Large Population Potential Games*

William H. Sandholm†

January 8, 2009

Abstract

We offer a parsimonious definition of large population potential games, provide some alternate characterizations, and demonstrate the advantages of the new definition over the existing definition, but also show the equivalence of the two definitions.

1. Introduction

Potential games admit a variety of practical applications and possess appealing theoretical properties, most notably the convergence of myopic adjustment processes to Nash equilibrium. Sandholm (2001), building on the work of Beckmann et al. (1956), Monderer and Shapley (1996), and Hofbauer and Sigmund (1988), defines and analyzes potential games in settings with large populations of anonymous agents.

In this paper, we provide a new definition of potential games for these large population settings. At first glance it may seem that the definition proposed here is more general but more complicated than the one from Sandholm (2001); however, we show that if population sizes are fixed, the two definitions are equivalent. Equally importantly, the new definition is both more parsimonious and easier to verify than the old one: while under the old definition, checking whether a game is a potential game requires constructing extended payoff functions that satisfy certain properties, the new definition can be verified directly.

While our analysis allows for multiple populations (i.e., multiple player roles), it is easier to describe the main ideas informally by focusing on the single population case. In this setting, an n-strategy population game is defined by a vector of payoff functions $F = (F_1, \ldots, F_n)$ defined on the set of population states—that is, the simplex X in $\mathbf{R}^n$.

*I thank an anonymous referee, an anonymous Associate Editor, and the Editors for helpful comments, and Katsuhiko Aiba for able research assistance. I also gratefully acknowledge financial support from NSF Grants SES-0092145 and SES-0617753.

†Department of Economics, University of Wisconsin, 1180 Observatory Drive, Madison, WI 53706, USA. e-mail: whs@ssc.wisc.edu; website: http://www.ssc.wisc.edu/~whs.

In Sandholm (2001), potential games are defined using standard concepts from multivariable calculus. A population game F is called a potential game if there is a scalar-valued function f, the potential function, whose gradient always equals the vector of payoffs; put differently, the payoff to strategy i must always be given by the ith partial derivative of f. By standard results, F is a potential game in this sense if and only if its derivative matrices $DF(x)$ are symmetric, so that corresponding cross partial derivatives of F are equal.

Because the statements above involve partial derivatives, they only make sense if the relevant functions are defined on a full-dimensional subset of $\mathbf{R}^n$—for instance, on the positive orthant $\mathbf{R}^n_+$. In particular, checking the symmetry of the derivative matrix $DF(x)$ requires us to extend the payoff functions from the simplex X to the orthant $\mathbf{R}^n_+$. Since points outside the simplex are not population states, extending payoffs in this way is not parsimonious. Worse still, this approach makes it difficult to decide whether a given population game is a potential game. Even if the extensions one has checked do not satisfy the symmetry condition, an extension that one has not checked may do so. Simple examples in Sections 3 and 4 show that this possibility is real.

We therefore offer a definition of potential games that does not require such extensions: payoffs and potential need only be defined on the simplex. Later on, in Theorem 4.3, we characterize potential games using a derivative symmetry condition, but one that continues to make sense when payoffs are only defined on the simplex.

From a game theoretic point of view, the key difference between the new and old definitions is this: While under the old definition, the gradient ∇f completely describes payoffs under F, under the new definition, the gradient need only describe relative payoffs under F—in other words, differences between the payoffs of pairs of strategies. In this sense, the new definition accords more closely with that of Monderer and Shapley (1996), as their normal form potential games are defined by the property that differences in payoffs are captured by differences in potential.

The new definition and characterization employ the orthogonal projection Φ. When applied to a payoff vector π, Φ returns the demeaned version of that vector, so that $\Phi\pi$ only retains information about relative payoffs. Using Φ in the definition of potential games is an uncomplicated way of ensuring that only relative payoffs are regulated by the definition.

What is the link between the old and new definitions of large population potential games? Surprisingly, Theorem 4.1 shows that the two definitions are equivalent: given a payoff function F and potential function f defined on the simplex, there is always an extension of these functions to $\mathbf{R}^n_+$ that satisfies the definition of potential games from Sandholm (2001). In particular, the information about payoffs not initially captured by f—namely, information about average payoffs—can always be incorporated into the extension of f. This result tells us that the old definition of potential games does not entail a loss of generality. But the new approach is superior: since it does not require extensions, it allows us to decide whether a population game is a potential game in a systematic way.

Section 2 introduces population games. Section 3 reviews the definition of potential games from Sandholm (2001), explains the inadequacies of this definition, and proposes a new one. Section 4 relates the old and new definitions, and also describes situations in which the old definition is more useful than the new one.

