# Learning to Forgive*

PRELIMINARY AND INCOMPLETE

Thomas W. L. Norman
All Souls College, Oxford OX1 4AL, UK

July 4, 2006

**Abstract**

If players learn to play an infinitely repeated game using Foster and Young's (*Games and Economic Behavior* **45**, 2003, 73–96) hypothesis testing, then their strategies almost always approximate equilibria of the repeated game. If, in addition, they are sufficiently "conservative" in their hypothesis revision, then almost all of the time is spent approximating an efficient subset of "forgiving" equilibria. *Journal of Economic Literature* Classification: C72; C12.

*Key Words:* Repeated games; Folk Theorem; Hypothesis testing; Stochastic stability.

> "To err is human; to forgive, divine."
> Alexander Pope, *An Essay on Criticism* (1711)

## 1 Introduction

Game theorists' teeth are cut on the Prisoner's Dilemma:

| | $C$ | $D$ |
|---|---|---|
| $C$ | 2<br>2 | 3<br>0 |
| $D$ | 0<br>3 | 1<br>1 |

With defection a dominant strategy, and thus the unique Nash equilibrium, we are left to wonder how players might cooperate, and thus realize a Pareto improvement. Repeating the game provides an intuitive and dramatic answer; the Folk Theorem for infinitely repeated games says that, if players are sufficiently patient, all feasible individually rational stage-game payoffs can be sustained in a (Nash or subgame-perfect) equilibrium of the repeated game (Aumann 1957, Aumann and Shapley 1976, Rubinstein 1979, Fudenberg and Maskin 1986). In Figure 1,

*Thanks are due in particular to Joe Perkins and Peyton Young for their indefatigable interest, and to seminar participants at Oxford University. Email thomas.norman@all-souls.ox.ac.uk.

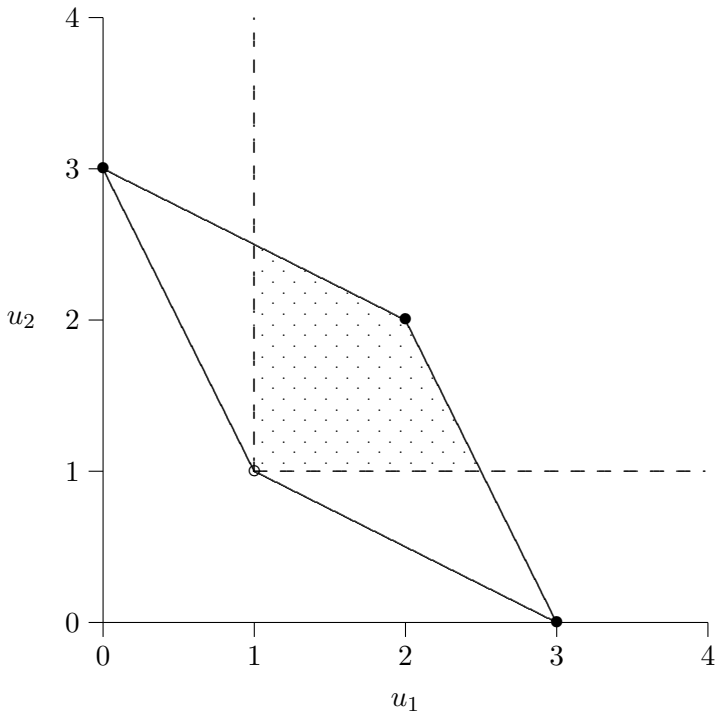


Figure 1: Equilibrium payoffs in the repeated Prisoner's Dilemma

the possibilities are thus widened from the stage Nash payoffs $(1, 1)$ to the entire shaded region.

The Folk Theorem thus provides an answer to the puzzle of cooperation in the Prisoner's Dilemma, but it also clearly leaves us with a rather profound equilibrium selection problem. Nevertheless, there is a general sense amongst practitioners that some of the equilibria attainable under the Folk Theorem are more appealing than others:

> "In applying repeated games, economists typically focus on one of the efficient equilibria, usually a symmetric one. This is due in part to a general belief that players may coordinate on efficient equilibria, and in part to the belief that cooperation is particularly likely in repeated games. It is a troubling fact that at this point there is no accepted theoretical justification for assuming efficiency in this setting."[1]

The idea of "renegotiation proofness" (Farrell and Maskin 1989, van Damme 1989, Pearce 1987, Abreu, Pearce, and Stacchetti 1993)—whereby a Pareto-dominated equilibrium in any subgame is "renegotiated" away—is one possible justification, but it sits a little uneasily with the non-cooperative approach in general, and the criticisms of Pareto optimality in static games in particular.

An alternative justification for efficiency in repeated games is provided by the evolutionary approach. Axelrod's (1981, 1984) celebrated evolutionary simulations of the repeated Prisoner's

[1] Fudenberg and Tirole (1991), p. 160.

Dilemma found selection pressure in favor of the strategy of "tit-for-tat," whereby a player cooperates in the first period and thereafter chooses the action his opponent took in the previous round. However, the outcome of such simulations is quite sensitive to the initial distribution of strategies upon which the selection process acts. On a theoretical level, meanwhile, the usual formulation of evolutionary stability suffers from severe existence problems in infinitely repeated games (Boyd and Lorberbaum 1987, Farrell and Ware 1988, Kim 1994), whilst a switch to neutral stability gives little sharpening of the predictions of the Folk Theorem. Fudenberg and Maskin (1990) and Binmore and Samuelson (1992) do find efficiency to be implied by modified versions of evolutionary stability, but such concepts too are subject to path dependence in their predictions.

The recent literature on stochastic evolution (Foster and Young 1990, Kandori, Mailath, and Rob 1993, Young 1993, Ellison 2000) offers up techniques for equilibrium selection that are insensitive to the initial distribution of strategies. The concept of stochastic stability picks out the equilibrium most likely to be played over the long-run evolution of a system made ergodic by the introduction of noise. However, such a system need not in general pick out a Nash equilibrium. Moreover, the learning interpretation of evolutionary models seems particularly strained in the case of repeated games; evolution requires a large number of repetitions of the *whole* game—a repeated repeated game, if you will—which may be unappealing in many cases.

The learning literature seems to provide the more natural analytical framework of learning over the course of a *single* repeated game. Furthermore, certain forms of convergence to equilibria of the repeated game have been demonstrated in this setting, both for Bayesian rational learning (Kalai and Lehrer 1993) and for hypothesis testing (Foster and Young 2003). The latter offers an interesting opportunity for equilibrium selection. For the noise inherent in the hypothesis-testing process means that, whilst it will spend most of its time approximating equilibria of the repeated game, it will not settle for so long on a *particular* equilibrium. Rather, any given equilibrium will be visited with a frequency determined by its attractiveness and persistence, as in the stochastic stability literature.

This paper investigates the implications of this observation for equilibrium selection in infinitely repeated games, under certain conditions on the hypothesis-testing process. In particular, it is found that, if players are sufficiently "conservative" in revising their hypotheses—in the sense that a rejected hypothesis is with high probability replaced by a "nearby" alternative—then the process spends most of its time approximating an efficient subset of equilibria of the repeated game. Furthermore, this subset has a "forgiving" property shared by a common modification of the "tit-for-tat" strategy.

## 2 Evolutionary Stability in Infinitely Repeated Games

Evolutionary stability has had limited success in selecting between the equilibria possible under the various Folk Theorems. Axelrod and Hamilton (1981) show that "always defect" is not an ESS in the repeated Prisoner's Dilemma with time-average payoffs, since it is vulnerable to invasion by tit-for-tat (though this breaks down under discounting). Axelrod (1981, 1984) argues in favor of tit-for-tat on the basis of his concept of a "collectively stable strategy," but this concept does not imply evolutionary stability and gives little sharpening of the Nash Folk Theorem. Moreover, tit-for-tat is not a subgame-perfect equilibrium strategy against itself, and thus is not even a candidate equilibrium under the perfect Folk Theorems.

