of the available actions and the payoffs are zero whenever the players do not coordinate, the Pareto optimal equilibrium is also $\frac{1}{2}$-dominant. In a game with the total bandwagon property, risk dominant equilibria are automatically $\frac{1}{2}$-dominant.[14] The result of Kim (1996) on the behaviour of the KMR model in a class of symmetric I player, two action games also follows from the argument above given the appropriate definition of $\frac{1}{2}$-dominance.[15]

4. STEP-BY-STEP EVOLUTION AND AN IMPROVED THEOREM

In this section a more powerful theorem is obtained by deriving a tighter bound on the time required for evolution toward a limit set to occur. The construction of the new bound, using a measure I call the modified coradius, follows a consideration of the extent to which large evolutionary changes will occur more quickly when they can be achieved by passing through a number of intermediate steady states. The improved theorem is relatively most useful when one wishes to analyse complex models with a large number of steady states.

4.1. *The modified coradius*

Suppose Ω is a union of limit sets and $(z_1, z_2, \ldots, z_T)$ is a path from x to Ω. Let $L_1, L_2, \ldots, L_r \subset \Omega$ be the sequence of limit sets through which the path passes consecutively (with the convention that a limit set can appear on the list multiple times but not successively.) Note that x may be an element of L_1 and that $L_i \notin \Omega$ for $i < r$. Define a *modified cost function*, c^*, by subtracting from the cost of the path the radius of the *intermediate* limit sets through which the path passes,

$$c^*(z_1, z_2, \ldots, z_T) = c(z_1, z_2, \ldots, z_T) - \sum_{i=2}^{r-1} R(L_i).$$

14. Note that what Kandori and Rob (1993) show is that risk dominant equilibria are selected in games satisfying both the total bandwagon property and the monotone share property. Evidently, the latter restriction is unnecessary.

15. The appropriate definition for his model being that an equilibrium $(A, A, \ldots, A)$ is $\frac{1}{2}$-dominant if each player strictly prefers to play A given any distribution of strategies in the population involving at least half of the other players playing A.

(Note that $R(L_1)$ is not included in the sum.) The definition of modified cost can be extended to a point-to-set concept by setting

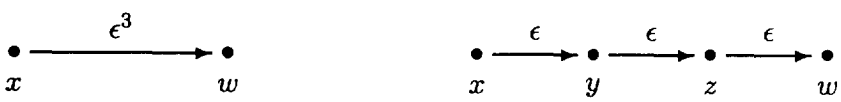
$$c^*(x, \Omega) = \min_{(z_1, \ldots, z_T) \in S(x, \Omega)} c^*(z_1, z_2, \ldots, z_T).$$

The *modified coradius* of the basin of attraction of Ω is then defined by

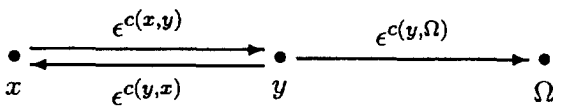
$$CR^*(\Omega) = \max_{x \notin \Omega} c^*(x, \Omega).$$

Note that $CR^*(\Omega) \leq CR(\Omega)$.

The definition is meant to capture the effect of intermediate steady states on expected waiting times. The diagram below illustrates how intermediate steady states can speed evolution. In the diagram arrows are used to indicate the possible transitions out of each state and their probabilities (with null transitions accounting for the remaining probabilities). In the Markov process on the left, the cost of the path (x, w) is three, and the expected wait until such a transition occurs is $1/\varepsilon^3$. While the cost of the path (x, y, z, w) in the Markov process on the right is also three, the expected wait until a transition from x to w occurs is only $1/\varepsilon + 1/\varepsilon + 1/\varepsilon = 3/\varepsilon$. In a biological example the difference is intuitive. The graphs can be thought of as representing two different environments in which three major genetic mutations are necessary to produce the more fit animal w from animal x. The graph on the left reflects a situation in which an animal with any one or two of these mutations could not survive. In this case, getting the large mutation we need seems very unlikely. In the situation on the right, each single mutation on its own provides an increase in fitness which allows that mutation to take over the population. The large cumulative change from x to w seems more plausible when such gradual change is possible.



While it is intuitive that intermediate limit sets can speed evolution, why the particular correction for this fact embodied in the modified coradius definition is appropriate is not obvious. (Note that the modified coradius of w is indeed three in the process on the left and one in the process on the right.) The following example is meant to provide some intuition for the correction.



In the three state process shown above, Ω is always the unique long run stochastically stable state because $R(\Omega) = \infty$. Suppose now that $c(x, y) > c(y, x)$ and $c(y, \Omega) > c(y, x)$. Then $CR^*(\Omega) = c(x, y) + c(y, \Omega) - c(y, x)$. To compute $W(x, \Omega, \varepsilon)$ (which will turn out to be the longest expected wait) note that

$$W(x, \Omega, \varepsilon) = E(N_x) + E(N_y),$$

with N_x and N_y being the number of times x and y occur before Ω is reached conditional on the system starting in state x. We can write $E(N_x)$ as the product of the expected length of the run of x's before an $x \to y$ transition occurs and the number of times the transition

$x \to y$ occurs. Hence,

$$E(N_x) \sim \varepsilon^{-c(x,y)} \frac{\varepsilon^{c(y,x)}}{\varepsilon^{c(y,\Omega)}} = \varepsilon^{-(c(x,y)+c(y,\Omega)-c(y,x))}.$$

Similarly, we can write $E(N_y)$ as the product of the length of each run of y's and the number of $x \to y$ transitions. This gives $E(N_y) \sim \varepsilon^{-(c(y,x)+c(y,\Omega)-c(y,x))}$, which is asymptotically negligible compared to $E(N_x)$. Hence we have

$$W(x, \Omega, \varepsilon) \sim \varepsilon^{-(c(x,y)+c(y,\Omega)-c(y,x))},$$

which is exactly $\varepsilon^{-CR^*(\Omega)}$.

Looking at this calculation, the subtraction of the radius of the intermediate limit set reflects the fact that because most of the time the system is in state x the ε order of the waiting time is determined by the wait to leave x and the probability of a transition from y to Ω *conditional on a transition out of y occurring*. While the unconditional probability of the transition from y to Ω is $\varepsilon^{c(y,\Omega)}$, the conditional probability is approximately $\varepsilon^{c(y,\Omega)-c(y,x)}$ (or $\varepsilon^{c(y,\Omega)-R(y)}$).

4.2. *The main theorem*

The following theorem strengthens the result of the previous section by using the modified coradius as the measure of persistence. In practice, this extension seems to greatly enhance the result's applicability.

Theorem 2. *Let $(Z, P, P(\varepsilon))$ be a model of evolution with noise, and suppose that for some set Ω which is a union of limit sets $R(\Omega) > CR^*(\Omega)$. Then*

(a) *the long-run stochastically stable set of the model is contained in Ω;*
(b) *for any $y \notin \Omega$, $W(y, \Omega, \varepsilon) = O(\varepsilon^{-CR^*(\Omega)})$ as $\varepsilon \to 0$.*

Proof. Given a Markov process $(Z, P(\varepsilon))$, write $N(A, B, x)$ for the expected number of times states in A occur (counting the initial period if $x \in A$) before the process reaches B (not counting the process as having immediately reached B if $x \in B$) when the process starts at x. Write $Q(A, B, x)$ for the probability that A is reached before B when the process starts at x (not counting what happens in the initial period if $x \in A$ or $x \in B$). A characterization of the steady-state distribution (presented as Lemma 1 in the Appendix) is that

$$\frac{\mu^\varepsilon(y)}{\mu^\varepsilon(\Omega)} = \frac{N(y, \Omega, y)}{\sum_{\omega \in \Omega} Q(\omega, \Omega - \omega, y) N(\Omega, y, \omega)}.$$

The numerator is bounded above by $W(y, \Omega, \varepsilon)$ so it will suffice for parts (a) and (b) of the theorem to show that

$$W(y, \Omega, \varepsilon) = O(\varepsilon^{-CR^*(\Omega)}) \qquad \forall y \notin \Omega, \tag{1}$$

and

$$1/N(\Omega, y, \omega) = O(\varepsilon^{R(\Omega)}) \qquad \forall \omega \in \Omega. \tag{2}$$

The second and sixth lemmas in the Appendix contain these two results.