2. Games Played by Large Populations

2.1 Population Games

Let $\mathcal{P} = \{1, ..., p\}$ be a *society* consisting of $p \geq 1$ *populations* of agents. Agents in population p form a continuum of *mass* $m^p > 0$.

The set of *strategies* available to agents in population p is denoted $S^p = \{1, \dots, n^p\}$, and has typical elements i, j, and (in the context of normal form games) s^p. We let $n = \sum_{p \in \mathcal{P}} n^p$ equal the total number of pure strategies in all populations.

During game play, each agent in population p selects a strategy from S^p. The set of *population states* (or *strategy distributions*) for population p is thus $X^p = \{x^p \in \mathbf{R}^{n^p}_+ : \sum_{i \in S^p} x^p_i = m^p\}$. The scalar $x^p_i \in \mathbf{R}_+$ represents the mass of players in population p choosing strategy $i \in S^p$. Elements of $X = \prod_{p \in \mathcal{P}} X^p = \{x = (x^1, \dots, x^p) \in \mathbf{R}^n_+ : x^p \in X^p\}$, the set of *social states*, describe behavior in all p populations at once.

We generally take the sets of populations and strategies as fixed and identify a game with its payoff function. A *payoff function* F: $X \to \mathbf{R}^n$ is a continuous map that assigns each social state a vector of payoffs, one for each strategy in each population. F^p_i: $X \to \mathbf{R}$ denotes the payoff function for strategy $i \in S^p$, while F^p: $X \to \mathbf{R}^{n^p}$ denotes the payoff functions for all strategies in S^p. When $p = 1$, we omit the redundant superscript p from all of our notation.

Example 2.1. Random matching in normal form games. A ***p**-player normal form game* is defined by a strategy set $S^p = \{1, \dots, n^p\}$ and a utility function $U^p : S \to \mathbf{R}$ for each player $p \in \mathcal{P} = \{1, ..., p\}$, $p \geq 2$. The domain of U^p is the set of pure strategy profiles $S = \prod_{q \in \mathcal{P}} S^q$.

Suppose that agents from p unit-mass populations are randomly matched to play the normal form game $U = (U^1, \dots, U^p)$, with one agent from each population q being drawn to serve in player role q. Then the probability that an agent from population p is matched with opponents who jointly play strategy profile $s^{-p} \in S^{-p} = \prod_{r \neq p} S^r$ is $\prod_{r \neq p} x^r_{s^r}$. The population game defined by this random matching procedure therefore has the multilinear payoff function

$$(1) \qquad F^p_{s^p}(x) = \sum_{s^{-p} \in S^{-p}} U^p(s^1, \ldots, s^p) \prod_{r \neq p} x^r_{s^r} \text{ for all } x \in X. \; \S$$

2.2 Full Population Games

In searching for interesting properties of population games, it seems natural to consider the marginal effect of adding new agents choosing strategy $j \in S^q$ on the payoffs of agents currently choosing strategy $i \in S^p$. This effect is captured by the partial derivative $\frac{\partial F^p_i}{\partial x^q_j}$. But if F^p_i is only defined on X, the payoffs that would arise were newcomers to enter the population are not defined. For this reason, even if F^p_i is differentiable (in a sense to be made explicit below), its partial derivatives do not exist.

To ensure that partial derivatives exist, we can extend the domain of payoff functions from the state space $X = \{x = (x^1, \ldots, x^p) \in \mathbf{R}^n_+ : \sum_{i \in S^p} x^p_i = m^p\}$ to the entire positive orthant $\mathbf{R}^n_+$. We call the resulting game $\tilde{F} : \mathbf{R}^n_+ \to \mathbf{R}^n$ a *full-dimensional population game*, or simply a *full population game*.

As suggested above, one can sometimes interpret this extended definition of payoffs as describing the values that payoffs would take were the population masses to change. For instance, this interpretation is the natural one in congestion games: here it makes sense to consider the effects of changing the number of drivers commuting between a certain origin/destination pair, and it is clear how one ought to define the payoffs at any newly feasible state $x \in \mathbf{R}^n_+ - X$.[1] But in other settings, there may be no one choice of extension that is obviously "correct" from a game-theoretic point of view.