Boyd and Lorberbaum (1987) show that no pure strategy can be evolutionarily stable in the infinitely repeated Prisoner's Dilemma, whilst Farrell and Ware (1988) extend this to finite mixtures of pure strategies. Kim (1994) generalizes these results to any strategies, and also to Selten's (1983) extensive-form concept of direct ESS. But Sugden (1986) and Boyd (1989) show that ESSs do exist if players occasionally make "mistakes" (as distinct from mutations)—an important notion throughout the rest of the paper. The existence problem for direct ESS is the possibility of mutation to strategies that differ from the existing ones only off the equilibrium path. Selten's notion of a limit ESS addresses this problem by perturbing the game—so that every information set is reached with positive probability—and finding the limit of a sequence of direct ESSs as the perturbations vanish. This gives a refinement of sequential equilibrium in symmetric extensive-form games (van Damme 1987). However, Kim proves that a Folk Theorem obtains for limit ESSs; the concept offers no sharpening of the predictions of subgame perfection in the infinitely repeated Prisoner's Dilemma.

A similar criticism can be levelled at the relaxation of evolutionary stability to neutral stability, even with time-average payoffs, where there exist neutrally stable strategies of the infinitely repeated Prisoner's Dilemma that are arbitrarily close to "always defect" (Fudenberg and Maskin 1990) for example. Modifications of evolutionary/neutral stability would thus seem to be required for significant refinements of the Folk Theorem. One such modification is Fudenberg and Maskin's (1990) explicit incorporation of mistakes into neutral stability, with players weighting events lexicographically in decreasing order of the number of mistakes required for the event to occur. When players employ finitely complex strategies and have time-average payoffs of this lexicographic form, Fudenberg and Maskin demonstrate that stability implies efficiency in the infinitely repeated Prisoner's Dilemma. The essential idea is that, when players make mistakes, the worst possible history for an inefficient strategy profile will eventually occur; such a profile is then vulnerable to invasion by a mutant that mimics the incumbent strategy except after this worst history, at which point it engages in the familiar evolutionary "secret handshake" (Robson 1990), having nothing to lose from punishment. Binmore and Samuelson (1992) also employ the secret handshake to destabilize inefficient profiles, but their modification to neutral stability is the introduction of complexity costs in the players' strategies, so that the

off-the-equilibrium-path punishments required to prevent secret handshakes cannot themselves form part of a stable strategy profile.

But static notions of evolutionary stability are usually justified with reference to their dynamic stability properties (see, e.g., Weibull 1995). And whilst it seems likely that one could specify dynamic processes that would justify these modifications of evolutionary stability, to do so would highlight an (arguably) unappealing feature of the evolutionary approach to repeated games in general—namely, that the process would involve many repetitions of the *whole* repeated game. This presents obvious difficulties in the zero-discounting case of time-average payoffs, and whilst the hazard-rate interpretation of the discounting case guarantees that the repeated game will end in finite time (Binmore and Samuelson 1992), this still seems a little unsatisfactory. Moreover, the simple behavior (e.g. imitation, myopia) assumed in evolutionary models seems out of place in the high-rationality world of repeated games. The more natural setting for exploring equilibrium selection in repeated games would instead seem to be provided by models of learning.

## 3 Hypothesis Testing and the Folk Theorem

Let us begin by recalling the details of Foster and Young's (2003) hypothesis-testing model. There is a finite, $n$-person stage game $G$ with players $i = 1, 2, \ldots, n$, action spaces $X_i$ and utility functions $u_i : X \to \mathbb{R}$, $X = \Pi X_i$. This stage game is infinitely repeated in discrete time $t = 1, 2, \ldots$, with public observation of play. A *history of play* is denoted $\omega \in \Omega$; $\omega^t = (\omega_1^t, \ldots, \omega_n^t) \in X$ then denotes the actions taken in period $t$, $\overline{\omega}^t = (\omega^1, \omega^2, \ldots, \omega^t)$ the *initial history* of actions taken in periods 1 through $t$ inclusive, and $\Omega(\overline{\omega}^t) = \{\alpha \in \Omega \mid \overline{\alpha}^t = \overline{\omega}^t\}$ the set of all *continuations* of the initial history $\overline{\omega}^t$. Let $\overline{\Omega}(\overline{\omega}^t) = \{\overline{\alpha}^{t'} \mid t' \geq t,\ \overline{\alpha}^t = \overline{\omega}^t\}$ be the set of possible *continued initial histories* following $\overline{\omega}^{t-1}$.

Each player then has a forecast $p_i^t(x_{-i} \mid \overline{\omega}^{t-1}, b_i)$ of his opponents' one-step-ahead behaviors, conditional on every possible initial history, determined by his *model* $b_i$. Moreover, this model has *memory at most* $m$ in that the conditional distributions satisfy

$$p_i^t(x_{-i} \mid \overline{\omega}^{t-1}, b_i) = p_i^t(x_{-i} \mid \omega^{t-m}, \ldots, \omega^{t-1}, b_i) \quad \text{for all } t > m.$$

Since there are $M = |X|^m$ possible length-$m$ histories, models with memory at most $m$ occupy the Euclidean space $\mathcal{B}_i = \prod_{j \neq i} \Delta_j^M$. In response to his model, player $i$ adopts a *behavioral response* $a_i$ with *memory at most* $m$ in that the conditional probability that $i$ plays action $x_i$ in period $t$, given the history $\overline{\omega}^{t-1}$, is

$$q_i^t(x_i \mid \overline{\omega}^{t-1}, a_i) = q_i^t(x_i \mid \omega^{t-m}, \ldots, \omega^{t-1}, a_i) \quad \text{for all } t > m.$$

Note that $a_i \in \mathcal{A}_i = \Delta_i^M$, and $\mathcal{B}_i = \prod_{j \neq i} \mathcal{A}_j$. Letting $\mathcal{A} = \prod \mathcal{A}_i$ and $\vec{a} = (a_1, \ldots, a_n) \in \mathcal{A}$,

we can define the mapping $B_i : \mathcal{A} \to \mathcal{B}_i$ from any response vector $\vec{a}$ to the correct model for $i$, $B_i(a) = \prod_{j \neq i} a_j$.

With a discount factor $\rho_i < 1$ for player $i$, $i$'s normalized discounted utility following the initial history $\overline{\omega}^{t-1}$ is $U_i^t(\omega) = (1-\rho_i) \sum_{t'=t}^{\infty} \rho_i^{t'-t} u_i(\omega^{t'})$. Letting $\nu_{a_i,b_i}$ be the probability measure over infinite histories induced by the response $a_i$ and the model $b_i$, we can define $i$'s expected utility $U_i^t(a_i, b_i) \equiv \mathrm{E}\left(U_i^t(\omega) \mid a_i, b_i, \overline{\omega}^{t-1}\right)$ at time $t$ over all continuation histories $\Omega(\overline{\omega}^{t-1})$ as

$$\mathrm{E}\left(U_i^t(\omega) \mid a_i, b_i, \overline{\omega}^{t-1}\right) = \int_{\Omega(\overline{\omega}^{t-1})} U_i^t(\omega) d\nu_{a_i,b_i} \Big/ \int_{\Omega(\overline{\omega}^{t-1})} d\nu_{a_i,b_i}.$$

Given $\sigma_i > 0$, $a_i$ is then a *$\sigma_i$-optimal response to $b_i$* if $U_i^t(a_i, b_i) \geq U_i^t(a_i', b_i) - \sigma_i$, $\forall t, \forall a_i'$. For each player $i$ there is a $\sigma_i$-optimal response function $A_i^{\sigma_i} : \mathcal{B}_i \to \mathcal{A}_i$ that is assumed to be *continuous* in $b_i$ and each payoff $u_i(x)$, and *diffuse* in the sense that each action is played with positive probability. Such an $A_i^{\sigma_i}$ is called a *$\sigma_i$-smoothed best response function* and $\{A_i^{\sigma_i} : \sigma_i > 0\}$ is a *family* of smoothed best response functions. Let $\mathcal{S}^{\vec{\sigma}}$ be the set of fixed points of $A^{\vec{\sigma}} \circ B$, with $\mathcal{S}$ the set of subgame-perfect equilibrium response vectors obtained when $\sigma_i = 0$, $\forall i$.