It is also worth noting that it follows from the second and third lemmas that the radius provides a precise measure of the epsilon-order of the time necessary to leave the

basin of attraction of a single limit set, *i.e.* Lemmas 2 and 3 imply that $W(l, Z - D(L), \varepsilon) \sim \varepsilon^{-R(L)}$ for any state l belonging to a limit set L. ||

In contrast with the now standard techniques developed by Foster and Young (1990), KMR, Young (1993*a*), and Kandori and Rob (1995) the main theorem of this paper is not universally applicable. Why then might one wish to use the theorem of this paper?[16] First, and most concretely, the theorem provides a bound on the convergence rate as well as a long-run limit. Second, the theorem provides intuition for why the long-run stochastically stable set of a model is what it is. (In this respect readers may also be surprised to find that in most previous papers all that is going on behind the tree construction is that there is a single limit set, or collection thereof, which is persistent and relatively attractive.) Finally, because a direct application of the standard tree construction algorithm requires that one identify all of the limit sets of a model and compute the cost of moving between them, it is difficult both to develop theorems which apply to broad classes of models (which contain members where the number of limit sets and the relations between them differ) and simply to analyse complex models with large numbers of steady states. It is in these situations that the radius-modified coradius theorem may be easier to apply.[17]

4.3. *Examples*

I now present a few simple examples which illustrate the use of the radius-modified coradius theorem and some of its limitations. I begin with an example which again provides geometric intuition for uniform matching models.

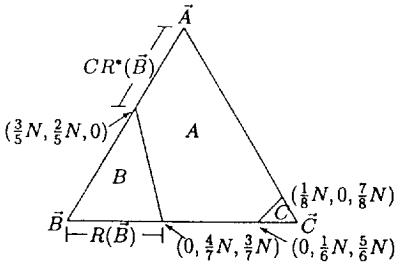
Example 2. *Consider a model in which N players are uniformly randomly matched to play the game G shown below at* $t = 0, 1, 2, \ldots$. *Let the base Markov process be that determined by the best-reply dynamics, i.e. with player i in period t playing a best response to the distribution of strategies in the population in period* $t - 1$. *Let the perturbed dynamics be those generated by assuming that rather than always following the best reply dynamics each player independently in each period trembles with probability* ε, *in which case he picks a strategy at random from a uniform distribution. Then, for N sufficiently large we have* $\mu^*(\vec{B}) = 1$.

	A	*B*	*C*
A	*7, 7*	*5, 5*	*5, 0*
B	*5, 5*	*8, 8*	*1, 0*
C	*0, 5*	*0, 1*	*6, 6*

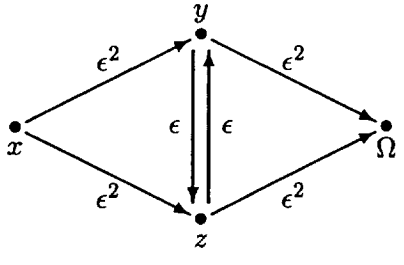
16. See Section 6 for a discussion of the theorem's applicability.

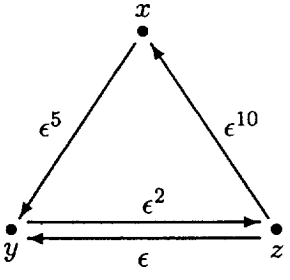
17. In Section 6.3 it is shown that whenever Theorem 2 applies to a single limit set Ω the long run stochastically stable set could have been found by applying "tree surgery" arguments in a fairly systematic way. The claim that application may be "easier" might thus be interpreted alternately as saying that the theorem systematizes the use of tree surgery arguments.

Proof. Again the result follows easily from looking at the structure of the best response regions in $(\sigma_A, \sigma_B, \sigma_C)$-space (which roughly correspond with basins of attraction under the best-response dynamics.)[18] Note that a simple radius–coradius argument does not suffice—$R(\vec{B})$ is about $\frac{3}{7}N$, while $CR(\vec{B})$ (the distance from $\vec{C}$ to $D(\vec{B})$) is about $\frac{4}{7}N$. However, $CR^*(\vec{B})$ is only about $\frac{2}{5}N$ because this is the modified cost of the path $(\vec{A}, \vec{B})$ and the modified cost of the indirect path $(\vec{C}, \vec{A}, \vec{B})$ is about $\frac{1}{8}N$. The result thus follows from Theorem 2.



The next two examples point out a couple of weaknesses of the theorem in order to provide a more complete understanding of how it works. First, the four-state Markov process whose dynamics are represented in the diagram on the left below illustrates that the upper bound which the theorem provides on the expected wait until the system reaches the long-run stochastically stable set is not necessarily tight. In the example, Ω is the long-run stochastically stable set and $CR^*(\Omega) = 3$. (Any path from x to Ω has a modified cost of at least three.) The expected wait until Ω is reached conditional on starting at x, however, is only $3/2\varepsilon^2$, because it takes $1/2\varepsilon^2$ periods to reach the set $\{y, z\}$ and then a further $1/\varepsilon^2$ periods to reach Ω. Intuitively, the reason why the modified coradius bound is not tight here is that the general calculations which derive this bound do not take full advantage of the relationships between the limit sets. In particular, the calculation is done always assuming that the "worst case" occurs whenever a transition to a new limit set does not correspond to the minimum modified cost path. In this case convergence is faster than such a worst case calculation indicates, because the transitions from y or z which do not reach Ω never take us back to x. A failure to provide tight bounds in some situations seems unavoidable in any computation which does not require a full identification of all of the limit sets and an analysis of the relationships between them.





The three-state Markov process whose dynamics are represented in the figure on the right above is intended to illustrate how (unlike the Freidlin–Wentzell characterization),

18. See Ellison (1995) for a discussion of how changing to a different Darwinian dynamic may change the shapes of the basins and the long-run stochastically stable set of this model.

a direct application of the main theorem sometimes fails to identify the long-run stochastically stable set. While the example has an unique long-run stochastically stable state, if we compute the radius and modified coradius of each of the limit sets we do not find it—$R(x) = 5 < 11 = CR^*(x)$, $R(y) = 2 < 5 = CR^*(y)$, and $R(z) = 1 < 5 = CR^*(z)$. That the theorem fails to identify the long-run stochastically stable set here is not a result of the modified coradius bound not being tight (the bounds are tight). Rather, the persistence-attractiveness description just does not capture fully what is going on as the system moves between limit sets.

This last example also provides a nice opportunity to point out a feature of the theorem which has not been emphasized so far—that the Ω in the theorem can be a union of limit sets rather than a single limit set. This feature is certainly practically important because in models involving extensive form games, *e.g.* Nöldeke and Samuelson (1993, 1997) and Huck and Oechssler (1995), the long-run stochastically stable set has often been found to be a union of several limit sets between which the system moves when single mutations occur. In addition this feature may at times be exploited to expand the theorem's power. In the example note that $R(\{x, y\}) = 2 > 1 = CR(\{x, y\})$, and that $R(\{y, z\}) = 10 > 5 = CR(\{y, z\})$. Applying the theorem twice then with $\Omega = \{x, y\}$ and with $\Omega = \{y, z\}$ we can conclude both that the long-run stochastically stable set is contained in $\{x, y\}$ and that the long-run stochastically stable set is contained in $\{y, z\}$. Clearly, this implies that y is the unique long-run stochastically stable state. It is not clear to me how useful the possibility of making arguments like this will be in practice.

5. LOCAL INTERACTION: TWO DIMENSIONAL LATTICES AND FAST VS. SLOW EVOLUTION

Evolutionary models with local interaction are intended to capture social situations in which players interact most often with a small stable set of friends, colleagues, or neighbours. Ellison (1993) argued that such models are interesting not only because such relationships exist in the real world, but also because it is only in the context of such models that convergence is fast enough to make evolutionary selection plausible. In this section I show that $\frac{1}{2}$-dominant equilibria are selected and evolution is fast in particular models with one and two dimensional local interaction structures. The latter model should be of interest both because it clarifies the role of "contagion" dynamics in producing fast evolution and as an illustration of where and how the radius-*modified* coradius theorem is most useful.