Example 2.2. Random matching in normal form games revisited. Suppose we would like to define a full population game version of the random matching game (1). Many extensions of definition (1) from X to $\mathbf{R}^n_+$ are equally good in principle. But of these, the easiest to write down is the extension that retains the multilinear functional form from (1):

$$(2) \qquad \tilde{F}^p_{s^p}(x) = \sum_{s^{-p} \in S^{-p}} U^p(s^1, \ldots, s^p) \prod_{r \neq p} x^r_{s^r} \text{ for all } x \in \mathbf{R}^n_+. \; \S$$

[1] In particular, the cost of choosing path i should equal the sum of the delays on the links in the path, where the delay on a link is described by a nondecreasing function of the number of drivers on that link—see Beckmann et al. (1956) or Sandholm (2001).

3. Large Population Potential Games

3.1 Full Potential Games

We can now present a definition of potential games used in Sandholm (2001).[2]

Definition. *We call the full population game* $\tilde{F} : \mathbf{R}^n_+ \to \mathbf{R}^n$ *a* full potential game *if there exists a continuously differentiable function* $\tilde{f} : \mathbf{R}^n_+ \to \mathbf{R}$ *satisfying*

$$\text{(FP)} \qquad \nabla \tilde{f}(x) = \tilde{F}(x) \ \textit{ for all } x \in \mathbf{R}^n_+.$$

Property (FP) can be stated more explicitly as

$$\frac{\partial \tilde{f}}{\partial x^p_i}(x) = \tilde{F}^p_i(x) \ \text{ for all } p \in \mathcal{P}, i \in S^p \text{ and } x \in \mathbf{R}^n_+.$$

We call the function $\tilde{f}$, which is unique up to an additive constant, the *full potential function* for the game $\tilde{F}$.

That payoffs are fully captured by the scalar-valued function $\tilde{f}$ is the source of full potential games' attractive properties. Nash equilibria of $\tilde{F}$ can be characterized in terms of Kuhn-Tucker conditions for maximizers of $\tilde{f}$. Moreover, reasonable evolutionary dynamics—namely, dynamics in which strategies' growth rates are positively correlated with their payoffs—ascend potential, and therefore converge to sets of rest points from all initial conditions. Since under many dynamics the rest points are identical to the Nash equilibria of $\tilde{F}$, this global convergence result provides a strong justification of the prediction of Nash equilibrium play.

Suppose that $\tilde{F}$ is continuously differentiable (C^1). Then by the standard integrability condition from multivariable calculus (see, e.g., Lang (1997)), $\tilde{F}$ is a full potential game if and only if it satisfies *full externality symmetry*:

$$\text{(FES)} \qquad D\tilde{F}(x) \text{ is symmetric for all } x \in \mathbf{R}^n_+.$$

More explicitly, $\tilde{F}$ is a full potential game if and only if

$$\text{(3)} \qquad \frac{\partial \tilde{F}^p_i}{\partial x^q_j}(x) = \frac{\partial \tilde{F}^q_j}{\partial x^p_i}(x) \ \text{ for all } i \in S^p, j \in S^q, \ p, q \in \mathcal{P}, \text{ and } x \in \mathbf{R}^n_+.$$

[2]Actually, while we suppose here that payoffs and potential are defined on $\mathbf{R}^n_+$, in Sandholm (2001) we assumed that these were defined on the smaller set $\bar{X} = \prod_{p \in \mathcal{P}} \bar{X}^p \subset \mathbf{R}^n_+$, where $\bar{X}^p = \{x^p \in \mathbf{R}^{n^p}_+ : \sum_{i \in S^p} x^p_i \in [m^p - \varepsilon, m^p + \varepsilon]\}$ for some $\varepsilon > 0$. As both of these domains are full-dimensional, they accomplish the same purpose.

Thus, full externality symmetry has a natural game-theoretic interpretation: the effect on the payoffs to strategy $i \in S^p$ of introducing new agents choosing strategy $j \in S^q$ must always equal the effect on the payoffs to strategy j of introducing new agents choosing strategy i.

3.2 An Example

The difficulties that the definition of full potential games allows are illustrated in the following example.

Example 3.1. Random matching in normal form potential games. The normal form game U is a *potential game* (Monderer and Shapley (1996)) if there is a *potential function* $V : S \to \mathbf{R}$ such that

$$(4) \qquad U^p(\hat{s}^p, s^{-p}) - U^p(s^p, s^{-p}) = V(\hat{s}^p, s^{-p}) - V(s^p, s^{-p}) \text{ for all } \hat{s}^p, s^p \in S^p, s^{-p} \in S^{-p}, \text{ and } p \in \mathcal{P}.$$

In words, U is a potential game if any unilateral deviation has the same effect on the deviator's payoffs as it has on potential. Equivalently, U is a potential game if there is a potential function V and auxiliary functions $W^p : S^{-p} \to \mathbf{R}$ such that

$$(5) \qquad U^p(s) = V(s) + W^p(s^{-p}) \text{ for all } s \in S \text{ and } p \in \mathcal{P},$$

so that each player's payoff is the sum of a common payoff term and a term that only depends on opponents' behavior.

