**Title:** The Castle on the Hill

**Author:** David K. Levine

Department of Economics

UCLA

Los Angeles, CA 90095

phone/fax 310-825-3810

email dlevine@ucla.edu

**Proposed Running Head:** Castle on the Hill

Forthcoming: *Review of Economic Dynamics*

# The Castle on the Hill[*]

First version: August 5, 1998

Final Version: January 4, 2000

David K. Levine

**Abstract:** A simple example of a stochastic game with irreversibility is studied and it is shown that the folk theorem fails in a robust way. In this game of Castle on the Hill, for a broad range of discount factors, including those close to one, equilibrium is unique. Moreover, the equilibrium for large discount factors is Pareto dominated by the equilibrium for low discount factors. A unique cyclic equilibrium is also possible for intermediate ranges of discount factors.

[*] I would like to thank NSF grant SBER 9617899 and the UCLA Academic Senate for financial support. An especial debt of gratitude is owed an anonymous referee who discovered an error in an earlier version of the main theorem.

## 1. Introduction

A simple example of a stochastic game with irreversibility is studied and it is shown that the folk theorem fails in a robust way. In this game of Castle on the Hill, for a broad range of discount factors, including those close to one, equilibrium is unique. Moreover, the equilibrium for large discount factors is Pareto dominated by the equilibrium for low discount factors. A unique cyclic equilibrium is also possible for intermediate ranges of discount factors.

Consider first an infinitely repeated stage-game, such as the prisoners' dilemma, with discounting and public randomization. A great deal is known about the structure of sub-game perfect equilibrium payoffs as a function of the discount factor. For zero discount factor, the equilibrium payoffs are the convex hull of static Nash payoffs, which will be a singleton if, as in the prisoners' dilemma, the static equilibrium is unique. As the discount factor increases, the set of equilibrium payoffs are non-decreasing convex sets, typically increasing in sudden jumps. If certain technical conditions are satisfied[1], as the discount factor approaches one, the folk theorem holds: the set of equilibrium payoffs approaches the entire individually rational socially feasible set. Even if we allow for imperfect public information, the work of Abreu, Pearce and Stacchetti [1988] and Fudenberg, Levine and Maskin [1994] show that this basic picture is unchanged for perfect public equilibrium.

An important generalization of repeated games are stochastic games. Here there is a finite collection of different states. In each state a particular stage-game is played, and that play determines not only stage-payoffs, but also determines Markov transition probabilities over which state occurs next. For an important class of these games, including those in which regardless of play every state has a positive probability of being reached in finite time, Dutta [1991] shows that the folk theorem continues to hold. As we shall see in this paper, monotonicity can fail, even for this class of games. However, there are important classes of games in which irreversibility is important, that is, in which it may be impossible when certain strategies are employed to move from one state to another. One type of irreversibility is the existence of absorbing states. This is the case, for example, in the Rubinstein-Stahl bargaining model and the war-of-attrition game, both of which have been studied extensively. Notice, however, that both of these are folk theorem games; this is established for the finite version of the Rubinstein-Stahl bargaining model by Binmore, Shaked and Sutton [1985]. To see that this is the case in the war-of-attrition, notice that a delay followed by a public randomization over one player winning, such that each player gets (including the delay) in expected value more than their minmax level, is in fact subgame perfect. These results, though similar to the purely repeated case, are not as robust. In the bargaining model the folk theorem is not robust to a continuum of actions, since in that case Rubinstein [1982] showed that there is a unique equilibrium. In the war-of-attrition, the folk theorem fails without public randomization. In neither case has the issue of monotonicity with public randomization been studied.

Another class of stochastic games to which Dutta’s theorem does not apply are games in which there are no absorbing states, but in which movement between different states is possible only with some degree of agreement among the players. The goal of this paper is to study a simple example of such a game, and show that these games can be radically different than repeated games. The type of game we study is one in which positional advantage is important, and expensive to defend. The specific example is a two-player two-state game, which is symmetric in the sense that the role of the two players is reversed in the two states. In each state, one player is a Lord and receives a high utility and the other a Serf who receives a lower utility. Each player has two actions: the Lord may defend his castle, which is costly, or not. If the castle is left undefended, the Serf may attack the castle, which is also costly, but results in his becoming Lord in the following period. Notice that regardless of the state, the other state is reachable, but only if the players agree, in the sense that the Lord fails to defend, and the Serf attacks. This may be thought of as a useful paradigm for situations in which position has an advantage, for example, the Lord might be a monopolist, and the Serf an innovator. Here the monopolist must decide whether to engage in costly patenting activity that will preempt the entrant, and prevent him from introducing an innovation that will result in the entrant assuming the monopoly position. This example has a number of features that differentiate it from a simple repeated game:

