A note on Bagwell’s paradox and forward induction in three classic games

José Luis Ferreira*
Universidad Carlos III de Madrid

October, 2010

Abstract

Stackelberg-like games show a situation where there is a clear advantage in moving first. In a thought provoking article, Bagwell (1995) shows that this advantage may not be robust against imperfect observations of the first move. We explore these ideas in the context of forward induction in three classic games, namely, the outside option game, the game of burning a dollar before the battle of the sexes and the beer-quiche game. JEL numbers: C72, G13, L13.

Keywords: Bagwell’s paradox, Noise, Battle of the sexes, Beer-Quiche game, Outside option, Forward induction.

*Departamento de Economía. Universidad Carlos III de Madrid, 28903 Getafe, Madrid, Spain. Tel. 34-91 6249580. jlferr@eco.uc3m.es. Financial support from grant 2008/00055/001 (Ministerio de Educación, Spain) is gratefully acknowledged. The author thanks two anonymous referees whose comments improved the analysis and the presentation of this work. Only the author is responsible for any errors that may remain. The last versions of the paper were written while the author was visiting the ESI at Chapman University.
1 Introduction

In a thought provoking article, Bagwell (1995) shows that the advantage of a first mover (e.g., in a Stackelberg duopoly) may disappear when her action is not perfectly observed by the player moving second. This result is remarkable because it still holds if the probability of not observing the first action is arbitrarily close to zero. The robustness of models *a la* Stackelberg seemed to be under question, but not for long since, after its publication, several works developed ways of modeling the imperfect observability that restore the first mover’s advantage. These include van Damme and Hurkens (1997), Güth *et al.* (1998), and Maggi (1999). Perhaps the most straightforward way of restoring the first mover advantage in pure strategies is presented in Adolph (1996), where it is noticed that mistakes can restore the commitment prediction if the probability of making a mistake is greater that the probability of observing the wrong signal. However, it should be noted that there are other interesting equilibrium selection procedures that do not solve the paradox, as shown in Oechsler and Schlag (2000).

It must also be stressed that the paradox, as stated originally, is about equilibria in pure strategies. Without noise there is a unique subgame perfect equilibrium. However, with noise there are multiple equilibria and, among them, a mixed Nash equilibrium that is close to the original subgame perfect equilibrium and that converges to it as the level of noise converges to zero.

Without noise, it is optimal for the first mover to make her choice publicly. With noise, this may not necessarily be optimal, as the course of the play depends on beliefs of what happens. In fact, the contribution of van Damme and Hurkens (1997) is to provide a theory of equilibrium selection that favors the mixed strategy equilibrium that is close to the original equilibrium.

Motivated by these works, we explore whether the Bagwell’s paradox is present in other games in which there is an advantage of playing first, namely the outside option game, the game of burning a dollar before the battle of the sexes and the Beer-Quiche game. In the outside option game we find that the selection of equilibria made by appropriate equilibrium concepts (like the iterative elimination of weakly dominated strategies and Kohlberg Mertens stable set) is not affected by the introduction of noise. The same equilibrium concepts fail to select the forward induction equilibrium when noise is added to the “burning a dollar” game. Hence, a paradox similar to Bagwell’s appears in the “burning a dollar” game, but not in the outside option games. In both games, the introduction of mistakes destroys the selection of the equilibrium that is compatible with forward induction arguments, and thus cannot restore the equilibrium selection when it is lost due to the introduction of noise. In the Beer-Quiche game we show that, for some
specifications of the relative sizes of the probabilities of the noisy signals, the “right” equilibrium is selected, but for other specifications other equilibria are also selected, and for even other specifications the “right” equilibrium is not selected. The first set of specifications include, as a particular case, a model already discussed in Carlsson and Dasgupta (1997), a work developed independently from Bagwell (1995) and the articles that followed.

The rest of the work is organized as follows. Section 2 discusses the general ideas in the motivating example in Bagwell (1995). Section 3 analyzes the three games and Section 4 concludes.

2 The paradox

In this section we present the example in Bagwell (1995), and then introduce mistakes (in addition to noise) as a new element. This example reviews the original paradox and the role of mistakes in its solution before moving to other classes of games. Consider the $2 \times 2$ game in Table 1.