5.1. *A one-dimensional model*

Suppose first as in Ellison (1993) that players $1, 2, \ldots, N$ are arranged sequentially around a circle and that they are repeatedly matched to play the game G. Consider a $2k$ neighbour version of the best-reply dynamic in which player i in period $t+1$ plays a best response to the distribution σ_{it} formed by taking the average of the play in period t of players $i-k$, $i-k+1, \ldots, i-1, i+1, \ldots, i+k$. Again take the ε-perturbed process to be that given by independent random trembles. The following Corollary shows that in this environment the selection of the risk dominant equilibrium in 2×2 games again generalizes to the selection of $\frac{1}{2}$-dominant equilibria whenever they exist.

Corollary 2. *Suppose* (A, A) *is a* $\frac{1}{2}$*-dominant equilibrium of* G*. For* N *sufficiently large the limit distribution* μ^* *corresponding to the* $2k$ *neighbours on a circle model with best-reply dynamics has* $\mu^*(\vec{A}) = 1$*, and for all* $z \neq \vec{A}$ *we have* $W(z, \vec{A}, \varepsilon) = O(\varepsilon^{-(k+1)})$.

Proof. The argument that $R(\vec{A}) > CR(\vec{A})$ follows directly from the proof of Theorem 1 of Ellison (1993). Any state in which $k+1$ adjacent players play A lies in $D(\vec{A})$ because the cluster of players playing A will grow contagiously. Hence $CR(\vec{A}) \leq k+1$.

For $N \geq (k+1)(K+2)$ a path from $\vec{A}$ out of $D(\vec{A})$ requires a mutation among players $1, 2, \ldots, k+1$, a second mutation among players $k+2, \ldots, 2(k+1)$, *etc.*, and a $k+2^{nd}$ among players $(k+1)^2+1, \ldots, (k+1)(k+2)$. Hence, $R(\vec{A}) \geq k+2$ and again Theorem 1 applies. ||

Evolution in this model may be regarded as "fast" in that the ε-exponent of the waiting time does not increase in the size of the population. The way in which evolution toward A typically occurs is that a small cluster of players playing A arises randomly, and playing A then spreads contagiously as players at the edge of the cluster choose to join in.

5.2. *A two-dimensional model*

Consider now a local interaction model in which players are situated at the vertices of a two-dimensional lattice and interact with their four nearest neighbours.[19] The best response dynamics of such a model are quite different from those of the one-dimensional model. Small clusters of players playing a $\frac{1}{2}$-dominant equilibrium need not grow contagiously, and there may be many heterogeneous steady states. The main result of this subsection is that coordinating on $\frac{1}{2}$-dominant equilibria is nonetheless still selected and that evolution remains fast.

To specify the model formally, suppose that $N_1 N_2$ players are located at the vertices of an $N_1 \times N_2$ lattice on the surface of a torus and are repeatedly matched to play a finite symmetric game G with strategy set S. Let a state $z \in Z$ of the system be a function $z: \{1, \ldots, N_1\} \times \{1, \ldots, N_2\} \to S$ with $z(i, j)$ interpreted as being the action taken by the player at location $\{i, j\}$. Let the unperturbed dynamics on Z be generated by assuming that in each period each player plays a deterministic best response to the strategies used by his *four* immediate neighbours in the previous period, *i.e.* with probability one z_t is followed by a state $b(z_t)$ which satisfies $b(z_t)(i, j) \in \text{Argmax}_{s \in S} g(s, \sigma_{ijt})$ where σ_{ijt} is the distribution putting probability $\frac{1}{4}$ on each of $z_t(i-1, j)$, $z_t(i+1, j)$, $z_t(i, j-1)$, and $z_t(i, j+1)$. Again take the perturbed process to be that which incorporates independent ε-probability mutations.

Corollary 3. *Suppose* (A, A) *is a* $\frac{1}{2}$*-dominant equilibrium of* G*. For* $N_1 > 3$ *and* $N_2 > 3$*, the limit distribution* μ^* *corresponding to the best-reply dynamics of the nearest neighbour matching model on an* $N_1 \times N_2$ *torus has* $\mu^*(\vec{A}) = 1$*, and for any* $z \neq \vec{A}$ *we have* $W(z, \vec{A}, \varepsilon) = O(\varepsilon^{-3})$ *as* $\varepsilon \to 0$.

Proof. First, I show that $R(\vec{A}) \geq \min(N_1, N_2)$. Let $g(z)$ be the smaller of the number of rows in which all players play A and the number of columns in which all players play

19. See Blume (1993) for a discussion of a two-dimensional Ising model which has a somewhat analogous conclusion about long run behaviour, and Anderlini and Ianni (1996) and Blume (1995) for analyses of non-ergodic models with two-dimensional interaction structures.

A. If a given row (column) has all players playing A in z, then all players in that row (column) have at least two neighbours playing A, and hence all of them play A in $b(z)$ as well. From this, we know that $g(b(z)) \geq g(z)$. Further, as each "mutation" can only break up one row and one column, we know that $g(z') \geq g(z) - c(z, z')$. Applying this relationship repeatedly, it follows that if $c(\vec{A}, z_2, \ldots, z_T) < \min(N_1, N_2)$ then $g(z_T) \geq 1$. To establish that there is no path from $\vec{A}$ to $Z - D(\vec{A})$ with cost less than $\min(N_1, N_2)$, it thus suffices to show that $g(z) \geq 1 \Rightarrow z \in D(\vec{A})$.

The intuition for why this is true is that if all players in a "cross" pattern play A, the set of players playing A will expand out from the centre of the cross until it encompasses the entire population. Formally, suppose $g(z) \geq 1$ and assume without loss of generality that $z(i, 1) = z(1, j) = A$ for all $i \in \{1, \ldots, N_1\}$ and all $j \in \{1, \ldots, N_2\}$. If in addition $z(i, j) = A$ for all (i, j) with $i + j \leq k$, it is easy to see that $b(z)$ also has the same "cross" of players playing A and has $b(z)(i, j) = A$ for all i, j with $i + j \leq k + 1$. The conclusion that $g(z) \geq 1 \Rightarrow z \in D(\vec{A})$ then follows by induction.

To complete the proof it now suffices to show that $CR^*(\vec{A}) \leq 3$. To do so, I explicitly construct for each $x \notin D(\vec{A})$ a path from x to $D(\vec{A})$. Let $z_1 = x$. Let $z_t = b^{t-1}(z_1)$ for $t = 2, \ldots, k_1 + 1$, where k_1 is the minimum index such that $b^{k_1}(z_1)$ is a member of a limit set. Each transition so far has cost zero. Let z_{2+k_1} be a state in which players $(1, 1)$ and $(2, 2)$ both play A, and in which the rest of the population plays as in the state $b(z_{1+k_1})$. Note that $c(z_{1+k_1}, z_{2+k_1}) \leq 2$.

Next, let $z_{t+k_1+2} = b^t(z_{2+k_1})$ for $t = 0, \ldots, k_2$, where k_2 is the minimum index such that $b^t(z_{2+k_1})$ is a member of a limit set. Again each of these transitions have cost zero. Because players $(1, 2)$ and $(2, 1)$ each have two neighbours playing A when t is even, and players $(1, 1)$ and $(2, 2)$ each have two neighbours playing A when t is odd, at least one of these pairs plays A in period $2 + k_1 + k_2$. Let $z_{2+k_1+k_2+1}$ be defined so that all of players $(1, 1)$, $(1, 2)$, $(2, 1)$, and $(2, 2)$ play A with the remainder of the players playing as in $b(z_{2+k_1+k_2})$. This transition adds at most one to the modified cost because at most two mutations are involved and the transition escapes a limit set (provided mutations are necessary at this step; otherwise skip this last transition and also set k_3 equal to zero).