- For a range of discount factors including all discount factors near one (and all discount factors near zero) the equilibrium is unique.
- The equilibrium for discount factors near one is Pareto dominated by the equilibrium with discount factors near zero. Basically, for low discount factors, the Lord will not bother to defend, nor the Serf to attack. However, for high discount factors, the Serf will wish to attack and this forces the Lord to defend. The result is that resources are wasted on defense.
- For intermediate ranges of discount factors and when it is much more costly to defend than to attack, there is a unique equilibrium. This equilibrium is cyclic – the two players alternate between being Lord and Serf. It is also inefficient.

We also consider how robust these results are to introducing a small exogenous probability that the roles are reversed regardless of play in the stage game. Here Dutta's conditions apply, so the folk theorem must hold for discount factors sufficiently close to one. However, if the exogenous probability of reversal is small enough, then there will still be a range of large discount factors for which there is still a unique equilibrium in which the Lord defends and the Serf attacks. For discount factors near zero, the unique equilibrium is still neither defense nor attack. This implies that monotonicity fails: the set of equilibria for the large discount factors is not a superset of that for small discount factors. Moreover, if the exogenous probability of reversal is small enough relative to the subjective interest rate, the irreversible model much more accurately captures the set of equilibria than does the folk theorem.

## *2. The Castle on the Hill*

In the two-player stage-game of *castle on the hill*, there are two player roles, that of a Lord and that of a Serf. The Lord lives in a castle at the top of a hill, and the Serf lives in a village beneath the hill. The Lord moves first and must choose between two options: defending the castle or consuming. If he defends the castle the game ends, he receives a payoff of $x - d$ where $d$ is the cost of defense, the Serf receives a payoff of 1, and next period the Lord retains the castle. If the Lord chooses to consume he receives a payoff of $x$, and the Serf must choose whether to attack the castle or consume. If he attacks the castle the Serf receives 0. However, since the castle is not defended the attack succeeds, so next period the role of the two players is reversed – the Serf occupies the castle and becomes a Lord, and the Lord is exiled to the village and becomes a Serf. If, on the other hand, the Serf chooses to consume he receives 1 and the Lord retains the castle. The extensive and normal forms of this game are illustrated below.

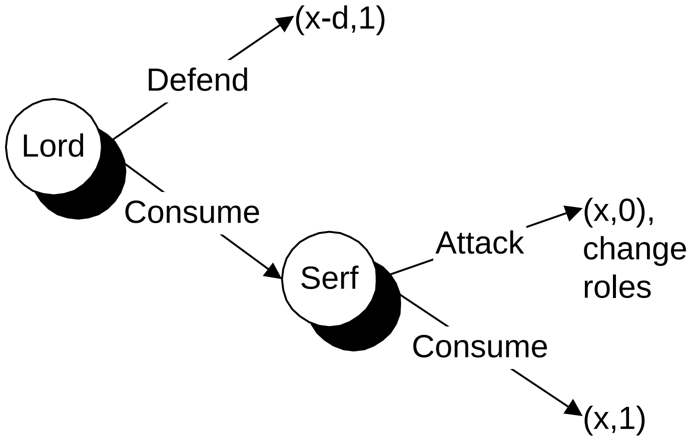

Figure 1: Extensive Form of Castle on the Hill

| Lord/Serf | A | C |
|---|---|---|
| D | $x-d$ ,1 | $x-d$ ,1 |
| C | $x$ ,0 [change roles] | $x$,1 |

Figure 2: Normal Form of Castle on the Hill

In the infinitely repeated game, both players maximize average present value of utility, and are equally patient with common discount factor $0 \le \delta < 1$.

We will examine the case in which it is considerably better to be Lord than Serf. In particular we assume that it is better to be a Lord and pay double the defense cost than it is to be a serf.