<table>
<thead>
<tr>
<th></th>
<th>$S$</th>
<th>$C$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>$S$</strong></td>
<td>5, 2</td>
<td>3, 1</td>
</tr>
<tr>
<td><strong>$C$</strong></td>
<td>6, 3</td>
<td>4, 4</td>
</tr>
</tbody>
</table>

Table 1

The unique Nash equilibrium of this game is $(C, C)$, with payoffs $(4, 4)$, whereas if Player 1 moves first in a perfectly observable fashion, the only subgame perfect equilibrium is $(S, (S, C))$, where $(S, C)$ is the strategy for Player 2 who chooses $S$ if $S$ is observed and $C$ if $C$ is observed, with payoffs $(5, 2)$. Now suppose that Player 2 does not observe 1’s action, but receives a signal $\phi \in \{s, c\}$ such that

$$Pr(\phi = s \mid S) = Pr(\phi = c \mid C) = 1 - \varepsilon,$$

where $\varepsilon \in (0, 1)$. In this case the only Nash equilibrium (and Bayesian equilibrium) in pure strategies is $(C, (C, C))$, where $(C, C)$ is the strategy for Player 2 who chooses $C$ regardless of the value of $\varepsilon$. Since the equilibrium is valid no matter how small $\varepsilon$ is, the interpretation of this fact is the source of the paradox: Player 1 can take no advantage in moving first if his action is not perfectly observable (for details, see Bagwell, 1995.)

To see that there is no equilibrium in which Player 1 moves $S$, suppose that $(S, (S, S))$ is proposed. Notice now that $(S, S)$ is Player 2’s unique best
reply against $S$, and that Player 1 knows that her profitable deviation from $S$ to $C$ will go unnoticed as long as $\varepsilon > 0$. Thus Player 1 cannot play $S$ in an equilibrium.

Consider an alternative source of uncertainty. When Player 1 intends to choose one action, she makes a mistake with probability $\lambda$ and chooses the other action instead. More precisely, Player 1 can choose between $\tilde{S}$ and $\tilde{C}$, where $\tilde{S} = (1 - \lambda)S + \lambda C$ and $\tilde{C} = (1 - \lambda)C + \lambda S$. Now, the only SPE is, again $(S, (S, C))$. If Player 1 chooses $S$ in equilibrium, action $C$ can still be chosen by mistake, and the reaction to $C$ is $C$. Therefore, Player 1 does not want to deviate. When both noise and mistakes are present, the equilibrium may or may not show a first mover advantage, depending on the relative probabilities of occurrence. For details see Adolph (1996).

3 Forward induction

In this section we explore the implications of noisy observability of a commitment action in the context of forward induction. We show three examples. The first one is a game with an outside option in which forward induction selects the equilibrium that is more beneficial for the player that can opt out. The second is the classic game of the battle of the sexes in which one player has the option of publicly burning a dollar before playing the game, and, paradoxically, gaining an advantage by being able to do so. The third is the Beer-Quiche game.

The interest of these examples lies in the fact that the addition of noise a la Bagwell has completely different consequences. In the first example, the addition of noise does not affect the selection of the equilibrium implied by forward induction, while in the second it does. In the third example, the way the addition of noise affects the equilibrium selection depends on the relative sizes of the noise. The addition of mistakes, that is useful to restore the first mover advantage in Bagwell’s example does not help in our games, as always destroys the equilibrium selection and, thus, cannot restore it.

Before we proceed we must be precise about the meaning of the forward induction concept. The intuitive idea says just that, when a player moves first, she must have the capability of inducing her preferred equilibrium when her action clearly indicates which one it is. However, it is hard to translate this idea into a formal definition that works for any game. We take the definition in van Damme (1987).

Definition 1 (Van Damme, 1987)): A solution concept $S$ is consistent with forward induction in the class of generic 2-person games if $p \in S$ for
any path $p$ for which there does not exist a player $i$ who by unilaterally
deviating from $p$ can enforce that a subgame is reached for which exactly one
solution (according to $S$) yields this player more than $p$ does and for which
all other solutions yields this player less.

The most widely used equilibrium refinement that has been used to for-
malize forward induction arguments is the Kohlberg Mertens stable set,
KMSS, (as defined in Kohlberg and Mertens, 1986). Although, in gen-
eral, it is not compatible with van Damme’s definition, they agree in the
three examples.