Repeating the process of adding successive best responses we obtain a state $z_{2+k_1+k_2+k_3}$ belonging to a limit set in which all four of those players play A. The state $z_{2+k_1+k_2+k_3}$ must either have all players in rows 1 and 2 playing A, or have players $(1, 1), \ldots, (1, a)$ and $(2, 1), \ldots, (2, a)$ playing A for some $a \geq 2$, with at least one of $z_{2+k_1+k_2+k_3}(1, a+1)$ and $z_{2+k_1+k_2+k_3}(2, a+1)$ different from $\vec{A}$. Let $z_{2+k_1+k_2+k_3+1}$ be defined by adding a single mutation to $b(z_{2+k_1+k_2+k_3})$ so that an extra one of the two players mentioned above plays A. This and all future transitions involving single mutations will not add to the modified cost of the path because the path at the same time exits another limit set (whose radius is 1). Let $z_{2+k_1+k_2+k_3+2}, \ldots, z_{2+k_1+k_2+k_3+k_4}$ be a sequence of best responses such that $z_{2+k_1+k_2+k_3+k_4}$ is again a member of a limit set. This limit set will have at least one of players $(1, a+1)$ and $(2, a+1)$ playing A if $z_{2+k_1+k_2+k_3}$ had neither playing A, and will have both playing A if $z_{2+k_1+k_2+k_3}$ had one of them playing A.

Repeating this process, we eventually obtain a path of modified cost at most three to a limit set in which all players in the first two rows play A. Following the same process of adding single mutations to the existing cluster, but now working on the first two columns, we obtain a path of modified cost at most three to a state in which all players in the first two columns also play A. By the result above on "crosses", such a state is in the basin of attraction of $\vec{A}$. Hence, $CR^*(\vec{A}) \leq 3$. ||

The two-dimensional local interaction model may have many limit sets. For example, if G is a 2×2 coordination game where the players receive a payoff of two from coordinating on A, one from coordinating on B and zero from miscoordinating, it is a steady state

for a cluster of four adjacent players (in a "square" configuration) to play A while the rest of the population plays B. A large number of steady states can then be obtained by having variously sized separated blocks of players play A while the remainder of the population plays B, or by having vertical or horizontal "stripes" where A is played. While this would greatly complicate an attempt to explicitly construct a minimum-order tree on the set of limit sets, it is precisely the kind of situation where the modified coradius is most useful.

What happens in the dynamics of the model is that rather than growing contagiously, small clusters of players playing the risk-dominant equilibrium grow by agglomeration as mutations at the edge of the clusters cause new players to join them. Fast evolution, however, is not so much about the contagious spread of strategies as it is about the ability of strategies to gain footholds in small areas. While not as fast as a contagion, the step-by-step agglomerative growth after a cluster forms does not add significantly to the overall length of the evolutionary process.

6. RELATIONSHIPS WITH THE LITERATURE

One of the primary motivations for this paper was a desire to develop a framework which would provide a more thorough understanding of the behaviour of evolutionary models with noise. It may thus be useful to discuss in some detail the relationship between this paper and the existing literature. In this section I discuss the extent to which the main theorem of this paper is applicable to models which have been previously studied, how the techniques of this paper build on previous work, and an alternate "tree-surgery" proof of part of the main theorem.

6.1. *Applications*

The question of where the framework is applicable can be interpreted in two very different ways: asking for which models in the literature the long-run stochastically stable set could have been identified using the main theorem of this paper, or alternatively asking where the theorem further makes the identification of the long-run stochastically stable set easier (or more general) than it would have been with the standard techniques.

The first interpretation is relevant when one wants to know when the radius-modified coradius theorem can be used to add convergence rates to existing characterizations or when it could have been used to identify the long-run stochastically stable set in the first place. With this interpretation, the theorem turns out to be applicable to almost every model for which the long-run stochastically stable set has previously been identified.[20] For example, this includes all of the cases where the long-run stochastically stable set is identified in Kandori, Mailath and Rob (1993), Ellison (1993), Evans (1993), Nöldeke and Samuelson (1993, 1997), Robson and Vega-Redondo (1996), Samuelson (1994) and Young (1993*a*, *b*).[21]

Cases where the theorem is not applicable include both models which do not fit into the framework of this paper and models which do fit but for which the theorem lacks

20. One should keep in mind here that this definition of applicability allows me to include many models where the easiest way to see that the main theorem of this paper applies, involves essentially constructing the minimum order Freidlin–Wentzell tree and reading the radius and modified coradius off of the tree diagram.

21. Most of these papers, in fact, require only a $R > CR$ theorem. Young (1993*b*) and the 3×3 game in Young (1993*a*) are examples where the modified coradius is necessary and can be easily determined from the minimum cost tree. The results of the Samuelson and Nöldeke–Samuelson analyses are most easily rederived by showing that their lemmas follow from a modified coradius computation.

power. Examples of the first type include models where the noise is sufficiently restricted so as to render the model non-ergodic (as in Anderlini and Ianni (1996) and some of Nöldeke and Samuelson's (1997) signalling games), and papers which employ variants of the solution concept (as in Binmore and Samuelson (1997) who look at an $N \to \infty$ limit). While it is easy to construct examples of Markov processes for which the radius-modified coradius calculation lacks power (as was done at the end of Section 4), such examples appear to be quite rare in the literature—the only examples I know of are a few of the differentiated price oligopoly games discussed in Kandori and Rob (1995).

To help assess whether the techniques of this paper may also facilitate the analysis of more complex models, it is useful also to discuss "applicability" in terms of where the techniques of this paper make the analysis easier. The cleanest examples I have found of this type are, not coincidentally, those that I have discussed in this paper. The result on $\frac{1}{2}$-dominance with its trivial proof generalizes several previous results including KMR's original theorem on 2×2 games, Kandori and Rob's (1995) result that Pareto optimal equilibria are selected in pure coordination games, and Kandori and Rob's (1993) conclusion that risk dominance is a sufficient condition for being long-run stochastically stable in games satisfying their total bandwagon and monotone share properties.[22] Similarly, the result on one dimensional local interaction models generalizes the analysis of 2×2 games in Ellison (1993).[23]

6.2. *Relationships with other papers*

The analysis of this paper combines and builds on several ideas which have been developed in previous papers. The single most basic idea behind the main theorem of this paper—that long-run stochastic stability is closely related to waiting times necessary for transitions between limit sets—is new as a basis for an algorithm, but has been clearly recognized as an intuitive description of long-run stochastic stability from the very beginning (see *e.g.* KMR).

While the focus on waiting times rather than trees is new, the basic intuitions behind the calculations—that evolution with noise tends to favour limit sets with larger basins of attraction, and that the presence of intermediate steady states may facilitate evolution—each have predecessors in the literature. That a set with a large enough basin of attraction is selected has been noted in a number of places. An explicit statement of the fact that $R(\Omega) > CR(\Omega)$ is a sufficient condition for long-run stochastic stability was independently given by Evans (1993) (see Lemma 3.1), with a tree surgery argument which establishes the result (though it is not stated explicitly) contained in the proof of Theorem 1 of Ellison (1993). Other authors have noted that the limit set with the largest basin of attraction is the long-run stochastically stable set in particular contexts, *e.g.* Kandori and Rob (1993) note that this is the case in 3×3 games satisfying their total bandwagon and monotone share properties. That the presence of intermediate steady states may speed evolution has not been stated so explicitly, but clearly plays a prominent role in the work of Samuelson (1994) and Nöldeke and Samuelson (1993, 1997). In those papers, the long-run stochastically stable set is identified using a lemma which (in my words) states that a component Ω of limit sets receives probability one in the limit distribution if $R(\Omega) > 2$ and Ω can be reached from any other limit set via a chain of single mutations (a condition which ensures

22. The result also generalizes the result of Maruta (1997) (which was developed later, but without knowledge of this paper) that $\frac{1}{2}$-dominant equilibria are selected in coordination games.