***Assumption:*** $x-2d>1$.

We will also be interested in whether or not $d>1$; if so, we will say that it is *more expensive to defend than to attack.*

### *3. Analysis of the Game*

If the discount factor is zero or equivalently the game is played only once, there is a unique subgame perfect equilibrium: the Serf will consume, and the Lord will consume. In fact any Nash equilibrium results in the same utility, as it is strictly dominant for the Lord to consume, and weakly dominant for the Serf to consume. As we shall see the unique subgame perfect equilibrium extends to discount factors close to zero.

To put the game in the context of the folk theorem, it is useful to begin by describing the socially feasible individually rational set. Because the game is not repeated, the socially feasible payoff set depends on the discount factor. With public randomization, when the discount factor is zero, it is the triangular region shown in Figure 3.

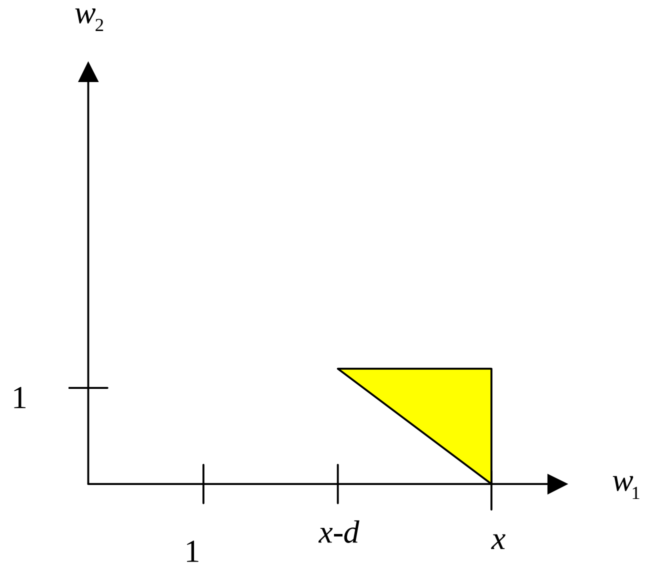

Figure 3: Socially Feasible Set for $\delta = 0$

As the discount factor increases, the socially feasible payoff set (with public randomization) is monotone, and in the limit as the discount factor approaches one, it approaches the convex hull of the triangular set and its mirror image as shown in Figure 4.

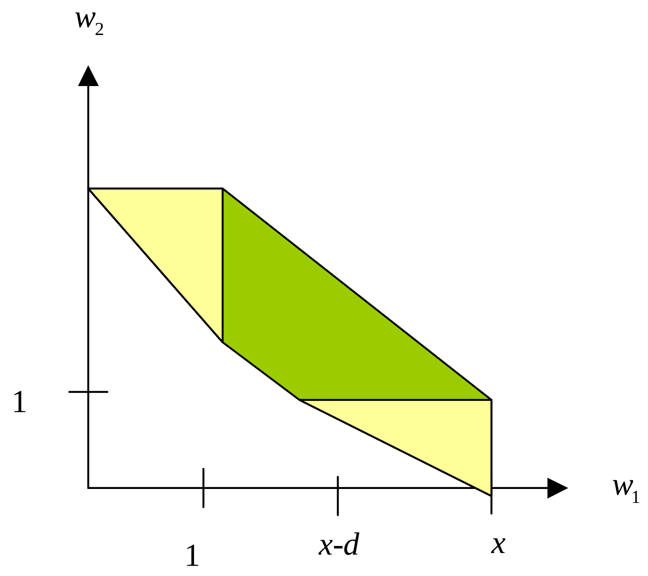

Figure 4: Socially Feasible Set for $\delta \to 1$

By way of contrast, the minmax points depend on the state, but are independent of the discount factor. The minmax for Lord is $x-d$; he can guarantee this amount by defending and can be held to this by the Serf attacking. The minmax for the Serf is 1; he can guarantee this amount by consuming and can be held to this amount by the Lord defending.

Our main theorem gives a partial characterization of equilibrium payoffs.

***Theorem:***

(A) If $0 \le \delta < \min\left\{\frac{d}{x-1+d}, \frac{1}{x}\right\}$ there is a unique Nash equilibrium: the Lord and Serf both consume regardless of the history.

(B) If $d > 1$ so it is more expensive to defend than to attack, if $x > (2d-1)/(d-1)$ and $\frac{1}{x-1} < \delta < \frac{d}{x-1+d}$ then there is a unique Nash equilibrium: the Lord consumes and the Serf attacks, regardless of the history. In this case, players alternate between being Lord and Serf.