Suppose that p populations of agents are randomly matched to play the normal form potential game U . The full population game we obtain from equation (2) is

$$(6) \qquad \tilde{F}^p_{s^p}(x) = \sum_{s^{-p} \in S^{-p}} (V(s) + W^p(s^{-p})) \prod_{r \neq p} x^r_{s^r} \text{ for all } x \in \mathbf{R}^n_+,$$

where V and W^p are the potential and auxiliary functions from equation (5).

Since the normal form game U admits a potential function, it seems natural to expect $\tilde{F}$ to admit one as well. But if we compute the partial derivatives of payoffs, we find that

$$\frac{\partial \tilde{F}^p_{s^p}}{\partial x^q_{s^q}}(x) = \sum_{s^{-\{p,q\}} \in S^{-\{p,q\}}} \left(V(s) + W^p(s^q, s^{-\{p,q\}})\right) \prod_{r \neq p,q} x^r_{s^r}, \text{ while}$$
$$\frac{\partial \tilde{F}^q_{s^q}}{\partial x^p_{s^p}}(x) = \sum_{s^{-\{p,q\}} \in S^{-\{p,q\}}} \left(V(s) + W^q(s^p, s^{-\{p,q\}})\right) \prod_{r \neq p,q} x^r_{s^r}.$$

Thus, without further assumptions about W^p and W^q, $\tilde{F}$ is not a full potential game.

If each W^p is identically zero, so that U is a *common interest game*, then the partial derivatives above are equal, and so $\tilde{F}$ is a full potential game. In this case, if we let $\tilde{f}$: $\mathbf{R}^n_+ \to \mathbf{R}$ equal aggregate payoffs in $\tilde{F}$ divided by p,

$$\tilde{f}(x) = \frac{1}{p} \sum_{p \in \mathcal{P}} \sum_{s^p \in S^p} x^p_{s^p} \tilde{F}^p_{s^p}(x) = \sum_{s \in S} V(s) \prod_{r \in \mathcal{P}} x^r_{s^r}, \tag{7}$$

we find that

$$\frac{\partial \tilde{f}}{\partial x^p_{s^p}}(x) = \sum_{s^{-p} \in S^{-p}} V(s) \prod_{r \neq p} x^r_{s^r} = \tilde{F}^p_{s^p}(x).$$

Thus, $\nabla \tilde{f}(x) = \tilde{F}(x)$ for all $x \in \mathbf{R}^n_+$, and so $\tilde{f}$ is a full potential function for $\tilde{F}$. §

This example seems to suggest that random matching in a normal form potential game does not generate a full potential game. But we will soon see that the example is misleading. While equation (6) is the only way to define the payoffs of the random matching game on the state space X, equation (6) implicitly uses extension (2) to define payoffs on the rest of $\mathbf{R}^n_+$. Theorem 4.1 shows that by choosing the extension more carefully, one can ensure that conditions (FP) and (FES) are satisfied.

More fundamentally, Example 3.1 illustrates the difficulties inherent in definitions that require the existence of extensions satisfying certain properties. If we only know how payoffs must be defined on the state space X, it is not obvious how to determine whether an extension to $\mathbf{R}^n_+$ satisfying conditions like (FP) and (FES) exists: the fact that we have not found an extension satisfying these properties does not necessarily mean that no such extension exists.

The most parsimonious approach to this problem is to work directly with payoff and potential functions defined on the state space X. Since in this case partial derivatives of payoffs and potential are not defined, the usual definitions from multivariate calculus do not apply, and so must be replaced by notions of differentiation for functions defined on affine spaces. In geometric terms, this means that derivatives can only be evaluated with respect to directions tangent to X, which represent feasible changes in the social state. As we will see, the game theoretic consequence of this restriction is that potential functions defined on X only provide information about relative payoff levels.

Once this parsimonious analysis has been carried out, it is not difficult to introduce alternative definitions that use payoff extensions, and thus avoid affine calculus, but that do not cause the difficulties illustrated in the example above. We show how this is done in Section 3.4 and Corollary 4.6 below.

3.3 Derivatives and Tangent Spaces

We now introduce the definitions we need to proceed with the analysis proposed above. For a more thorough treatment of these ideas, see Akin (1990).