23. Here it is instructive also to compare the tree-surgery proof in Ellison (1993) with the fully constructive proof which was given in the working paper version of that paper only for 2-neighbour interaction.

that $CR^*(\Omega) = 1$). This paper adds to this lemma the idea that a path deserves "credit" for $R(L)$ mutations (rather than one) when passing through the limit set L and combines it with the previous one so that it can be applied in models where the nonselected limit sets are not so unstable as to be upset by just a single mutation.

The final departure of this paper from the literature is its provision of a general description of medium-run behaviour in the form of a characterization of waiting times. While no previous papers have provided any general discussions of medium-run behaviour, the topic has been discussed in a few particular models. Ellison (1993) argues that convergence rates are an important consideration and discusses convergence rates both via an eigenvalue computation and in terms of $N \to \infty$ asymptotics of waiting times. The two subsequent papers which explicitly discuss convergence rates, those of Binmore and Samuelson (1997) and Robson and Vega-Redondo (1996), both take the approach of using minimum T-period transition probabilities to bound waiting times. Though neither paper states a theorem at a level of generality which is greater than that necessary to apply to the model in question, the arguments they give are easily generalized to establish that the waiting time to reach a long-run stochastically stable state Ω which satisfies $R(\Omega) > CR(\Omega)$ is at most $O(\varepsilon^{-CR(\Omega)})$.

6.3. *A Freidlin–Wentzell "tree surgery" argument*

While I have presented the radius-modified coradius theorem as exploiting insights gained from thinking about waiting times, for those familiar with the previous literature it may also be useful to note that a result similar to the long-run stochastic stability part of the theorem can also be derived using a "tree-surgery" argument. The proof is not too involved, and thus might be used as a starting point when trying to analyse models to which the theorem of this paper fails to apply.

Theorem 3. *Let $(Z, P, P(\varepsilon))$ be a model of evolution with noise. If for some limit set Ω and some state $x \notin \Omega$ we have $R(\Omega) > c^*(x, \Omega)$ then $\mu^*(x) = 0$. If $R(\Omega) = c^*(x, \Omega)$ then $\mu^*(x) > 0$ implies $\mu^*(\Omega) > 0$.*

It should be noted that unlike Theorems 1 and 2, the Ω in the statement of this proposition is required to be a single limit set rather than a union of limit sets. When Ω is a single limit set, part (a) of Theorem 2 follows from the first conclusion of Theorem 3 because $c^*(z, \Omega) \leq CR^*(\Omega)\ \forall z \notin \Omega$, and hence the first conclusion implies $\mu^*(z) = 0\ \forall z \notin \Omega$, which in combination with the fact that $\mu^*(Z) = 1$ implies that $\mu^*(\Omega) = 1$. Theorem 3 adds to Theorem 2 a partial characterization which may be informative even when there is no set Ω with $R(\Omega) > CR^*(\Omega)$.

Proof. An x-tree t is a function $t: Z \to Z$ such that $t(x) = x$ and such that $\forall z \neq x$ there exists k with $t^k(z) = x$. Define the cost of an x-tree t, $c(t)$, by $c(t) = \sum_{z \neq x} c(z, t(z))$. A consequence of Freidlin and Wentzell's results is that long-run stochastically stable states correspond to minimum cost trees. Specifically, x is *not* long-run stochastically stable if for some $w \neq x$ there exists a w-tree t such that all x-trees t' have $c(t') > c(t)$.

To show that $x \notin \Omega$ is not in the long-run stochastically stable set when $R(\Omega) > c^*(x, \Omega)$, I show how, given an x-tree t' one can (for some $w \in \Omega$) construct a w-tree t'' which has a strictly lower cost. Suppose t' is an x-tree. Let $(z_1 (= x), z_2, \ldots, z_T (= w))$ be a path from x to Ω with minimum modified cost. This cost is at most $c^*(x, \Omega)$.

Let $(L_1, L_2, \ldots, L_r = \Omega)$ be the set of limit sets crossed along this path. With an appropriate choice of path we may assume that these limits sets are distinct. Note that for any limit set Ω' of (Z, P) and for any state $w' \in \Omega'$ we may define a w'-tree b: $D(\Omega') \to D(\Omega')$ such that $c(z, b(z)) = 0 \ \forall z \neq w'$. Let $Y = \bigcup_{i=1}^{r} D(L_i)$. Combining several maps like that above we may choose b: $Y \to Y$ such that for some $k > 0$ we have $b^k(z) \in \{z_1, z_2, \ldots, z_T\}$ for all $z \in Y$, such that $c(z, b(z)) = 0$ if $z \notin \{z_1, z_2, \ldots, z_T\}$, and such that $b(w) = w$.

Define a w-tree t'' by

$$t''(z) = \begin{cases} z_{t+1} & \text{if } z = z_t \text{ for some } t \in \{1, 2, \ldots, T-1\}, \\ b(z) & \text{if } z \in Y,\ z \notin \{z_1, z_2, \ldots, z_{T-1}\}, \\ t'(z) & \text{otherwise.} \end{cases}$$

To see that this is a w-tree note that for any state z the path $(z, t''(z), t''^2(z), \ldots)$ coincides with that of t' until it reaches either an element of Y or an element of $\{z_1, z_2, \ldots, z_T\}$. Because t' is an x-tree (and $z_1 = x$), this ensures that one of these sets will be reached. If the second set is reached, the path clearly leads to w. If the first set is reached then, by the definition of b, the second set will be reached within k periods and again the path leads to w.

Because the trees t' and t'' are identical at all states covered by the "otherwise" line of the definition of t'', we may cancel the costs of transition for all such states when comparing the costs of t' and t''. In addition, we know $c(z, b(z)) = 0 \ \forall z \in Y$ with $z \notin \{z_1, z_2, \ldots, z_T\}$. Using these facts we can write

$$c(t'') - c(t') = \sum_{t=1}^{T-1} c(z_t, z_{t+1}) - \sum_{z \in Y} c(z, t'(z)).$$

Now, consider each of the two sums on the right-hand side of this expression. The first is at most $c^*(x, \Omega) + \sum_{i=2}^{r-1} R(L_i)$. To bound the second, note that, because t' is an x-tree, corresponding to any limit set Ω' with $x \notin D(\Omega')$ there exists a $k > 0$ and a $w' \in \Omega'$ such that $(w', t'(w'), \ldots, t'^k(w'))$ is a path from Ω' out of $D(\Omega')$. The cost of such a path is at least $R(\Omega')$. If $\Omega' \in \{L_2, \ldots, L_r = \Omega\}$, the cost of each of the transitions in such a path appears as a term in the second sum on the right-hand side. Further, the terms are distinct for distinct limit sets. Hence, that sum is at least $R(\Omega) + \sum_{i=2}^{r-1} R(L_i)$. Hence, we know

$$c(t'') - c(t') \leqq c^*(x, \Omega) + \sum_{i=2}^{r-1} R(L_i) - (R(\Omega) + \sum_{i=2}^{r-1} R(L_i)) < 0.$$

Thus, t'' is a lower cost w-tree as desired.

As for the second conclusion of the theorem, note simply that if $c^*(x, \Omega) = R(\Omega)$, then the argument above shows that the minimum cost x-tree is no cheaper than a w-tree for some $w \in \Omega$. ||

7. CONCLUSION

In this paper, I have discussed the behaviour of stochastic models both in general and in several particular examples. With regard to the former, the paper outlines an approach which involves describing the basins of attraction with two new measures, the radius and coradius. The main theorem is applicable to many of the models which have been previously studied and expands our understanding of these models in two ways: it provides a clear intuitive argument which may eliminate much of the mystery left in the wake of Freidlin–Wentzell tree constructions and provides a measure of the rate at which a model converges. The approach is tractable in the examples discussed here despite the fact that

the games involved may have best response cycles, and that in one case (the two-dimensional local interaction model) the dynamics are so complex as to render even a listing of the stable limit sets difficult. As the literature on stochastic evolution continues to grow, economic interest continues to draw researchers to more complex games. Among the recent notable examples are Young's (1993*b*) analysis of bargaining, Nöldeke and Samuelson's (1997) analysis of signalling games, and Binmore and Samuelson's (1997) "muddling" model. I hope that the argument of this paper will spur further research into such topics in the future both by reducing the burden of carrying out proofs and by making analyses more complete and more transparent.