Player $i$ periodically tests his *null hypothesis* that "the real process generating the actions from time $t$ on is described by the pair $(A_i^{\sigma_i}(b_i^t), b_i^t)$." If $i$ is not conducting a test at the start of a given period, he begins a new test with probability $1/s_i$. He then collects data on realized actions over the next $s_i$ periods, at the end of which he either accepts $H_0$ or rejects it. If it is rejected, he chooses a new model according to a probability measure $f_i^{t+s_i+1}(b_i \mid \overline{\omega}^{t+s_i})$. The hypothesis tests that the players employ are assumed to be "powerful," in that they accept the null with high probability when the null is correct, and reject with high probability when the null is not correct; see Foster and Young (2003) for the formal details of this concept.

The relevant formulation of the (perfect) Folk Theorem for this framework is Fudenberg and Maskin's (1986, 1991) discounting case, which also allows for unobservable mixed strategies.[2] The theorem says that, when players can observe each others' payoffs and actions, all feasible individually rational payoffs can be sustained in a subgame-perfect equilibrium of the infinitely repeated game if the discount factor is sufficiently close to one and a "full dimensionality" condition is satisfied. Foster and Young's (2003) Theorem 2 then says that, if players—lacking knowledge of their opponents' payoffs—engage in hypothesis testing about opponents' strategies, then there exist parameters of this process such that the strategies are $\varepsilon$-close to being a subgame-perfect equilibrium of the infinitely repeated game at least $1 - \varepsilon$ of the time.

But we need not stop here; not all subgame-perfect equilibria are equally appealing under hypothesis testing. Rather than settling on a particular equilibrium, for given $\varepsilon > 0$ the process will perpetually bounce around between equilibria given long enough. But it will spend more time in some equilibria than in others, according to how likely they are to be entered and exited.

[2] Also relevant will be Aumann and Shapley's (1976) time-average case, where players are arbitrarily patient.

The crucial variables in this respect are, for our purposes, the probabilities $f_i(\cdot|\overline{\omega}^t)$ with which new models are adopted upon hypothesis rejection.

# 4 Conservatism and Forgiveness

The main assumption made on the hypothesis-revision densities $f_i(\cdot|\overline{\omega}^t)$ by Foster and Young (2003) is *flexibility*, whereby, for each $\tau_0 > 0$, the $f_i$-measure of any $\tau_0$-ball of hypotheses is bounded below by a strictly positive number $f_*(\tau_0) > 0$. This is quite a weak assumption, allowing a wide range of possible models to be adopted at any given revision. In particular, absent further assumptions, it need not be more difficult for a player to adopt a model further away in the Euclidean model space $\mathcal{B}_i = \prod_{j \neq i} \Delta_j^M$, as would be the case in most traditional evolutionary analyses. This role for the "distance" between models is instead captured by Foster and Young's "conservatism," under which the new hypothesis lies within $\lambda_i$ of the old hypothesis with probability at least $1 - \lambda_i$, where $\lambda_i$ is positive and close to zero. Foster and Young do not use this assumption in their main convergence results; rather, they show that if players are sufficiently conservative, have sufficiently sharp best responses and employ sufficiently powerful hypothesis tests with sufficiently fine tolerances, then at all times the hypothesis testing strategies are $\epsilon$-best responses to their *beliefs* (as distinct from models).

We will modify conservatism slightly, in order to allow us to strengthen the concept somewhat.

**Definition 1** *Model revision is* conservative *if, following rejection, player $i$ adopts a new model that is within $\lambda_i$ of his previous model with probability at least $1 - \Lambda_i$.*

Conservatism thus still captures the idea that local model revisions are highly probable, but just how probable is no longer tied to the size of the neighborhood; this will be key in proving the main result.

We will also depart from the Foster and Young framework in our notion of smoothed best response. First, we will begin with a *static* notion of $\sigma_i$-optimality which requires only that $U_i^t(a_i, b_i) \geq U_i^t(a_i', b_i) - \sigma_i$, $\forall a_i'$, in the period $t$ when $a_i$ is selected, before moving on later to Foster and Young's full extensive-form $\sigma_i$-optimality. Second, we will allow player $i$ to adopt any $\sigma_i$-smoothed best response function following the adoption of a new model, rather than always employing the same $A_i^{\sigma_i}$. Once they have chosen such a response though, they must still retain it until the next time they reject their model. This modification of the Foster and Young framework is necessary to allow the experimentation with efficiency exploited in the proof of the main result. Individual response vectors are, however, doomed to instability even under this assumption, given the presence of many alternative $\sigma_i$-optimal strategies following adoption even of a model arbitrarily close to the rejected one. Hence, we consider the stability not of individual response vectors, but of a class of response vectors sharing certain properties.

### 4.1 Example: Prisoner's Dilemma

Consider two players playing the infinitely repeated Prisoner's Dilemma according to (slightly noisy) "trigger" strategies: cooperate if and only if your opponent has always cooperated in the past. Suppose that one of the players has made a mistake, so that the players are locked in to perpetual defection. And suppose that one of the players now tests and then rejects his current (correct) hypothesis in favor of a local alternative placing small probability on his opponent playing "perfect tit-for-tat"—whereby a player cooperates in the first period and thereafter cooperates if and only if either both players cooperated or both players defected in the previous period—starting in some particular future period $t'$. Then, if that player is sufficiently patient, it is a smoothed best response (according to our static notion of optimality) for him also to play perfect-tit-for-tat starting in $t'$, and continuing as long as there have been no more than $l$ deviations from perfect tit-for-tat since $t'$, where $l$ is a positive integer. To see this, note that it is optimal to play perfect tit-for-tat if it turns out that the opponent is playing perfect tit-for-tat, and if not, reversion to the trigger strategy after $l$ deviations yields only a small loss if the player is sufficiently patient. Crucially, things can be no worse for the player upon reversion to the trigger strategy, since they were already locked in to the worst possible scenario of perpetual defection.

Suppose now that the opponent too rejects his null hypothesis in favor of the same local alternative. Then it is a smoothed best response for him to experiment with perfect tit-for-tat starting in $t'$ in the same way. Cooperation will thus begin in period $t'$, and continue for some time if mistakes are unlikely. During this cooperative phase, if a player again rejects his hypothesis, and adopts a local alternative with somewhat more probability on his opponent playing perfect tit-for-tat, then it is a smoothed best response for that player to continue playing perfect tit-for-tat as long as there have been no more than $l'$ deviations from perfect tit-for-tat since $t'$, where $l' > l$; more mistakes are forgiven, since it is now less likely that the opponent is playing the trigger strategy. Indeed, further model rejections and revisions may occur before reversion to trigger strategies takes place; and ultimately, enough model revisions may occur for the system to arrive in a state where both players are almost certain that their opponent is playing perfect tit-for-tat, a smoothed best response to which is to also play perfect tit-for-tat.