(C) If $\delta > \max\left\{\frac{1}{x-2d}, \frac{d}{x-1-d}\right\}$ there is a unique subgame perfect equilibrium: the Lord defends and the serf attacks regardless of the history.

*Proof:* If the Lord consumes, he gets at least $(1-\delta)x+\delta$; if he defends now, he gets no more than $(1-\delta)(x-d)+\delta x$. So consumption is dominant if $(1-\delta)x+\delta > (1-\delta)(x-d)+\delta x$, or $(1-\delta)d > \delta(x-1)$. Suppose that this is the case, so that it is always optimal for the Lord to consume. If the Serf consumes, he gets 1. If he attacks, then he gets at most $0, x, x, \ldots$ yielding a present value of $\delta x$, and he gets no less than $0, x, 0, x, \ldots$ yielding a present value of $\delta x/(1+\delta)$. So it is certainly optimal for the

Serf to consume if $1>\delta x$. Combining this condition, for the condition for the Lord to consume gives (A).

On the other hand, if $\delta x/(1+\delta)>1$ then certainly the Serf will attack. A necessary condition for a discount factor that satisfies this inequality together with the condition for the Lord consuming is

$$\frac{1}{x-1}<\frac{d}{x-1+d}.$$

This is equivalent to $2d-1<(d-1)x$; if $d<1$ this condition is inconsistent with $x-2d>1$, so it is necessary that $d>1$ and $x>(2d-1)/(d-1)$, which give (B).

To prove (C) recall that the minmax for Lord is $x-d$. Let $w_L, w_S$ be the Lord and Serf average present value respectively in some subgame perfect. Social feasibility implies that $w_L+w_S\le x+1$. Since the Lord gets at least his minmax $w_L\ge x-d$ it must be that $w_S\le 1+d$.

Suppose now that the Lord chooses to consume. If the Serf chooses to attack, he gets at least $\delta(x-d)$, since he will be Lord next period. If he consumes, he gets at best $(1-\delta)+\delta[1+d]$. Since $\delta>1/(x-2d)$, this means that it is better to attack than consume. So in any subgame perfect equilibrium, the Serf must attack.

Turning to the Lord, if he consumes he gets at most $(1-\delta)x+\delta(1+d)$, while he can get at least $x-d$. Since $\delta>d/(x-1-d)$, he must defend.

☑

## *4. Robustness to Noise*

How robust are these results? The condition that every state be reached with positive probability from every other state is a generic one, so irreversibility of the type considered here is non-generic. Suppose we consider the generic case. In particular, let us introduce a small exogenous probability that the roles are reversed regardless of play in the stage game. Then Dutta's condition applies, so the folk theorem must hold for discount factors sufficiently close to one. However, fixing $\delta$ in any of the cases (A), (B) or (C) above, if the exogenous probability of reversal is small enough, then because the inequalities in the proof are strict, they will not be violated, and the unique equilibrium will be unchanged and still unique. In other words, there will still be a range of large discount factors for which there is still a unique equilibrium in which the Lord defends and the Serf attacks. For discount factors near zero, the unique equilibrium is still neither defense nor attack. Notice that this implies that monotonicity fails: the set of equilibria for the large discount factors is not a superset of that for small discount factors. Moreover, if the exogenous probability of reversal is small enough relative to the subjective interest rate, the irreversible model much more accurately captures the set of equilibria than does the folk theorem.

## *References*

Abreu, D., D. Pearce and E. Stacchetti (1988), “Towards a Theory of Discounted Repeated Games with Imperfect Monitoring,” Princeton.

Abreu, D., P. K. Dutta and L. Smith (1992), “Folk Theoems for Repeated Games: A Neu Condition,” Unpublished Manuscript.

Binmore, K., A. Shaked and J. Sutton (1985), “Testing Non-Cooperative Bargaining Theory: A Preliminary Study,” *American Economic Review*, Vol. 75, No. 5, 1178-1180.

Dutta, P. K. (1991), “A Folk Theorem for Stochastic Games,” University of Rochester, Rochester Center for Economic Research, Working Paper No. 293.

Fudenberg, D. and E. Maskin (1986), “The Folk Theorem for Repeated Games with Discounting and Incomplete Information,” *Econometrica*, 54: 533-54.

Fudenberg, D., D. K. Levine and E. Maskin (1994), “The Folk Theorem with Imperfect Public Information,” *Econometrica*, 62: 997-1039.