3.1 The outside option game

Consider the following game. When Player 1 chooses $O$ the game ends with
payoffs $<2, 2>$, while if she chooses $I$ a battle of the sexes game follows. The
normal form of this game (after identifying $OF$ and $OT$ as the same strategy
$O$) is presented in Table 2.

<table>
<thead>
<tr>
<th>Player 1</th>
<th>$F$</th>
<th>$T$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$O$</td>
<td>2, 2</td>
<td>2, 2</td>
</tr>
<tr>
<td>$IF$</td>
<td>3, 1</td>
<td>0, 0</td>
</tr>
<tr>
<td>$IT$</td>
<td>0, 0</td>
<td>1, 3</td>
</tr>
</tbody>
</table>

Table 2

There are two pure-strategies equilibria in this game: $(O, T)$ and $(IF, F)$,
but only the second satisfies the forward induction criterium. Iterative elimi-
nation of (weakly) dominated strategies ($IEDS$) selects this equilibrium,
which also constitutes the only Kohlberg Mertens stable set.

The introduction of noise produces the extensive form that can be seen
in Figure 1. The normal form is shown in Table 3.

<table>
<thead>
<tr>
<th></th>
<th>$FF$</th>
<th>$FT$</th>
<th>$TF$</th>
<th>$TT$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$O$</td>
<td>2, 2</td>
<td>2, 2</td>
<td>2, 2</td>
<td>2, 2</td>
</tr>
<tr>
<td>$IF$</td>
<td>3, 1</td>
<td>$3 - 3\varepsilon; 1 - \varepsilon$</td>
<td>$3\varepsilon; \varepsilon$</td>
<td>0, 0</td>
</tr>
<tr>
<td>$IT$</td>
<td>0, 0</td>
<td>$\varepsilon; 3\varepsilon$</td>
<td>$1 - \varepsilon; 3 - 3\varepsilon$</td>
<td>1, 3</td>
</tr>
</tbody>
</table>

Table 3

The extensive form clearly shows that the noise-game has no subgames
other than the whole game, hence van Damme’s definition of forward induction
is not applicable. For any $\varepsilon \in (0, 1)$, the pure-strategies equilibria are
Figure 1: The Outside option game with noise

(O, TF), (O, TT) and (IF, FF), but only the last one survives IEDS (in a first round IT is eliminated, in a second round FT, TF and TT, and finally, O). Thus, we see that, although the formal definition of forward induction is not applicable, the concept of IEDS still captures its intuition (in the sense that the equilibrium outcome does not vary when introducing a small probability of noise.) Since the Kohlberg Mertens stable set is consistent with IEDS and always exists, it follows that (IF, FF) is the only pure strategy equilibrium in a KMSS.

We now introduce mistakes the following way. When Player 1 wants to play I she actually plays \((1 - \lambda) I + \lambda O\), and when she wants to play O she plays \(\lambda I + (1 - \lambda) O\). After I or O is actually played, the game follows as in the noise-game case. As with the addition of noise, the addition of mistakes makes the definition of forward induction inapplicable, and also makes the refinements not to select the same forward induction outcome. The normal form (where the matrix has been split for convenience) is depicted in Table 4.
In this game both (IF; FF) and (OT; TT) are strict equilibria, and a fortiori survive the process of IEDS, and belong to different KMS. Thus the forward induction outcome is not selected. If there is perfect observability ($\varepsilon = 0$), the strategies FF and FT become strategy F, and strategies TF and TT become strategy T. Again (IF; F) and (OT; T) are strict equilibria. Finally, if $\lambda = 0$, we are back in the case of imperfect observability and no mistakes analyzed above. As a conclusion we have that, in this game, the concept of IEDS preserves the intuition of forward induction in the presence of imperfect observability, but not in the presence of mistakes.