Once each player correctly believes perfect tit-for-tat is being played by his opponent, no local model revision can destabilize it. For there is no initial history such that perfect tit-for-tat is strongly Pareto-dominated by some other response vector; it is a response vector that *forgives mistakes*. Whilst a player can still place small probability on his opponent experimenting with a response that would leave the player better off (e.g. the "always cooperate" strategy, against which defection goes unpunished), the opponent has no incentive to adopt such a response, and the experimentation is doomed to break down, resulting in reversion to perfect tit-for-tat or an equivalent forgiving strategy. If such forgiving strategies are to be upset then, we require a (less probable) nonlocal model revision, or a sequence of (improbable) model rejections and revisions.

Hence, the set of forgiving response vectors is easy to enter and difficult to leave, so that the system will spend a lot of time there (or thereabouts).

### 4.2 General Case

This logic applies quite generally to favor response vectors that share the forgiving property of perfect tit-for-tat.

**Definition 2** $(\vec{a}, \overline{\omega}^{t-1})$ *is* weakly $\theta$-efficient *if* $(\mathrm{E}(U_i^t(\omega)|a_i, B_i(\vec{a}), \overline{\omega}^{t-1}))_{i=1,2}$ *is weakly Pareto-*$\vec{\theta}$*-undominated in the set of equilibrium response vectors.*

*The* worst-case scenario *for player* $i$ *under* $(\vec{a}, \overline{\omega}^{t-1})$ *is the initial history* $\overline{\omega}_i^{t''-1} = \arg\min_{\overline{\omega}^{t'-1} \in \overline{\Omega}^{s_i}(\overline{\omega}^{t-1})} \mathrm{E}(U_i^{t'}(\omega)|a_i, B_i(\vec{a}), \overline{\omega}^{t'-1})$.

$(\vec{a}, \overline{\omega}^{t-1})$ *is* $\theta$-forgiving *if, for any worst-case scenario* $\overline{\omega}_i^{t''-1}$, $(\vec{a}, \overline{\omega}_i^{t''-1})$ *is either weakly* $\theta$*-efficient or has probability at most* $\theta$ *following* $\overline{\omega}^{t-1}$.

Thus, if $(\vec{a}, \overline{\omega}^{t-1})$ is $\theta$-forgiving, then any worst-case scenario such that some equilibrium response vector strongly Pareto-$\vec{\theta}$-dominates $(U_i^{t'}(a_i, B_i(\vec{a})))_{i=1,2}$ is reached with probability at most $\theta$. If $\theta = 0$, then we call $(\vec{a}, \overline{\omega}^{t-1})$ simply *efficient* or *forgiving*. Intuitively, $(\vec{a}, \overline{\omega}^{t-1})$ is forgiving if any finite number of mistakes brings no reduction in expected payoffs for at least one player. In the repeated Prisoner's Dilemma, perfect tit-for-tat following mutual cooperation (or defection) is forgiving in this sense. This is slightly different (though related) to Axelrod's (1984, p. 36) informal notion of forgiveness in the Prisoner's Dilemma as the "propensity to cooperate in the moves after the other player has defected."

Let us begin with a result for our static notion of $\sigma_i$-optimality, under which experimentation with forgiveness is easier to foster.

**Theorem 1 (Static)** *Suppose that two sufficiently patient players adopt hypotheses with finite memory, have* $\sigma_i$*-smoothed static best response functions, employ powerful hypothesis tests with comparable amounts of data, and are flexible and conservative in the adoption of new hypotheses. Given any* $\epsilon > 0$*, if the* $\sigma_i$ *are small (given* $\epsilon$*), if the test tolerances* $\tau_i$ *are sufficiently fine (given* $\epsilon$ *and* $\sigma_i$*), if the amounts* $s_i$ *of data collected are sufficiently large (given* $\epsilon$*,* $\sigma_i$ *and* $\tau$*) and if the degrees* $1 - \Lambda_i$ *of conservatism are sufficiently high, then there exists* $\theta > 0$ *such that the repeated-game strategies are within* $\epsilon$ *of* $\theta$*-forgiving* $(\vec{a}, \overline{\omega}^{t-1})$ *at least* $1 - \epsilon$ *of the time.*

A state is within $\epsilon$ of $(\vec{a}, \overline{\omega}^{t-1})$ if its initial history is $\overline{\omega}^{t-1}$ and its response vector is within $\epsilon$ of $\vec{a}$. The proof is relegated to the appendix.

The intuition behind the result is that, if $(\overline{\omega}^{t-1}, \vec{a})$ is not forgiving, then it is vulnerable to experimentation initiated following the worst mistakes that will go unforgiven, since the experimenter has nothing to lose and both players have something to gain. This is similar to the reasoning employed by Fudenberg and Maskin (1990), except that a one-off secret handshake followed by permanent efficiency or reversion can no longer work, since the continued possibility

of mistakes makes the secret handshake's signal imprecise.[3] Instead, the players must first experiment a little, then gradually tolerate more and more mistakes as the experimentation is reciprocated, until they finally become completely efficient and forgiving.

On the other hand, if $(\overline{\omega}^{t-1}, \vec{a})$ is forgiving, then experimentation cannot lead to a payoff profile strictly preferred by both players, and hence it will fail. Unlike in the symmetrized setting of Fudenberg and Maskin (1990), however, such failed experimentation can occur, and when it does the experimenter may revert not just to his previous response $a_i$, but to any alternative smoothed best response to his model. Hence, any given response vector cannot possibly be stable; rather, because a smoothed best response to a forgiving model must itself be $\theta'$-forgiving for some $\theta' > 0$, there exists a $\theta > \theta'$ such that the set of $\theta$-forgiving response vectors is stable for some long period of time. These observations, along with the probabilistic techniques of Foster and Young (2003), are exploited to give the result on the long-run behavior of the dynamical system.

The idea of experimentation that is doomed to failure might seem a little odd; why should a player adopt a model (of reciprocated experimentation) under which his opponent fails to act in his own interests in such an elaborate manner? Should player $i$ not place low probability on his opponent playing a response that is $\sigma_j$-optimal, $j \neq i$, under no possible models? The answer is that such reasoning can have no place in a learning model such as Foster and Young's, since it would require players to have knowledge of opponents' payoffs and would thus sacrifice the "uncoupled" nature of the process.[4]

One deficiency of Theorem 1 is the static notion of $\sigma_i$-optimality employed. This facilitates experimentation, since it is not clear that Fudenberg and Maskin's worst-case scenario will endure long enough to sustain extended experimentation as $\sigma_i$-optimal in all subgames. Demonstrating that it will in fact do so if players are sufficiently patient is the key to the following result, which returns to Foster and Young's full extensive-form notion of $\sigma_i$-optimality.

**Theorem 2 (Extensive-form)** *Suppose that two sufficiently patient players now have $\sigma_i$-smoothed extensive-form best response functions, but are otherwise as before. Given any $\epsilon > 0$, there again exist values of the learning parameters and $\theta > 0$ such that the repeated-game strategies are within $\epsilon$ of $\theta$-forgiving $(\vec{a}, \overline{\omega}^{t-1})$ at least $1 - \epsilon$ of the time.*

The proof is again relegated to the appendix.

The downside with this extensive-form $\sigma_i$-optimality result is that it is likely to require a much higher discount factor than Theorem 1; experimentation must now remain $\sigma_i$-optimal in every subgame where it continues, so that less departure from time-average payoffs and their disregard for initial histories can be tolerated. Whether or not the $\theta$ parameter is tighter in Theorem 2 is not so clear; whilst the set of extensive-form $\sigma_i$-optimal responses to a $\theta'$-forgiving

[3] Moreover, the noise that is explicit in the Foster and Young set-up allows an extension of Fudenberg and Maskin's reasoning from time-average payoffs to the discounting case.