This paper is also a paper about models of evolution with noise. In this area, the primary result of the paper is that the selection of risk dominant equilibria in 2×2 games generalizes to the selection of $\frac{1}{2}$-dominant equilibria whenever they exist. Other examples illustrate that the long-run stochastically stable outcome need not involve Nash equilibrium play, even when the Nash equilibrium of a game is unique, and that despite lacking contagion dynamics, models with two dimensional local interaction can feature rapid convergence.

The paper may be of more general interest for the insight it provides into the circumstances in which evolutionary change is likely to be rapid. The argument that shifts from one equilibrium to another are most likely to be observed in systems which are amenable to gradual change may be applicable in a wide variety of economic contexts—both where gradual change takes the form of shifts which occur first in small subsets of the population and where it involves a continuous variable adjusting slowly between two extremes.

In the future, the results of this paper might be extended in a number of directions. Among the most promising topics to explore are the possibility of extending the applicability of the theorem by grouping limit sets, the possibility of developing tighter bounds on waiting times by taking advantage of specific features of the dynamics rather than always assuming the worst case, and the use of tree-surgery arguments in situations where the main theorem of this paper does not apply.

APPENDIX

Lemma 1. *Suppose* $(Z, P, P(\varepsilon))$ *is a model of evolution with noise. If* $y \in Z$ *and* $\Omega \subset Z$ *with* $y \notin \Omega$ *then*

$$\frac{\mu^\varepsilon(y)}{\mu^\varepsilon(\Omega)} = \frac{N(y, \Omega, y)}{\sum_{\omega \in \Omega} Q(\omega, \Omega - \omega, y) N(\Omega, y, \omega)}.$$

Proof. It is a standard result (see *e.g.* Theorem 6.2.3 of Kemeny and Snell (1960)) that for any two states y and ω of an ergodic Markov process $\mu^\varepsilon(\omega) = \mu^\varepsilon(y) N(\omega, y, y)$. Summing over all $\omega \in \Omega$ gives $\mu^\varepsilon(\Omega) = \mu^\varepsilon(y) N(\Omega, y, y)$. Hence,

$$\begin{aligned}
\frac{\mu^\varepsilon(\Omega)}{\mu^\varepsilon(y)} &= N(\Omega, y, y) \\
&= \sum_{\omega \in \Omega} Q(\omega, y \cup \Omega - \omega, y) N(\Omega, y, \omega) \\
&= \sum_{\omega \in \Omega} Q(\omega, \Omega - \omega, y) Q(\Omega, y, y) N(\Omega, y, \omega) \\
&= Q(\Omega, y, y) \sum_{\omega \in \Omega} Q(\omega, \Omega - \omega, y) N(\Omega, y, \omega).
\end{aligned}$$

The desired result is then a consequence of the fact that

$$N(y, \Omega, y) = 1/Q(\Omega, y, y). \quad \|$$

Lemma 2. *Suppose $(Z, P, P(\varepsilon))$ is a model of evolution with noise and that Ω is a union of limit sets of (Z, P). Then, for any $\omega \in \Omega$ and any $y \notin D(\Omega)$*

$$\frac{1}{N(\Omega, y, \omega)} \leq \frac{1}{N(\Omega, Z - D(\Omega), \omega)} = O(\varepsilon^{R(\Omega)}).$$

Proof. The first inequality is obvious. To prove the second note first that because

$$\min_{\omega \in \Omega} N(\Omega, Z - D(\Omega), \omega) \geq \max_{\omega' \in \Omega} Q(Z - D(\Omega), \Omega, \omega') \cdot 1 + \left(1 - \max_{\omega' \in \Omega} Q(Z - D(\Omega), \Omega, \omega')\right)\left(1 + \min_{\omega \in \Omega} N(\Omega, Z - D(\Omega), \omega)\right),$$

we have

$$\frac{1}{N(\Omega, Z - D(\Omega), \omega)} \leq \max_{\omega' \in \Omega} Q(Z - D(\Omega), \Omega, \omega').$$

For $z \in Z$ define $e(z) = C(z, Z - D(\Omega))$. Note that $e(z) = 0$ if $z \notin D(\Omega)$, $e(z) \geq R(\Omega)$ if $z \in \Omega$, and that $C(z, z') - e(z') \geq e(z)$ for all z, z'.

The desired result follows if we show that

$$Q(Z - D(\Omega), \Omega, z) = O(\varepsilon^{e(z)}) \qquad \forall z \in Z.$$

I establish this by showing inductively for $k = 0, 1, 2, \ldots, R(\Omega)$ that

$$\max_{z | e(z) \geq k} Q(Z - D(\Omega), \Omega, z) = O(\varepsilon^k).$$

For $k = 0$ the result is trivial because probabilities are always bounded above by one.

Suppose now that the result holds for $k = 0, 1, 2, \ldots, s$. For each z with $e(z) \geq s + 1$ let $E_{z,s-1} = \{\varepsilon \in (0, \infty) | z \in \arg\max_{z | e(z) \geq s+1} Q(Z - D(\Omega), \Omega, z)$.[24] Suppose z' is a state such that $E_{z',s-1}$ is not bounded away from zero. Choose T such that there is a positive probability of reaching Ω from z' in period T without leaving $D(\Omega)$ in the unperturbed process (Z, P). Write $Q(A, B, x, T)$ for the probability that A is reached before B when the process starts at x (again not counting what happens in the initial period if $x \in A$ or $x \in B$) and that this occurs in the first T periods. Then for $\varepsilon \in E_{z',s+1}$ we have

$$Q(Z - D(\Omega), \Omega, z') = Q(Z - D(\Omega), \Omega, z', T) + \sum_{z \in D(\Omega) - \Omega} r_{z'z} Q(Z - D(\Omega), \Omega, z),$$

where $r_{z'z}$ is the probability of going from z' to z in exactly T periods without hitting either Ω or $Z - D(\Omega)$. Subtracting $\sum_{z \in D(\Omega) - \Omega, e(z) \geq s+1} r_{z'z} Q(Z - D(\Omega), \Omega, z')$ from both sides and using the fact that $Q(Z - D(\Omega), \Omega, z') \geq Q(Z - D(\Omega), \Omega, z)$ for all other z gives

$$\left(1 - \sum_{z \in D(\Omega) - \Omega, e(z) \geq s+1} r_{z'z}\right) Q(Z - D(\Omega), \Omega, z') \leq Q(Z - D(\Omega), \Omega, z', T) + \sum_{z \in D(\Omega) \quad \Omega, e(z) \leq s} r_{z'z} Q(Z - D(\Omega), \Omega, z). \tag{A1}$$

for all $\varepsilon \in E_{z',s+1}$. Because the probability of reaching Ω from z' in period T is positive in the unperturbed model, $\left(1 - \sum_{z \in D(\Omega) - \Omega, e(z) \geq s+1} r_{z'z}\right)$ is bounded away from zero as $\varepsilon \to 0$.

Because $Q(Z - D(\Omega), \Omega, z', T)$ is the sum of the probabilities of a number of T period or shorter paths from z' to $Z - D(\Omega)$ it is $O(\varepsilon^{C(z', Z - D(\Omega))}) = O(\varepsilon^{s+1})$.

For any $z \in D(\Omega) - \Omega$ with $e(z) \leq s$, $r_{z'z}$ is similarly $O(\varepsilon^{C(z',z)}) = O(\varepsilon^{e(z') - e(z)})$ and by the inductive hypothesis $Q(Z - D(\Omega), \Omega, z) = O(\varepsilon^{e(z)})$.