It is interesting to understand why (OT; TT) is a strict equilibrium in the game with mistakes. In the game without mistakes, the follower thinks that if the leader chooses to proceed to the second stage she will aim at her preferred equilibrium. As in the previous section, this fact will be understood by the follower no matter what is the signal. But in the game of mistakes, the leader may proceed to the second stage by mistake, and when she learns that she mistakenly proceeded to the second stage, she can then decide which action to play. To put it in other words, there are two theories of why the game proceeds to the second stage. The first one is that Player 1 wanted it because she is aiming for her preferred equilibrium. The second one is that she did not want it, but made a mistake, in which case there is no presumption of aiming at a particular equilibrium. When both theories are possible, the result is that forward induction has a competitor, and thus loses its prediction power. For forward induction to work, the leader’s action has to convey her intention, but the possibility of mistakes weakens the link between actions and intentions.
The features of this example can be extended to the general class of outside option games as defined in van Damme (1989).  

### 3.2 Burning a dollar before playing Battle-of-the-sexes

Consider the battle of the sexes game as shown in Table 5.

<table>
<thead>
<tr>
<th></th>
<th>F</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>Player 1 (F)</td>
<td>3, 1</td>
<td>0, 0</td>
</tr>
<tr>
<td>Player 1 (T)</td>
<td>0, 0</td>
<td>1, 3</td>
</tr>
</tbody>
</table>

Table 5

Suppose now that, previous to this game, Player 1 decides whether to burn a dollar (B) or not (N) in an action that is perfectly observed by Player 2. The normal form of this new game (after deletion of redundant strategies) is shown in Table 6, where the first (alt. second) letter in the strategy of Player 2 indicates his action when B (N) is observed.

<table>
<thead>
<tr>
<th></th>
<th>FF</th>
<th>FT</th>
<th>TF</th>
<th>TT</th>
</tr>
</thead>
<tbody>
<tr>
<td>BF</td>
<td>2, 1</td>
<td>2, 1</td>
<td>-1, 0</td>
<td>-1, 0</td>
</tr>
<tr>
<td>BT</td>
<td>-1, 0</td>
<td>-1, 0</td>
<td>0, 3</td>
<td>0, 3</td>
</tr>
<tr>
<td>NF</td>
<td>3, 1</td>
<td>0, 0</td>
<td>3, 1</td>
<td>0, 3</td>
</tr>
<tr>
<td>NT</td>
<td>0, 0</td>
<td>1, 3</td>
<td>0, 0</td>
<td>1, 3</td>
</tr>
</tbody>
</table>

Table 6

Although, as it is written, the extensive form shows a non-generic game, the payoffs can easily be changed to make a generic one. There are four Nash equilibria in pure strategies, namely (BF, FT), (NF, FF), (NF, TF), and (NT, TT). However, according to Definition 1, only (NF, FF) is consistent with forward induction, which is also the only Nash equilibrium that survives IEDS (first eliminate BT, then TF and TT, next NT, FT and BF in that order).

In order to study the consequences of introducing noise, suppose that after Player 1 chooses her first action (not observed by Player 2), Player 2 observes the signal $\phi \in \{b, n\}$, with $\Pr(\phi = b \mid B) = \Pr(\phi = n \mid N) = 1 - \varepsilon$. This definition is natural in the new setting. Figure 2 shows the extensive form of the game. The normal form (after grouping redundant strategies) of the game is appears in Table 7.

---

1 Also, other equilibrium refinements may be used in this class of games. An interesting one is a version of evolutionary stability in Hauk and Hurkens (1986) that agrees with forward induction in this class of games.
The first (alt. second) letter in the strategy of Player 2 indicates his action when $b$ (alt. $n$) is observed. In this new game, van-Damme’s definition is no longer applicable, as there are no proper subgames. Nevertheless, we can still look for equilibria that survive $IEDS$. Iterative elimination of weakly dominated strategies stops after deletion of $BT$ (weakly dominated by $NT$) and $TF$ (weakly dominated by $\alpha FT + (1 - \alpha)FF$, where $\alpha = \frac{\epsilon}{2\delta}$). The only Nash equilibria in pure strategies are $(NF, FF)$ and $(NT, TT)$. Both equilibria are strict and cannot be eliminated by standard refinements based on mistakes or other perturbations. This example shows that, unlike in the outside option game, the introduction of noisy observability may destroy the
first mover advantage. Also, because both equilibria are strict, the introduction of errors cannot restore the original equilibrium selection when the size of the mistakes goes to zero.