[4] On uncoupled learning processes, see Hart and Mas-Colell (2003, 2004), and Foster and Young (2005).

model is narrower than the set of static $\sigma_i$-optimal responses, we are constrained to reaching only a $\theta'$-forgiving great state initially, rather than a forgiving great state.

# 5 Conclusion

If patient players learn according to Foster and Young's (2003) hypothesis testing then, and are sufficiently conservative in their adoption of new hypotheses, almost all time is spent approximating an efficient set of "strategies" that have an intuitive forgiving property. For example, in the Prisoner's Dilemma, almost all time is spent close to the Pareto frontier of the shaded region of Figure 1, enforced by equilibrium strategies such as perfect tit-for-tat that forgive finite numbers of mistakes. Intuitively, strategies that do not forgive mistakes are vulnerable to experimentation with efficiency once mistakes have been made. And whilst any *given* forgiving strategies are unstable in the face of alternative best replies, the *set* of forgiving strategies is stable; a change in long-run behavior requires both players to "agree" on their new behavior, which they cannot do if they are already playing efficiently.

The noise inherent in the hypothesis-testing process thus provides a tool for selecting among the myriad possibilities of the Folk Theorem, and under conservatism it gives support for the notion that efficiency is likely to emerge in repeated games. The hypothesis-testing model has precisely the elements required of an evolutionary refinement of the Folk Theorem: a tractable metric space for (approximations of) strategies and beliefs to occupy; endogenous "mistakes" occurring with small probability; and a technology for rejection and revision of "beliefs." However, it is likely that our results would extend to other stochastic dynamic models of evolution and learning in repeated games that share the key features of noisy best response and conservatism.

# Appendix

**Proof of Theorem 1.** As in Foster and Young (2003),

$$\forall i, \quad \sigma_i \leq \frac{\epsilon}{2}$$

and, for any given $A_i^{\sigma_i}$,

$$\exists \delta > 0, \quad \forall \vec{u}, t, i, \quad \forall b_i, b_i', \quad \left|b_i - b_i'\right| \leq \delta \;\Rightarrow\; \left|A_i^{\sigma_i}(b_i) - A_i^{\sigma_i}(b_i')\right| \leq \frac{\epsilon}{4}, \quad \text{and} \quad \delta < \frac{\epsilon}{4}. \tag{1}$$

Furthermore, fix $\tau > 0$ such that

$$\tau \leq \frac{\delta}{6};$$

for each $\tau$ there exist functions $k(\tau)$ and $r(\tau)$ such that whenever a player's model is within $c(\tau)$ of the correct model, he rejects with probability at most

$$k(\tau)\mathrm{e}^{-r(\tau)s_*}.$$

In the proof of their Claim 2, Foster and Young demonstrate that there is a $\gamma > 0$ such that, for a given model fixed point $\vec{b}^f = B(\vec{a}^f)$,

$$\forall i, \quad \left|b_i - b_i^f\right| < \gamma \;\Rightarrow\; \left|b_i - B_i\left(A^{\vec{\sigma}}(\vec{b})\right)\right| < c(\tau).$$

Finally, recall Foster and Young's notion of a *great state*—where, for every player $i$, $i$'s model is within $\gamma$ of a fixed point and no player is currently in a test phase.

Beginning at a $\theta$-unforgiving $\vec{a}$ at time $t$ for given $\theta > 0$, there exists a worst-case scenario $\overline{\omega}_1^{t''-1} = \arg\min_{\overline{\omega}^{t'-1} \in \overline{\Omega}^{s_1}(\overline{\omega}^{t-1})} \mathrm{E}(U_1^{t'}(\omega)|a_1, B_1(\vec{a}), \overline{\omega}^{t'-1})$, with probability at least $\theta$, such that $(\mathrm{E}(U_i^{t''}(\omega)|a_i, B_i(\vec{a}), \overline{\omega}^{t''-1}))_{i=1,2}$ is strongly Pareto-$\vec{\theta}$-dominated by some forgiving $\vec{a}^* \in \mathcal{S}$. Letting $\overline{\omega}_2^{t'''-1} := \arg\min_{\overline{\omega}^{t'-1} \in \overline{\Omega}(\overline{\omega}^{t''-1})} \mathrm{E}(U_2^{t'}(\omega)|a_2, B_2(\vec{a}), \overline{\omega}^{t'-1})$, it follows that $\mathrm{E}(U_2^{t'''}(\omega)|a_2, B_2(\vec{a}), \overline{\omega}^{t'''-1}) \leq \mathrm{E}(U_2^{t''}(\omega)|a_2, B_2(\vec{a}), \overline{\omega}^{t''-1})$. Let the response $a_1^{l_1}$ continue to play $a_1$ unless $\overline{\omega}_2^{t'''-1}$ is realized, in which case it plays $a_1^*$ if and only if there have been at most $l_1$ deviations from $\vec{a}^*$ (by either player) since $t'''$; otherwise, it reverts to $a_1$ forever. This response clearly does not have memory $m$—and in fact requires infinite memory over the whole game—but if it is $\sigma_1'$-optimal, $\sigma_1' < \sigma_1$, given some memory-$m$ model, then we can always choose $m$ sufficiently large such that there exists a $\sigma_1$-optimal memory-$m$ response constructed from $a_1^{l_1}$ using the procedure described in Foster and Young (2003, p. 81). Similarly, let $a_2^{l_2}$ respond to $\overline{\omega}_2^{t'''-1}$ by playing $a_2^*$ if and only if there have been at most $l_2$ deviations from $\vec{a}^*$ since $t'''$, otherwise reverting to $a_2$ forever. Finally, let $a_i^\infty := \lim_{l_i \to \infty} a_i^{l_i}$ be the response that plays $a_i^*$ for all $\overline{\omega}^{t'-1} \in \overline{\Omega}(\overline{\omega}_1^{t''-1})$, and $b_i' = (1 - \lambda_i + \iota_i)b_i + (\lambda_i - \iota_i)a_j^\infty$, $\iota_i < \min\{\delta, \lambda_i/2\}$, $j \neq i$. Consider the following sequence of events leading to a great state with a response vector within $\min\{\gamma, \delta\}$ of $\vec{a}^\infty$.

**Step 1.** Play proceeds in accordance with $\overline{\omega}_1^{t''-1}$. Player 2 does not start a test between periods $(t'' - \max\{s_1, 2s_2\})$ and $(t'' - 1)$; player 1 does not start a test between periods $(t'' - 2s_1)$ and $(t'' - (s_1 + 1))$, but does so in period $(t'' - s_1)$. After 1's test phase is completed, he rejects his current hypothesis and adopts a model within $\iota_1$ of $b_1'$. Now, fixing $\sigma_1' \in ((1 - \lambda_1 + \iota_1)\sigma_1, \sigma_1)$, whilst $\mathrm{E}(U_1^{t'}(\omega)|a_1^{l_1}, b_1, \overline{\omega}_1^{t'-1})$ is nonincreasing in $l_1$ for given $\rho_1 < 1$ and any $\overline{\omega}^{t'-1} \in \overline{\Omega}(\overline{\omega}^{t''-1})$, if $\rho_1$ is sufficiently high there exists a maximum $\bar{l}_1 > 0$ such that $a_1^{\bar{l}_1}$ is a $\sigma_1^+$-optimal response to $b_1$ for all $l_1 \leq \bar{l}_1$ and $\sigma_1^+ := \sigma_1'/(1 - \lambda_1 + \iota_1) > \sigma_1$. To see this, note that, by eventually reverting to $a_1$, $a_1^{l_1}$ gives the same time-average payoff as $a_1$ against $b_1$ following $\overline{\omega}_1^{t''-1}$. And since $a_1$ is a $\sigma_1$-optimal response to $b_1$, it follows that $a_1^{l_1}$ must be a $\sigma_1^+$-optimal response to $b_1$ under discounting given $l_1$ sufficiently low and $\rho_1$ sufficiently high. Moreover, since $a_1^*$ is

