Handout on Evolutionary Stability in pure strategies

Ben Polak, Econ 159a/MGT522a

October 9, 2007

Suppose that strategies or behavior in games are not chosen by reasoning people, but instead are 'hard-wired' by the players' genes. Suppose further that those strategies that are relatively successful (or rather, the genes associated with those strategies and behaviors) grow while less successful strategies die out. We might want to ask what strategies will be selected by such an evolutionary process. This question has led biologists to use game theory to study animal behavior.

A related question concerns firms that compete in the market place. Perhaps the firms' policies are not chosen by sophisticated game theorists, but rather are associated with 'rules of thumb'. In this case, those firms with rules of thumb that are worse (given the rules of thumb of the other firms) might go bankrupt, leaving only a 'population' of firms with more successful rules. Such competition might mimic rational choice in that the outcome might be that only 'well-run' firms survive.

We will limit our discussion for the week to a simplified case.

- We will look only at symmetric 2-player games.
- We will assume that there is a very large population of players each of whom is 'hard-wired' to play a particular strategy. Thus the population could (potentially) involve a mix of strategies.
- The players are randomly matched into pairs.
- We look at the average payoffs attained by strategies across these pairings.
- We assume that those strategies whose average payoffs are higher than others grow relative to those others in the population mix. (Notice that this says nothing about how the population does as a whole).
- Implicitly, we are assuming only asexual reproduction. That is, we will ignore the many interesting questions that involve interaction of animals with the same genes, or involve pairings of dominant and recessive genes in each particular animal.

1. Ideas and Examples

The key idea will be the following.

Evolutionary Stability (very loose definition) Consider a large population all of whom are playing the same strategy. The strategy is called evolutionarily stable if any small mutation playing a different strategy would die out.

Example 1. Prisoners' Dilemma: strictly dominated strategies are not ES.

	cooperation	non-cooperation
cooperation	$2, 2$	$0, 3$
non-cooperation	$3, 0$	$1, 1$

Suppose everyone in the population is hard-wired to play cooperation. Now suppose that there is a small mutation hard-wired to play non-cooperation. The population mix is then $(1 - \varepsilon)$ cooperators and ε non-cooperators. Each cooperator and each non-cooperator will be randomly paired with another animal, so

each will have a $(1-\varepsilon)$-chance of being paired with cooperator and an ε-chance of being paired with a non-cooperator. The average payoffs to the incumbent cooperators is then

$$(1-\varepsilon)[2]+\varepsilon[0],$$

while the average payoff to the mutant non-cooperators is

$$(1-\varepsilon)[3]+\varepsilon[1].$$

Clearly, the non-cooperative mutants do better (on average) than the cooperative incumbents. This mutation will not die out. Thus, a population that consists 100% of cooperators is not evolutionarily stable.

Conversely, suppose everyone in the population was hard-wired to play non-cooperation. And now suppose that there is a small mutation hard-wired to play cooperation. The population mix is then $(1-\varepsilon)$ non-cooperators and ε cooperators. Each cooperator and each non-cooperator will be randomly paired with another animal, so each will have a $(1-\varepsilon)$-chance of being paired with non-cooperator and an ε-chance of being paired with a cooperator. The average payoffs to the incumbent non-cooperators is then

$$(1-\varepsilon)[1]+\varepsilon[3],$$

while the average payoff to the mutant cooperators is

$$(1-\varepsilon)[0]+\varepsilon[2].$$

Clearly, the cooperative mutants do worse (on average) than the non-cooperative incumbents. This mutation will die out. Thus, a population that consists 100% of non-cooperators is evolutionarily stable.

Thus, the first lesson of this part of the course is as follows.

Lesson *'Evolution can suck'. Evolutionary stability does not imply nice or good or efficient.*

This example also illustrates (but does not prove) a general idea.

Lesson *Strictly dominated strategies cannot be evolutionary stable.*

Try to convince yourself of this at home. [Hint: consider a strictly dominating mutation.]

Example 2. Evolutionary Stability implies Nash.

	a	b	c
a	$2,2$	$0,0$	$0,0$
b	$0,0$	$0,0$	$1,1$
c	$0,0$	$1,1$	$0,0$

Suppose everyone in the population was hard-wired to play strategy c. Strategy c is not dominated but nevertheless (c,c) is not a NE. In particular, strategy b does strictly better against c than c does against itself. This suggests that a mutation of b's will not die out if it invades the all-c population. To show this,

consider a small mutation hard-wired to play b so that the population mix is then $(1-\varepsilon)$ playing c and ε playing b. The average payoff to the incumbents hard-wired to play c is then

$$(1-\varepsilon)[0]+\varepsilon[1],$$

while the average payoff to the mutant b's is then

$$(1-\varepsilon)[1]+\varepsilon[0].$$

Since ε is small, the mutant b's will do better (on average) than the incumbent c's, and not die out. Thus, a population that consists 100% of c's is not evolutionarily stable. (Notice that the mutation here, b, is itself not evolutionarily stable – a 100% population of b's can be invaded by c's.)

This suggests the following generalization:

Lesson *If $(\hat{s},\hat{s})$ is not a NE then $\hat{s}$ is not evolutionarily stable. Or, equivalently: if $\hat{s}$ is evolutionarily stable then $(\hat{s},\hat{s})$ is a NE.*

Try to convince yourself of this at home. [Hint: consider a mutation that is hard-wired to play what would be a profitable deviation.]

In this example, it is easy to check that a is evolutionarily stable. Notice that (a,a) is also a NE. But unfortunately not all Nash strategies are evolutionarily stable.