### 3.3 The Beer-Quiche game

In the Beer-Quiche game, Player 1 can be of two types, \( w \) and \( s \), and has to choose between \( Q \) and \( B \). Player 2 observes the choice of Player 1, but not her type, and chooses between \( F \) and \( NF \). The game has the following normal form shown in Table 8 (see Figure 3 with \( \varepsilon_w = \varepsilon_s = 0 \) for the extensive form.)

<table>
<thead>
<tr>
<th>Player 1</th>
<th>((Q,Q))</th>
<th>((Q,B))</th>
<th>((B,Q))</th>
<th>((B,B))</th>
</tr>
</thead>
<tbody>
<tr>
<td>((NF,NF))</td>
<td>2.1, 0.9</td>
<td>3.0, 0.9</td>
<td>2.0, 0.9</td>
<td>2.0, 0.9</td>
</tr>
<tr>
<td>((NF,F))</td>
<td>2.1, 0.9</td>
<td>1.2, 0.1</td>
<td>2.8, 1.0</td>
<td>1.0, 0.1</td>
</tr>
<tr>
<td>((F,NF))</td>
<td>0.1, 0.1</td>
<td>1.8, 1.0</td>
<td>0.2, 0.0</td>
<td>0.0, 0.1</td>
</tr>
<tr>
<td>((F,F))</td>
<td>0.1, 0.1</td>
<td>0.9, 0.1</td>
<td>2.9, 0.9</td>
<td>0.9, 0.1</td>
</tr>
</tbody>
</table>

Table 8

The game has two Nash equilibria in pure-strategies, \(((Q,Q) ; (NF,F))\) and \(((B,B) ; (F,NF))\), that correspond to the Bayesian (and sequential) equilibria of the extensive form version. The first equilibrium requires that the type of Player 2 (column) that observes \( Q \) (and decides the first component of his strategy) believes that the probability of Player 1 (row) being of Type \( w \) (the type that decides the second letter of the strategy) is no greater than \( \frac{1}{2} \). For the second equilibrium, this probability must be no smaller than \( \frac{1}{2} \). However, intuition dictates that only \(((B,B) ; (F,NF))\) is reasonable, as Player 1 will be willing to signal her type by playing \( B \) if she happens to be the type \( s \). Indeed, we can even extend van Damme’s definition of forward induction in a way that it can be applied to this game and selects this equilibrium.

**Definition 2:** A solution concept \( S \) is consistent with forward induction in the class of generic 2-person games if \( p \in S \) for any path \( p \) for which there does not exist a player \( i \) who by unilaterally deviating from \( p \) can enforce that an information set is reached for which exactly one solution (according to \( S \)) is compatible with reaching this information set and yields this player more than \( p \) does and for which all other solutions yields this player less.

See that Definition 1 is a special case of this Definition 2, and that the only difference is that the deviating player induces, not a new subgame, but a
new information set. Now, the problem is that the deviation to reach the new equilibrium may require various types of the player to deviate. Reconsider the “bad” equilibrium in the Beer-Quiche game, \( ((Q, Q); (NF, F)) \) both \( w \) and \( s \) should deviate from \( Q \) to \( B \) to induce the “good” equilibrium. Payoffs for Player 1 are 2.9 and 2.1 in the good and bad equilibria, respectively.

One can object that, as we stated before, only \( s \) gains something in this deviation (her payoff goes up from 2 to 3, while the payoff of Type \( w \) goes down from 3 to 2). But if indeed \( s \) gains and deviates, Player 2 should understand that, if \( B \) is not played, he is facing Type \( w \), and will not play \( N \), but \( F \), and Type \( w \) will receive a payoff of 1, not 3. The consequence of this is that it pays Type \( w \) to mimic the deviation by Type \( s \) and play \( B \) as well to see his payoff increase from 1 to 2.

Indeed only \( ((B, B); (F, NF)) \) is part of the unique Kohlberg Mertens stable set of this game, which contains as elements the strategy profiles of the form

\[
((B, B); (\alpha [NF, NF] + (1 - \alpha) [F, NF]))
\]

with \( \alpha \in \{0, \frac{1}{2}\} \).