an optimal response to $a_2^*$, mistakes have zero probability under $\vec{a}^*$ and hence $a_1^{\bar{l}_1}$ must be an optimal response to $a_2^\infty$. Hence,

$$\begin{aligned}
\mathrm{E}(U_1^{t''}(\omega)|a_1^{\bar{l}_1}, b_1', \overline{\omega}_1^{t''-1}) &= (1-\lambda_1+\iota_1)\,\mathrm{E}(U_1^{t''}(\omega)|a_1^{\bar{l}_1}, b_1, \overline{\omega}_1^{t''-1}) \\
&\qquad + (\lambda_1-\iota_1)\,\mathrm{E}(U_1^{t''}(\omega)|a_1^{\bar{l}_1}, a_2^\infty, \overline{\omega}_1^{t''-1}), \\
&\geq (1-\lambda_1+\iota_1)(\sup_{a_1'}\mathrm{E}(U_1^{t''}(\omega)|a_1', b_1, \overline{\omega}_1^{t''-1}) - \sigma_1^+) \\
&\qquad + (\lambda_1-\iota_1)\sup_{a_1'}\mathrm{E}(U_1^{t''}(\omega)|a_1', a_2^\infty, \overline{\omega}_1^{t''-1}), \\
&\geq \sup_{a_1'}\mathrm{E}(U_1^{t''}(\omega)|a_1', b_1', \overline{\omega}_1^{t''-1}) - (1-\lambda_1+\iota_1)\sigma_1^+ \\
&= \sup_{a_1'}\mathrm{E}(U_1^{t''}(\omega)|a_1', b_1', \overline{\omega}_1^{t''-1}) - \sigma_1',
\end{aligned}$$

so that $a_1^{\bar{l}_1}$ is a $\sigma_1'$-optimal response to $b_1'$. (1) then implies that player 1 has a $\sigma_1$-optimal response to his new model within $\epsilon/4$ of (a $\sigma_1$-optimal memory-$m$ response to $b_1'$ appropriately constructed from) $a_1^{\bar{l}_1}$. (Duration: $(t''-t)$ periods.)

**Step 2.** Play proceeds in accordance with $\overline{\omega}_2^{t'''-1}$. Prior to period $(t'''-s_2)$, player 2 starts a test period, at the end of which he rejects and adopts a model within $\iota_2 < \delta, \lambda_2/2$ of $b_2'$. By the argument in Step 1, there is then a $\sigma_2$-optimal response to $b_2'$ within $\epsilon/4$ of (a $\sigma_2$-optimal memory-$m$ response to $b_2'$ appropriately constructed from) $a_2^{\bar{l}_2}$. (Duration: $(t'''-t'')$ periods.)

**Step 3.** Each player $i$ conducts successive non-overlapping tests, rejecting at the end of each and adopting a $\lambda_i$-close model within $\iota_i$ of a linear combination of $b_i$ and $a_j^\infty$, $j \neq i$, that maximizes the weight on $a_j^\infty$ subject to its $\iota_i$-ball being contained in the rejected model's $\lambda_i$-ball, until he adopts a model within $\min\{\gamma, \delta\}$ of $a_j^\infty$. Throughout this process, there are no deviations from $\vec{a}^*$. (1) then implies the existence of a $\sigma_i$-optimal memory-$m$ response within $\epsilon/4$ of (a $\sigma_i$-optimal memory-$m$ response to $a_j^\infty$ appropriately constructed from) $a_i^\infty$. (Duration: at most $\varrho = \lceil 2(1-\min\{\gamma,\delta\}-\lambda_*+2\iota^*)/(\lambda_*-2\iota^*)\rceil s^*$ periods, where $\lambda_* := \min_i \lambda_i$, $\iota^* := \max_i \iota_i$ and $s^* := \max_i s_i$.)

**Step 4.** If the number of periods in Steps 1–3 is $T' < t'''-t+\varrho$, no player begins a test for the next $t'''-t+\varrho-T'$ periods.

The duration of the whole sequence is exactly $t'''-t+\varrho$.

We now calculate the probability of a particular such sequence. To begin with, $\overline{\omega}_1^{t''-1}$ must be realized, which occurs with probability $h_{\vec{a}}(\overline{\omega}_1^{t''-1} \mid \overline{\omega}^{t-1})$ say. There must be no rejections between $t$ and $t''-1$, which occurs with probability at least $(1-1/s_*)^{2(t''-s_*-t)}$, where $s_* := \min_i s_i$. Player 1's first test phase must then begin in period $(t''-s_1)$, and player 2 must not test

during this phase, which occurs with probability at least $(1/s^*)(1-1/s_*)^{s^*}$. Player 1 must then reject his null hypothesis, and adopt a new one within a target of radius $\iota_1$ in his model space, which occurs with probability at least $(1-\nu^*)f_*$, where $f_* = f_*(\iota_*)$ and $\nu^*$ is the maximum probability of a player accepting his null hypothesis.

Next, $\overline{\omega}_2^{t'''-1}$ must be realized, which occurs with probability $h_{\vec{a}}(\overline{\omega}_2^{t'''-1} \mid \overline{\omega}_1^{t''-1})$. There must be no rejections between $t''$ and $t'''-1$, followed by a player-2 test phase starting at $(t'''-s_2)$ and with no simultaneous testing by player 1, which occurs with probability at least $(1/s^*)(1-1/s_*)^{2(t'''-1-s_*-t'')+s^*}$. Player 2 must then reject his null hypothesis, and adopt a new one within a target of radius $\iota_2$ in his model space; this event has probability at least $(1-\nu^*)f_*$.

The Step 3 test phases must each begin at a specific time and no player can be testing during the other's phase; the probability of this is at least $\left((1/s^*)(1-1/s_*)^{s^*}\right)^{\varrho/s^*}$. Each of these tests must end with rejection and subsequent adoption within a model-space target of radius $\iota_i$; we can choose the test parameters such that the rejections occur with probability at least $1/2$, so that the event has probability at least $(f_*/2)^{\varrho/s^*}$. There must be no deviations from $\vec{a}^*$ from $\overline{\omega}_2^{t'''-1}$ to the end of Step 4; let the probability of this event be $h_{\vec{a}}(\omega^* \mid \overline{\omega}_2^{t'''-1})$, and let $H = h_{\vec{a}}(\omega^* \mid \overline{\omega}_2^{t'''-1}).h_{\vec{a}}(\overline{\omega}_2^{t'''-1} \mid \overline{\omega}_1^{t''-1}).h_{\vec{a}}(\overline{\omega}_1^{t''-1} \mid \overline{\omega}^{t-1})$.

Finally, there must be no further tests before period $t'''+\varrho$; this occurs with probability at least $(1-1/s_*)^{2(t''-t+\varrho)}$.