Example 3. Nash does not imply Evolutionary Stability.

	a	b
a	$2,2$	$0,0$
b	$0,0$	$0,0$

In this example, both (a,a) and (b,b) are NE. But b is not evolutionarily stable. Consider a population in which everyone was hard-wired to play b and consider a small ε-mutation hard-wired to play a. The average payoff of the incumbent b's would then be

$$(1-\varepsilon)[0]+\varepsilon[0]$$

while the average payoff of the mutant a's would be

$$(1-\varepsilon)[0]+\varepsilon[2].$$

Clearly, the mutation does better than the incumbent and would not die out, and hence a population that consists 100% of b's is not evolutionarily stable.

Notice the reason the mutant a's did better than the incumbent b's. When matched with incumbent b's (which happened with chance $(1-\varepsilon)$) both a's and b's did equally well: both got a payoff of 0. What made the mutant a's more successful was that they did better when matched with other mutants (which happened with chance ε).

In this example, it is easy to check that a is evolutionarily stable (do so at home). What is different about the (a,a) equilibrium in this game? It is a *strict* NE. This suggests the following generalization.

Lesson *If* $(\hat{s}, \hat{s})$ *is a* strict *NE then* $\hat{s}$ *is evolutionarily stable.*

Try to convince yourself of this at home. [Hint: consider any mutation and notice that it does strictly worse whenever it is matched with an incumbent (which happens with chance $(1-\varepsilon)$).]

2. Formal definitions

The time has come to give a fairly formal definition to Evolutionary Stability.

Formal Definition 1 In a 2-player, symmetric game, the pure strategy $\hat{s}$ is *evolutionarily stable in pure strategies* if there is a (small) mutation size $\bar{\varepsilon}$ such that for all mutations of size ε smaller than $\bar{\varepsilon}$ hard-wired to play some other strategy s'

$$(1-\varepsilon)\,u(\hat{s},\hat{s}) + \varepsilon u(\hat{s},s') > (1-\varepsilon)\,u(s',\hat{s}) + \varepsilon u(s',s').$$

Look at the left of the inequality. It is the average payoff of the incumbent strategy $\hat{s}$ against the mixed population that has $(1-\varepsilon)$ incumbents hard-wired to play $\hat{s}$ and ε mutants hard-wired to play s'. To the right of the inequality is the average payoff of the mutant strategy s' against the same mix. The **strict** inequality tells us that if $\hat{s}$ is evolutionarily stable then the mutation must do strictly worse. The part about $\bar{\varepsilon}$ just says we don't care about large mutations but do care about small mutations.

Now let me provide another equivalent definition.

Formal Definition 2 In a 2-player, symmetric game, the pure strategy $\hat{s}$ is *evolutionarily stable in pure strategies* if

- **(a)** $(\hat{s},\hat{s})$ is a NE; that is., $u(\hat{s},\hat{s}) \geq u(s',\hat{s})$ for all s'; AND
- **(b)** If $(\hat{s},\hat{s})$ is **not** a strict NE (that is, there is some $s' \neq \hat{s}$ such that $u(\hat{s},\hat{s}) = u(s',\hat{s})$), then $u(\hat{s},s') > u(s',s')$.

Compare this second definition to the examples and lessons above. Part (a) says that any evolutionarily stable strategy must be Nash. Part (b) says two things. First, if a strategy is strict Nash, then there is nothing else to check: it is evolutionarily stable. But if a strategy is Nash but not strict Nash, then (and only then) we need to check a second condition. The second condition says: if the mutation does as well against the incumbent as the incumbent does against itself, then to be evolutionarily stable the incumbent must do **strictly** better against the mutant than the mutant does against itself. This was the condition that strategy b failed in the third example above.

There are two reasons why the second definition is interesting. First, as we shall discover, in purely practical terms, it is much easier to check than the first definition. Second, on more intellectual terms, it is a remarkable fact that a key concept from modern economics, *Nash equilibrium*, should be so closely related to a key concept from modern biology, *evolutionary stability*. For nerds like me, there is something almost awe-inspiring about this coincidence.

For this course, you **do** need to know the two definitions above but you do not need to know the proof. It is just for nerds.

Sketch of a proof. We can rewrite the inequality in definition 1 as follows

$$(1-\varepsilon)\left[u\left(\hat{s},\hat{s}\right)-u\left(s',\hat{s}\right)\right]+\varepsilon\left[u\left(\hat{s},s'\right)-u\left(s',s'\right)\right]>0.$$

The first [term] compares the payoff of the incumbent and the mutant against the incumbent. The second [term] compares the payoff of the incumbent and the mutant against the mutant. The first term has weight $(1-\varepsilon)$ since this is the chance of being paired with an incumbent. The second term has weight ε since this is the chance of being paired with a mutant. Since the inequality must hold for all ε (smaller than some $\bar{\varepsilon}$), if the first term is strictly negative we are in trouble. By choosing ε arbitrarily small, we can make the weight of the second term arbitrarily small, and hence make the entire left side negative. Thus we need the first term to be weakly bigger than zero. This is exactly what part (a) of definition 2 says. Conversely, if the first term is strictly positive, we are done. The first definition allows us to choose ε (or more formally $\bar{\varepsilon}$) to be as small as we like, so we can choose the weight ε on the second term arbitrarily small and ensure that the entire left side is positive. Thus if the first term is strictly positive for all s' then $\hat{s}$ is evolutionary stable. This is exactly saying that strict NE is sufficient. Finally, if (and only if) the first term is exactly zero, then we have to look at the second term to check if $\hat{s}$ is evolutionarily stable. But this is exactly what part (b) of the second definition says.