We introduce imperfect observability the following way. If Player 1 is of type \( t \), with \( t \in \{s, w\} \), Player 2 observes the signal \( \phi \in \{b, q\} \), with

\[
\Pr(\phi = b \mid B) = \Pr(\phi = q \mid Q) = 1 - \varepsilon_t.
\]

As it was the case with the previous two examples (see tables 3 and 4), with this structure, the first and fourth rows of the normal form of the Beer-Quiche game do not change (in these two columns Player 2 plays the same action regardless of what he observes, and therefore players’ payoffs are not affected by the probability of misobservation). The extensive form can be seen in Figure 3.

We will show that the strategy profile \( ((Q, Q); (NF, F)) \) is a strict equilibrium in this new game for some values of \( \varepsilon_t \). To see this notice that, in the game without noise, Player 1 is already using a strict best response, and that Player 2 would also be using a strict best response if his payoff in this strategy were higher than in \( ((Q, Q); (NF, F)) \) (in the game without noise both strategy profiles give this player a payoff of 0.9.) Since the introduction of noise would only affect the payoffs of the second and third columns by introducing terms in \( \varepsilon \), we only need to know Player 2’s payoff in \( ((Q, Q); (NF, F)) \). Simple calculations show that this payoff is 0.9 + 0.1\( \varepsilon_s \) – 0.9\( \varepsilon_w \), which is greater than 0.9 if \( \varepsilon_s > 9\varepsilon_w \). This means that, in the noise-game, \( ((Q, Q); (NF, F)) \) is a strict equilibrium if \( \varepsilon_s \) is sufficiently larger than \( \varepsilon_w \) (and both are small enough). Since any strict equilibria is also an \( KMSS \) we conclude that this

---

2 This famous game, and its Kohlberg Mertens stable sets, is analyzed in many works. For example, see Myerson (1991).
Figure 3: The Beer-Quiche game with noise
concept does not always select the “right” equilibrium when noise is added even if the “right” equilibrium was selected without noise. Using the same arguments as before, one can also check that if $\varepsilon_s \leq 9\varepsilon_w$, $((Q, Q); (NF, F))$ no longer belongs in an $KMSS$.

Finally, the only Nash equilibrium in a Kohlberg Mertens stable set in the Bear-Quiche game may not be stable when adding noise. To see this notice first that $((B, B); (F, NF))$ would be a strict equilibrium if not for strategy $((B, B); (NF, NF))$, that gives Player 2 the same payoff. Therefore, to see whether $((B, B); (F, NF))$ is still an equilibrium in the noise-game, we only need to compare Player 2’s payoffs. Calculations now show that Player 2’s payoffs in $((B, B); (NF, NF))$ and $((B, B); (F, NF))$ are, respectively, 0.9 and $0.9 + 0.1\varepsilon_w - 0.9\varepsilon_s$. Thus $((B, B); (F, NF))$ fails to be an equilibrium in the noise-game if $\varepsilon_w < 9\varepsilon_s$. Hence, the concept of stable sets of equilibria is not robust with respect to this kind of perturbations. However, if we are restricted to the case that $\varepsilon_w = \varepsilon_s$, the refinement is robust.

This last observation agrees with a result in Carlsson and Dasgupta (1997). These authors define noise equilibria for a class of signaling games that include the Bear-Quiche game. In in their model, when a message $m$ is sent, the receiver observes $m' = m + Y$, where $Y$ is a random variable. They also study a particular case of our model, namely, when $\varepsilon_t = \varepsilon_{t'}$ for all types $t$ and $t'$, and conclude that all stable sets of equilibria can be approximated by noise-proof equilibria. In particular this means that the stable set equilibria selects the same outcomes in the noise-game and in the original game. Notice that the condition in Carlsson and Dasgupta implies $\varepsilon_s = \varepsilon_w$ in our example.

4 Conclusion

There are many situations in which, by playing first, a player may significantly condition the outcome of a game. These situations include Stackelberg-like models, outside option games and signaling games. If the first mover’s action is not perfectly observed, and if players can make mistakes when choosing their actions, we obtain different results regarding the equilibrium selection. We have analyzed three games in which forward induction and the Kohlberg Mertens stable set select an equilibrium out of different possibilities. In all three examples, mistakes, no matter how small, destroy the selection. Further, the selection is robust to the addition of noise in one example, is not in a second one, and depends on the particular parameters of the noise in a third one.
References