In summary, the probability of Steps 1–4 is at least

$$H.(1-\nu^*)^2(1-1/s_*)^{3\varrho+2(t'''+t''-1-2t+s^*-2s_*)}(f_*/2s^*)^{2+\varrho/s^*}.$$

Thus there are constants $\alpha, \beta \in (0,1)$ such that the probability of Steps 1–4 is at least $\alpha\beta^{\varrho}$, establishing the following fact: *If $(\vec{a}, \overline{\omega}^{t-1})$ is $\theta$-unforgiving, the probability of being in a forgiving great state at time $t'''+\varrho$ is at least $\alpha\beta^{\varrho}$.*

Now suppose that the process is in a forgiving great state at time $t$. Letting $T$ be a large positive integer, the probability that any player rejects a test over the next $T$ periods is bounded above by

$$\lceil 2T/s_* \rceil k_0 \mathrm{e}^{-r_0 s_*} < T\mathrm{e}^{-4rs_*},$$

where the inequality holds for all sufficiently large $s_*$ and some $r > 0$. The probability, conditional on rejection, that player $i$ will adopt a new model $b_i''$ such that $A^{\vec{\sigma}}(b_i'', b_j)$ gives a $\theta'$-unforgiving state is bounded above by $\Lambda$, for some $\theta' \geq 0$. To see this, note that otherwise there would have to be a player $i$ and a model $b_i' = (1-\zeta)b_i + \zeta b_i''$, $\zeta \leq \lambda_i$, $b_i'' \in \mathcal{B}_i$, that induced a $\sigma_i$-optimal $\theta'$-unforgiving response $a_i'$; choose $\theta'$ such that no such model exists. But $\theta'$ must be sufficiently low to then avoid a worst-case-scenario escape of the sort seen above; choose $\vec{\sigma}$ sufficiently low that mistakes—and thus a worst-case scenario escape—are sufficiently unlikely. There is then a constant $\eta \in (0,1)$ and a $\theta > \theta'$—which fixes the level of $\theta$ at the beginning of the proof—such that the following fact holds: *If $(\vec{a}, \overline{\omega}^{t-1})$ is a forgiving great state,*

*the probability of being in a $\theta$-unforgiving state within $T$ periods is at most $\eta\Lambda T$.*[5] Moreover, $\theta'$—and thus $\theta$—vanishes with $\vec{\sigma}$.

We can now use the above bounds to show that the fraction of times that the process is not in a $\theta$-forgiving state is very small for small $\lambda_*$. Starting from time $t$, let $\mathcal{E}$ be the event "the realized states in at least $\varepsilon T$ of the periods $t+1,\ldots,t+T$ are $\theta$-unforgiving." Let $\mathcal{E}'$ be the sub-event of $\mathcal{E}$ in which no forgiving great state is realized before the last $\theta$-unforgiving state, and let $\mathcal{E}'' = \mathcal{E} - \mathcal{E}'$. We shall bound the conditional probabilities of $\mathcal{E}'$ and $\mathcal{E}''$ from above independently of the state at time $t$.

Let $t^{\max}$ be the maximum over all $\theta$-unforgiving $\vec{a}$ and all $t' \geq t$ of the time interval between $t'$ and the realization of both players' worst remaining initial histories. If $\mathcal{E}'$ occurs, there are at least $\lfloor \epsilon T/(t^{\max}+\varrho)\rfloor = k$ distinct times $t < t_1 < \cdots < t_k \leq t+T$ such that the following hold:

- $t_{j+1} - t_j \geq t^{\max} + \varrho$ for $1 \leq j < k$,
- the state at time $t_j$ is $\theta$-unforgiving for $1 \leq j < k$,
- no forgiving great state occurs from $t_1$ to $t_k$.

By the preceding, the probability of this event is at most $(1-\alpha\beta^{\varrho})^{k-1} \leq \mathrm{e}^{-\alpha\beta^{\varrho}(k-1)}$. Letting $T = (t^{\max}+\varrho)(1+\beta^{-2\varrho})/\epsilon$, we have

$$P(\mathcal{E}') \leq \exp(-\alpha\beta^{\varrho}(\lfloor \epsilon T/(t^{\max}+\varrho)\rfloor - 1)) = \exp(-\alpha\beta^{-\varrho}),$$

where $\lfloor x\rfloor := \max\{z \in \mathbb{Z} \mid z \leq x\}$. This can be made as small as we wish when $\lambda_*$ is small; in particular it can be made less than $\epsilon/2$.

If $\mathcal{E}''$ occurs, the process does *not* stay in $\theta$-forgiving states for at least $T$ periods after entering a forgiving great state. So from above, and letting $\Lambda = \epsilon\beta^{3\varrho}/(t^{\max}+\varrho)$,

$$P(\mathcal{E}'') \leq \eta\Lambda T = \eta\Lambda(t^{\max}+\varrho)(1+\beta^{-2\varrho})/\epsilon = \eta(\beta^{3\varrho}+\beta^{\varrho}),$$

which can also be made less than $\epsilon/2$ when $\lambda_*$ is sufficiently small. Putting all of this together we conclude that, for all sufficiently small $\lambda_*$,

$$P(\mathcal{E}) = P(\mathcal{E}') + P(\mathcal{E}'') \leq \epsilon.$$

Now divide all times $t$ into disjoint blocks of length $T$, and let $Z_k$ be the fraction of $\theta$-unforgiving times in the $k$th block. We have just shown that $P(Z_k \geq \epsilon) \leq \epsilon$ for all $k$. Hence

$$E(Z_k) \leq P(Z_k \geq \epsilon)\cdot 1 + P(Z_k < \epsilon)\cdot\epsilon \leq 2\epsilon.$$

[5] A sequence of improbable rejections followed by local model revisions—an alternative means of escape from a forgiving great state—becomes arbitrarily unlikely as $\lambda_*$ becomes small.

It follows that the proportion of times that the process is in a $\theta$-unforgiving state is almost surely less than $2\epsilon$. Rerunning the entire argument with $\epsilon/2$ yields the desired conclusion, namely that $(\vec{a}, \overline{\omega}^{t-1})$ is $\theta$-forgiving at least $1-\epsilon$ of the time. ∎

**Proof of Theorem 2.** The proof proceeds in the same way as that of Theorem 1, except that $b_i' = (1-\lambda_i+\iota_i)b_i+(\lambda_i-\iota_i)a_j^l$, $\iota_i < \min\{\delta, \lambda_i/2\}$, $j \neq i$, for some $l \in \mathbb{N}$ such that $a_i^l$ is an extensive-form $\sigma_i$-optimal response to $b_i'$. To see that such an $l$ exists, suppose that there is some $\sigma_1^- < \sigma_1$ such that $a_1^l$ is a $\sigma_1^-$-optimal response to $a_2^l$, and fix $\sigma_1' \in ((1-\lambda_1+\iota_1)\sigma_1 + (\lambda_1-\iota_1)\sigma_1^-, \sigma_1)$.[6] Whilst $\mathrm{E}(U_1^{t'}(\omega)|a_1^{l_1}, b_1, \overline{\omega}_1^{t'-1})$ is nonincreasing in $l_1$ for given $\rho_1 < 1$ and any $\overline{\omega}^{t'-1} \in \overline{\Omega}(\overline{\omega}^{t''-1})$, if $\rho_1$ is sufficiently high there exists a maximum $\bar{l}_1 > 0$ such that $a_1^{\bar{l}_1}$ is a $\sigma_1^+$-optimal response to $b_1$ for all $l_1 \leq \bar{l}_1$ and $\sigma_1^+ := (\sigma_1' - (\lambda_1-\iota_1)\sigma^-)/(1-\lambda_1+\iota_1) > \sigma_1$. To see this, note that, under time-average payoffs, for any given hypothesis $(a_i, b_i)$, $\mathrm{E}(U_i^{t'}(\omega)|a_i, b_i, \overline{\omega}^{t'-1})$ is the same for all $\overline{\omega}^{t'-1}$; otherwise, there would exist a worst length-$m$ history and a strictly preferred continuation, a contradiction. Hence, by eventually reverting to $a_1$, $a_1^{l_1}$ gives the same time-average payoff as $a_1$ against $b_1$ following $\overline{\omega}^{t''-1}$, and indeed following any $\overline{\omega}^{t'-1} \in \overline{\Omega}(\overline{\omega}^{t''-1})$. And since $a_1$ is a $\sigma_1$-optimal response to $b_1$, it follows that $a_1^{l_1}$ must be a $\sigma_1^+$-optimal response to $b_1$ under discounting given $l_1$ sufficiently low and $\rho_1$ sufficiently high. Hence, for $l \leq \bar{l}_1$,