Rubinstein, A. (1982), “Perfect Equilibrium in a Bargaining Model,” *Econometrica*, 97-110.

# Figures

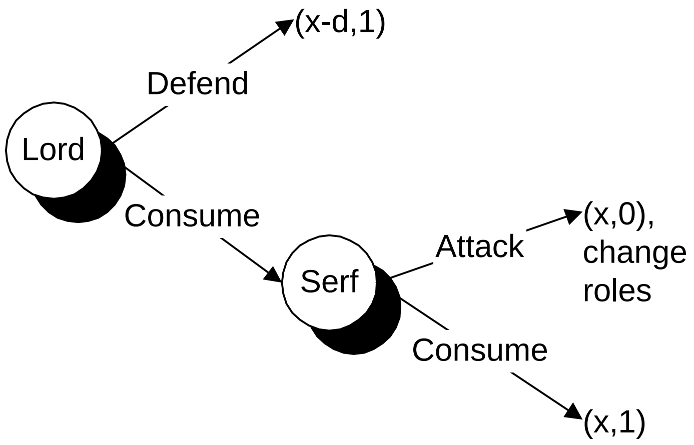

Figure 1: Extensive Form of Castle on the Hill

| Lord/Serf | A | C |
|---|---|---|
| D | $x-d$ ,1 | $x-d$ ,1 |
| C | $x$ ,0 [change roles] | $x$,1 |

Figure 2: Normal Form of Castle on the Hill

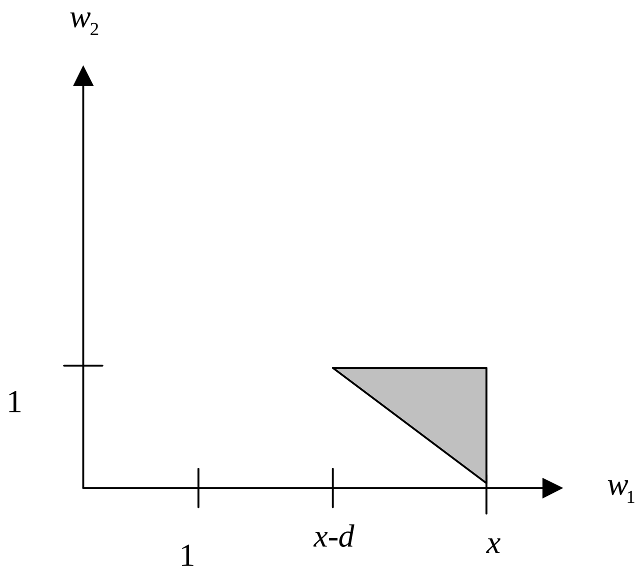

Figure 3: Socially Feasible Set for $\delta = 0$

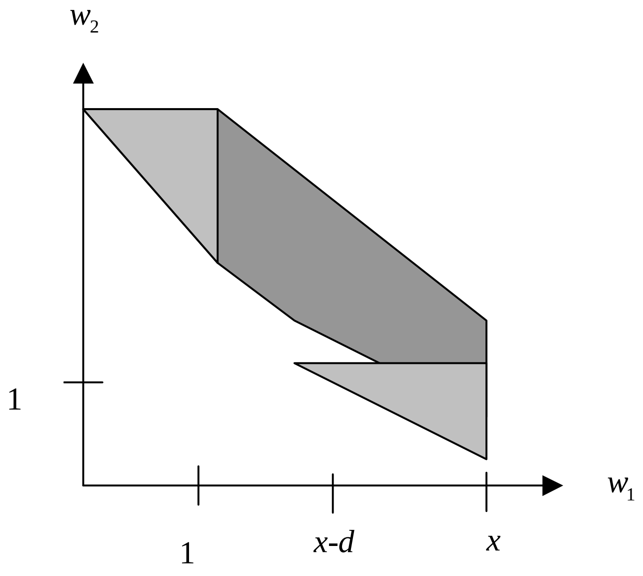

Figure 4: Socially Feasible Set for $\delta \to 1$

# Symbol Table

$d$ lower case latin italic dee

$x$ lower case latin italic ex

$\delta$ lower case greek delta

$w_1$ lower case latin w sub one

$w_2$ lower case latin w sub two

# Endnotes

[1] See Abreu, Dutta and Smith [1992] or Fudenberg and Maskin [1986].