Substituting each of these into (A1) we find that $\max_{z | e(z) \geq s+1} Q(Z - D(\Omega), \Omega, z) = Q(Z - D(\Omega), \Omega, z') = O(\varepsilon^{s+1})$ whenever $\varepsilon \in E_{z',s+1}$. Because the state space is finite, we can take a union over all sets $E_{z',s+1}$ and conclude that

$$\max_{z | e(z) \geq s+1} Q(Z - D(\Omega), \Omega, z) = O(\varepsilon^{s+1}).$$

The desired result follows by induction. ||

24. Recall that Q and N are always functions of ε although we have suppressed the ε argument to shorten the equations.

Lemma 3. *Suppose* $(Z, P, P(\varepsilon))$ *is a model of evolution with noise and that* L *is a limit set of* (Z, P). *Then,* $W(l, Z - D(L), \varepsilon) = O(\varepsilon^{-R(L)})$ *for all* $l \in L$.

Proof. Given L we can find a T and a $k > 0$ such that for any $z \in D(L)$ there exists a path $z = z_1, z_2, \ldots, z_T$ with $z_T \in Z - D(L)$ such that the product of the transition probabilities along the path is at least $k\varepsilon^{R(L)}$. Conditioning on the outcome of the first T periods we have for any $z \in D(L)$ that

$$W(z, Z - D(L), \varepsilon) \leq T + (1 - k\varepsilon^{R(L)}) \max_{z' \in D(L)} W(z', Z - D(L), \varepsilon).$$

Taking the maximum of the LHS over $z \in D(L)$ gives

$$\max_{z \in D(L)} W(z, Z - D(L), \varepsilon) \leq (T/k)\varepsilon^{-R(L)}$$

as desired. ||

Lemma 4. *Suppose* $(Z, P, P(\varepsilon))$ *is a model of evolution with noise. Let* $\mathcal{L}$ *be the union of the limit sets of* (Z, P) *and suppose* L *is a single limit set. Then for any* $l \in L$, $W(l, \mathcal{L} - L, \varepsilon) = O(\varepsilon^{-R(L)})$.

Proof. For each $l \in L$ let E_l be the set of values of ε for which $W(l, \mathcal{L} - L, \varepsilon) = \max_{l' \in L} W(l', \mathcal{L} - L, \varepsilon)$. If E_l is not empty then for all $\varepsilon \in E_l$ we have

$$W(l, \mathcal{L} - L, \varepsilon) \leq W(l, Z - D(L), \varepsilon) + \sum_{z \in Z - D(L)} Q(z, Z - D(L) - z, l) W(z, \mathcal{L}, \varepsilon)$$
$$+ \sum_{z \in Z - D(L), z \notin \mathcal{L}} Q(z, Z - D(L) - z, l) Q(L, \mathcal{L} - L, z) W(l, \mathcal{L} - L, \varepsilon).$$

Collecting terms we have

$$(1 - \sum_{z \in Z - D(L), z \notin \mathcal{L}} Q(z, Z - D(L) - z, l) Q(L, \mathcal{L} - L, z)) W(l, \mathcal{L} - L, \varepsilon)$$
$$\leq W(L, Z - D(L), \varepsilon) + \max_{z \in \mathcal{L}} W(z, \mathcal{L}, \varepsilon).$$

The first term on the LHS of the expression above is bounded away from zero because $Q(L, \mathcal{L} - L, z)$ is uniformly bounded away from one for any $z \notin D(L)$. The first term on the RHS is $O(\varepsilon^{-R(L)})$ by Lemma 3. The second term on the RHS is finite. Together these observations give that the desired result holds when $\varepsilon \in E_l$. Taking the union of these sets it holds for all ε. ||

Lemma 5. *Suppose* $(Z, P, P(\varepsilon))$ *is a model of evolution with noise. Let* $\mathcal{L}$ *be the union of the limit sets of* (Z, P), *and suppose that* L *and* L' *are two limit sets such that there exists a path from* L *to* L' *of cost* $C(L, L')$ *not passing through any other limit sets. Then for any* $l \in L$,

$$\frac{1}{Q(L', \mathcal{L} - (L \cup L'), l)} = O(\varepsilon^{R(L) - C(L, L')}).$$

Proof. Recall that $C(X, Y)$ is the minimum cost of any path from X to Y. Write $|L|$ for the number of elements of L. For $l \in L$ we have

$$Q(L', \mathcal{L} - (L \cup L'), l) \geq \frac{1}{|L|} \sum_{t=1,\ldots,\infty, l' \in L} \text{Prob}\{\{z_1, \ldots, z_t\} \cap (\mathcal{L} - L) = \emptyset, z_t = l', L' \text{ is reached and this occurs before } \mathcal{L} - (L \cup L') \text{ is reached, and } z_{t'} \in L \text{ in at most } |L| \text{ periods after } t \text{ before } L' \text{ is reached} | z_1 = l\},$$

because the probability of each path reaching L' before $\mathcal{L} - (L \cup L')$ is counted at most $|L|$ times in the summation on the RHS. This gives

$$Q(L', \mathcal{L} - (L \cup L'), l) \leq \frac{1}{|L|} \sum_{t=1,\ldots,\infty, l' \in L} \text{Prob}\{\{z_1, \ldots, z_t\} \cap (\mathcal{L} - L) = \emptyset, z_t = l' | z_1 = l\}$$
$$\cdot \text{Prob}\{L' \text{ is reached before } \mathcal{L} - (L \cup L') \text{ and at most } |L| \text{ periods are spent in } L | z_1 = l'\}.$$

Because there exists a path from L to L' of cost at most $C(L, L')$ which does not pass through any other limit sets the second terms on the RHS above are uniformly bounded below by $k\varepsilon^{C(L, L')}$ for ε small for some $k > 0$. The summation over t and l' of the first terms on the RHS is bounded below by $N(L, \mathcal{L} - L, l)$. Hence

we have

$$Q(L', \mathcal{A}-(L\cup L'), l) \geq \frac{1}{|L|} N(L, \mathcal{A}-L, l) k\varepsilon^{C(L,L')}$$

$$\geq \frac{1}{|L|} N(L, Z-D(L), l) k\varepsilon^{C(L,L')}$$

$$\geq k'\varepsilon^{C(L,L')-R(L)},$$

for some $k'>0$ using the result of Lemma 2. ||

Lemma 6. *Let $(Z, P, P(\varepsilon))$ be a model of evolution with noise and suppose that Ω is a union of limit sets of (Z, P). Then*

$$\max_{y\notin\Omega} W(y, \Omega, \varepsilon) = O(\varepsilon^{-CR^*(\Omega)}).$$

Proof. Write $\mathcal{A}$ for the union of the model's limit sets. First, note that for $y\notin\mathcal{A}$

$$W(y, \Omega, \varepsilon) = W(y, \mathcal{A}, \varepsilon) + \sum_{l\in\mathcal{A}-\Omega} Q(l, \mathcal{A}-l, y) W(l, \Omega, \varepsilon).$$

The first term on the RHS has a finite limit as $\varepsilon\to 0$, so it will suffice to establish that

$$\overline{W}(\Omega, \varepsilon) \equiv \max_{y\in\mathcal{A}-\Omega} W(y, \Omega, \varepsilon) = O(\varepsilon^{-CR^*(\Omega)}).$$

Given $y\in\mathcal{A}-\Omega$ let $y=z_1, z_2, \ldots, z_T$ be a path of modified cost $c^*(y, \Omega)$. We may choose this path so that the limit sets $L_1, L_2, \ldots, L_r$ through which it passes are distinct and so that the path is contained in each of these limit sets for a set of successive periods. (Given any path from y to Ω one can always obtain a path of equal or lower modified cost by replacing segments between the first and last visits to each limit set and between the first and last period in each limit set by paths which remain within the limit set in question and have zero cost.)