$$\begin{aligned}
\mathrm{E}(U_1^{t''}(\omega)|a_1^l, b_1^l, \overline{\omega}_1^{t''-1}) &= (1-\lambda_1+\iota_1)\,\mathrm{E}(U_1^{t''}(\omega)|a_1^l, b_1, \overline{\omega}_1^{t''-1}) \\
&\qquad + (\lambda_1-\iota_1)\,\mathrm{E}(U_1^{t''}(\omega)|a_1^l, a_2^l, \overline{\omega}_1^{t''-1}), \\
&\geq (1-\lambda_1+\iota_1)(\sup_{a_1}\mathrm{E}(U_1^{t''}(\omega)|a_1, b_1, \overline{\omega}_1^{t''-1}) - \sigma_1^+) \\
&\qquad + (\lambda_1-\iota_1)(\sup_{a_1}\mathrm{E}(U_1^{t''}(\omega)|a_1, a_2^l, \overline{\omega}_1^{t''-1}) - \sigma_1^-), \\
&\geq \sup_{a_1}\mathrm{E}(U_1^{t''}(\omega)|a_1, b_1^l, \overline{\omega}_1^{t''-1}) - \big((1-\lambda_1+\iota_1)\sigma_1^+ + (\lambda_1-\iota_1)\sigma_1^-\big) \\
&= \sup_{a_1}\mathrm{E}(U_1^{t''}(\omega)|a_1, b_1^l, \overline{\omega}_1^{t''-1}) - \sigma_1',
\end{aligned}$$

so that $a_1^l$ is a $\sigma_1'$-optimal response to $b_1^l$. A similar sequence of events to that in Theorem 1 then leads to a great state with a response vector within $\min\{\gamma, \delta\}$ of $\vec{a}^l$. Now choose $\vec{\sigma}$ sufficiently low that mistakes are sufficiently unlikely that $\vec{a}^l$ is $\theta'$-forgiving for some $\theta' < \theta$ such that the result holds. ∎

# References


Abreu, D., D. Pearce, and E. Stacchetti (1993): "Renegotiation and Symmetry in Repeated Games," *Journal of Economic Theory*, 60, 217–240.


[6] Since $\vec{a}^*$ can be chosen to be $\vec{\sigma}^-$-optimal, this supposition will fail if and only if $a_1$ is exactly $\sigma_1$-optimal in response to $a_2$. In this case, it can be shown that the modified $\tilde{a}_2^l$ that switches to a $\sigma_1^-$-optimal response to $b_1$ rather than reverting to $a_1$ is a $\sigma_1'$-optimal response to $b_1^l$.

Aumann, R. J. (1957): "Acceptable Points in General Cooperative $n$-Person Games," in *Contributions to the Theory of Games IV,* Annals of Mathematics Study 40, ed. by R. D. Luce, and A. W. Tucker, pp. 287–324. Princeton University Press, Princeton NJ.

Aumann, R. J., and L. Shapley (1976): "Long-Term Competition—A Game-Theoretic Analysis," Mimeo, Hebrew University. Reprinted in *Essays in Game Theory*, ed. by N. Megiddo (1994). Springer-Verlag, New York.

Axelrod, R. (1981): "The Emergence of Cooperation Among Egoists," *American Political Science Review*, 75, 306–318.

——— (1984): *The Evolution of Cooperation.* Basic Books, New York.

Axelrod, R., and W. Hamilton (1981): "The Evolution of Cooperation," *Science*, 211, 1390–1396.

Binmore, K. G., and L. Samuelson (1992): "Evolutionary Stability in Repeated Games Played by Finite Automata," *Journal of Economic Theory*, 57, 278–305.

Boyd, R. (1989): "Mistakes Allow Evolutionary Stability in the Repeated Prisoner's Dilemma Game," *Journal of Theoretical Biology*, 136, 47–56.

Boyd, R., and J. Lorberbaum (1987): "No Pure Strategy is Evolutionarily Stable in the Repeated Prisoners' Dilemma Game," *Nature*, 327, 58–59.

Ellison, G. (2000): "Basins of Attraction, Long-Run Stochastic Stability, and the Speed of Step-by-Step Evolution," *Review of Economic Studies*, 67, 17–45.

Farrell, J., and E. Maskin (1989): "Renegotiation in Repeated Games," *Games and Economic Behavior*, 1, 327–360.

Farrell, J., and R. Ware (1988): "Evolutionary Stability in the Repeated Prisoner's Dilemma Game," *Theoretical Population Biology*, 36, 161–166.

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

——— (2003): "Learning, Hypothesis Testing, and Nash Equilibrium," *Games and Economic Behavior*, 45, 73–96.

——— (2005): "Regret Testing: A Simple Payoff-Based Procedure for Learning Nash Equilibrium," Working Paper.

Fudenberg, D., and E. Maskin (1986): "The Folk Theorem in Repeated Games with Discounting or with Incomplete Information," *Econometrica*, 54, 533–554.

——— (1990): "Evolution and Cooperation in Repeated Games," *American Economic Review Papers and Proceedings*, 80, 274–279.

——— (1991): "On the Dispensability of Public Randomization in Discounted Repeated Games," *Journal of Economic Theory*, 53, 428–438.

Fudenberg, D., and J. Tirole (1991): *Game Theory*. The MIT Press, Cambridge, Massachusetts.

Hart, S., and A. Mas-Colell (2003): "Uncoupled Dynamics Do Not Lead to Nash Equilibrium," *American Economic Review*, 93, 1830–1836.

——— (2004): "Stochastic Uncoupled Dynamics and Nash Equilibrium," Working Paper.

Kalai, E., and E. Lehrer (1993): "Rational Learning Leads to Nash Equilibrium," *Econometrica*, 61, 1019–1045.

Kandori, M., G. J. Mailath, and R. Rob (1993): "Learning, Mutation and Long-Run Equilibria in Games," *Econometrica*, 61, 29–56.

Kim, Y. (1994): "Evolutionary Stable Strategies in the Repeated Prisoner's Dilemma," *Mathematical Social Sciences*, 28, 167–197.

Pearce, D. (1987): "Renegotiation-Proof Equilibria: Collective Rationality and Intertemporal Cooperation," Cowles Foundation Discussion Paper No. 855.

Robson, A. J. (1990): "Efficiency in Evolutionary Games: Darwin, Nash and the Secret Handshake," *Journal of Theoretical Biology*, 144, 379–396.

Rubinstein, A. (1979): "Equilibrium in Supergames with the Overtaking Criterion," *Journal of Economic Theory*, 21, 1–9.

Selten, R. (1983): "Evolutionary Stability in Extensive Two-Person Games," *Mathematical Social Sciences*, 5, 269–363.

Sugden, R. (1986): *The Economics of Rights, Cooperation and Welfare*. Basil Blackwell, Oxford.

van Damme, E. (1987): *Stability and Perfection of Nash Equilibria*. Springer Verlag, Berlin.

——— (1989): "Renegotiation-Proof Equilibria in Repeated Prisoner's Dilemma," *Journal of Economic Theory*, 47, 206–217.

Weibull, J. W. (1995): *Evolutionary Game Theory*. The MIT Press, Cambridge, Massachusetts.

Young, H. P. (1993): "The Evolution of Conventions," *Econometrica*, 61, 57–84.