Because $y\in L_1$ we have

$$W(y, \Omega, \varepsilon) = W(y, \mathcal{A}-L_1, \varepsilon) + \sum_{z\in\mathcal{A}-(L_1\cup\Omega)} Q(z, \mathcal{A}-(L_1\cup\Omega\cup\{z\}), y) W(z, \Omega, \varepsilon).$$

Write q_{12} for $\min_{y'\in L_1} Q(L_2, \mathcal{A}-(L_1\cup L_2), y')$ and $W(A, B, \varepsilon)$ for $\max_{x\in A} W(x, B, \varepsilon)$. If y is the element of $\mathcal{A}-\Omega$ for which the wait to reach Ω is largest for some subset of ε values, then on this set of ε values we have

$$\overline{W}(\Omega, \varepsilon) \leq W(y, \mathcal{A}-L_1, \varepsilon) + q_{12} W(L_2, \Omega, \varepsilon) + (1-q_{12})\overline{W}(\Omega, \varepsilon).$$

Repeating the argument above to bound successively $W(L_2, \Omega, \varepsilon)$, $W(L_3, \Omega, \varepsilon)$ *etc.* we find

$$\begin{aligned}
\overline{W}(\Omega, \varepsilon) \leq\ & W(L_1, \mathcal{A}-L_1, \varepsilon) + (1-q_{12})\overline{W}(\Omega, \varepsilon) \\
& + q_{12}(W(L_2, \mathcal{A}-L_2, \varepsilon) + (1-q_{23})\overline{W}(\Omega, \varepsilon)) \\
& + q_{12}q_{23}(W(L_3, \mathcal{A}-L_3, \varepsilon) + (1-q_{34})\overline{W}(\Omega, \varepsilon)) \\
& + \cdots \qquad \cdots \\
& + q_{12}q_{23}\cdots q_{r-2,r-1}(W(L_{r-1}, \mathcal{A}-L_{r-1}, \varepsilon) + (1-q_{r-1,r})\overline{W}(\Omega, \varepsilon)) \\
=\ & W(L_1, \mathcal{A}-L_1, \varepsilon) + q_{12} W(L_2, \mathcal{A}-L_2, \varepsilon) + q_{12}q_{23} W(L_3, \mathcal{A}-L_3, \varepsilon) \\
& + \cdots + (1-q_{12}q_{23}\cdots q_{r-1,r})\overline{W}(\Omega, \varepsilon)
\end{aligned}$$

Collecting terms this gives

$$\overline{W}(\Omega, \varepsilon) \leq \frac{W(L_1, \mathcal{A}-L_1, \varepsilon)}{q_{12}q_{23}\cdots q_{r-1,r}} + \cdots + \frac{W(L_{r-1}, \mathcal{A}-L_{r-1}, \varepsilon)}{q_{r-1,r}}.$$

It now suffices to show that each of the expressions on the RHS of the above equation is $O(\varepsilon^{-CR^*(\Omega)})$.

Because the path $z_1, \ldots, z_T$ had the minimum possible modified cost, its segment connecting L_j and L_{j+1} must have cost $C(L_j, L_{j-1})$. Hence by Lemma 5 we know that $1/q_{jj+1} = O(\varepsilon^{R(L_j)-C(L_j,L_{j-1})})$. Using this and the

result of Lemma 4 we have

$$\frac{W(L_i, \ldots, L_i, \varepsilon)}{q_{i,i+1}q_{i+1,i+2}\cdots q_{r-1,r}} = O(\varepsilon^{-(\sum_{j=i}^{r-1} C(L_j, L_{j+1}) - R(L_j))}\varepsilon^{-R(L_r)})$$

Because $R(L_i) + \sum_{j=i}^{r-1} C(L_j, L_{j-1}) - R(L_j))$ is the minimum modified cost of any path from L_i to Ω we have $R(L_i) + \sum_{j=i}^{r-1} C(L_j, L_{j-1}) - R(L_j)) \leqq CR^*(\Omega)$. Hence,

$$\frac{W(L_i, \ldots, L_i, \varepsilon)}{q_{i,i+1}q_{i+1,i+2}\cdots q_{r-1,r}} = O(\varepsilon^{-CR^*(\Omega)}),$$

as desired. ‖

Acknowledgements. I would like to thank Sara Fisher Ellison, Drew Fudenberg, Julian Jamison, David Levine, Eric Maskin, Roger Myerson, Georg Nöldeke, Tomas Sjöström, Peter Sorensen, the editor and three anonymous referees for their comments. Financial support was provided by National Science Foundation grant SBR-9515076 and by a Sloan Research Fellowship.

REFERENCES

ANDERLINI, L. and IANNI, A. (1996), "Path Dependence and Learning from Neighbors", *Games and Economic Behavior*, **13**, 141–177.

BERGIN, J. and LIPMAN, B. (1996), "Evolution with State-Dependent Mutations", *Econometrica*, **64**, 943–956.

BINMORE, K. and SAMUELSON, L. (1997), "Muddling Through: Noisy Equilibrium Selection", *Journal of Economic Theory*, **74**, 235–265.

BLUME, L. (1993), "The Statistical Mechanics of Strategic Interaction", *Games and Economic Behavior*, **5**, 387–423.

BLUME, L. (1995), "The Statistical Mechanics of Best-Response Strategy Revision", *Games and Economic Behavior*, **11**, 111–145.

CANNING, D. (1992), "Average Behavior in Learning Models", *Journal of Economic Theory*, **57**, 442–472.

ELLISON, G. (1993), "Learning, Local Interaction, and Coordination", *Econometrica*, **61**, 1047–1071.

ELLISON, G. (1995), "Basins of Attraction, Long Run Equilibria, and the Speed of Step-by-Step Evolution" (Working paper, MIT).

EVANS, R. (1993), "Observability, Imitation and Cooperation in the Repeated Prisoners' Dilemma" (Working paper, University of Cambridge).

FOSTER, D. and YOUNG, H. P. (1990), "Stochastic Evolutionary Game Dynamics", *Theoretical Population Biology*, **38**, 219–232.

FREIDLIN, M. and WENTZELL, A. (1984) *Random Perturbations of Dynamical Systems* (New York: Springer Verlag).

HAHN, S. (1995), "The Long Run Equilibrium in an Asymmetric Coordination Game" (Mimeo, Harvard University).

HARSANYI, J. and SELTEN, R. (1988) *A General Theory of Equilibrium Selection in Games* (Cambridge: MIT Press).

HUCK, S. and OECHSSLER, J. (1995), "The Indirect Approach to Explaining Fair Allocations" (Mimeo, Humboldt University).

KANDORI, M., MAILATH, G. and ROB, R. (1993), "Learning, Mutation, and Long Run Equilibria in Games", *Econometrica*, **61**, 29–56.

KANDORI, M. and ROB. R. (1995), "Evolution of Equilibria in the Long Run: A General Theory and Applications", *Journal of Economic Theory*, **65**, 383–414.

KANDORI, M. and ROB, R. (1993), "Bandwagon Effects and Long Run Technology Choice" (University of Tokyo Discussion Paper 93-F-2).

KEMENY, J. and SNELL, J. L. (1960) *Finite Markov Chains* (Princeton: Van Nostrand).

KIM, Y. (1996), "Equilibrium Selection in *n*-Person Coordination Games", *Games and Economic Behavior*, **15**, 203–227.

MARUTA, T. (1997), "On the Relationship between Risk-Dominance and Stochastic Stability", *Games and Economic Behavior*, **19**, 221–234.

MORRIS, S., ROB, R. and SHIN, H. (1995), "*p*-Dominance and Belief Potential", *Econometrica*, **63**, 145–157.

NÖLDEKE, G. and SAMUELSON, L. (1993), "An Evolutionary Analysis of Backward and Forward Induction", *Games and Economic Behavior*, **5**, 424–454.

NÖLDEKE, G. and SAMUELSON, L. (1997), "A Dynamic Model of Equilibrium Selection in Signaling Markets", *Journal of Economic Theory*, **73**, 118–156.

ROBSON, A. and VEGA-REDONDO, F. (1996), "Efficient Equilibrium Selection in Evolutionary Games with Random Matching", *Journal of Economic Theory*, **70**, 65–92.

SAMUELSON, L. (1994), "Stochastically Stable Sets in Games with Alternative Best Replies", *Journal of Economic Theory*, **64**, 35–65.
YOUNG, H. P. (1993*a*), "The Evolution of Conventions", *Econometrica*, **61**, 57–84.
YOUNG, H. P. (1993*b*), "An Evolutionary Model of Bargaining", *Journal of Economic Theory*, **59**, 145–168.