The *tangent space* TX of the state space X is the smallest subspace of $\mathbf{R}^n$ containing all directions of motion through X. One can verify that $TX = \prod_{p \in \mathcal{P}} TX^p$, where the set $TX^p = \{z^p \in \mathbf{R}^{n^p} : \sum_{i \in S^p} z_i^p = 0\}$ is the tangent space of X^p.

Let $f : X \to \mathbf{R}$ be a scalar-valued function on X. Then f is *differentiable* at $x \in X$ if there is a linear map $Df(x) : TX \to \mathbf{R}$, the *derivative* of f at x, satisfying

$$f(x+z) = f(x) + Df(x)z + o(z) \quad \text{for all } z \in TX.$$

The *gradient* of f at x is the unique vector $\nabla f(x) \in TX$ such that $Df(x)z = \nabla f(x)'z$ for all $z \in TX$. It is worth emphasizing that the uniqueness requirement here is with respect to the set TX. More generally, if $v \in \mathbf{R}^n$ is any vector that represents $Df(x)$ in the sense above, then $\nabla f(x) = \mathbf{\Phi} v$, where the matrix $\mathbf{\Phi} \in \mathbf{R}^{n \times n}$ is the orthogonal projection of $\mathbf{R}^n$ onto TX defined below.

Analogous definitions apply to vector-valued functions on X. If $F : X \to \mathbf{R}^n$ is C^1, its derivative at $x \in X$ is a linear map from TX to $\mathbf{R}^n$; while many matrices in $\mathbf{R}^{n \times n}$ can represent this derivative, applying the logic above to each component of F shows that there is a unique such matrix, the *derivative matrix* $DF(x)$, whose rows are in TX.

A leading role in our analysis is played by the orthogonal projection onto TX^p, represented by the symmetric matrix $\Phi \in \mathbf{R}^{n^p \times n^p}$. This matrix can be written explicitly as

$$\Phi = I - \tfrac{1}{n^p}\mathbf{1}\mathbf{1}',$$

where I is the identity matrix and $\mathbf{1} \in \mathbf{R}^{n^p}$ is the vector of ones. Since $TX = \prod_{p \in \mathcal{P}} TX^p$, it follows that the orthogonal projection onto TX is given by the block diagonal matrix $\mathbf{\Phi} = \text{diag}(\Phi, \ldots, \Phi) \in \mathbf{R}^{n \times n}$.

If $\pi^p \in \mathbf{R}^{n^p}$ is a payoff vector for population p, then the projected payoff vector

$$\Phi \pi^p = \pi^p - \mathbf{1} \cdot \tfrac{1}{n^p} \sum_{i \in S^p} \pi_i^p \equiv \pi^p - \mathbf{1}\, \bar{\pi}^p$$

subtracts the average payoff $\bar{\pi}^p$ from each component of π^p. In other words, Φ preserves the differences between components of π^p while normalizing their sum to zero. The residual from the projection, $(I - \Phi)\pi^p = \mathbf{1}\,\bar{\pi}^p$, is a constant vector whose components all equal $\bar{\pi}^p$, the average payoff of strategies belonging to population p. By the same logic, if $\pi = (\pi^1, \ldots, \pi^p) \in \mathbf{R}^n$ is a payoff vector for the society, then $\mathbf{\Phi}\pi = (\Phi\pi^1, \ldots, \Phi\pi^p)$ normalizes each of the p pieces of the vector π separately, and $(I - \mathbf{\Phi})\pi = (\mathbf{1}\bar{\pi}^1, \ldots, \mathbf{1}\bar{\pi}^p)$ describes average payoffs in each population.

3.4 Potential Games

With these preliminaries in hand, we present our new definition.

Definition. *Let $F\colon X \to \mathbf{R}^n$ be a population game. We call F a* potential game *if it admits a* potential function*: a C^1 function $f : X \to \mathbf{R}$ that satisfies*

$$\text{(P)} \qquad \nabla f(x) = \mathbf{\Phi} F(x) \quad \textit{for all } x \in X.$$

Since the potential function f has domain X, the gradient vector $\nabla f(x)$ is by definition an element of the tangent space TX. Condition (P) requires that this gradient vector always equal $\mathbf{\Phi}F(x)$, the projection of the payoff vector $F(x)$ onto TX. Put differently, the gradient of the potential function f must capture relative payoffs in F.

At the cost of sacrificing parsimony, one can avoid affine calculus by using a function defined throughout $\mathbf{R}^n_+$ to play the role of the potential function f. To do so, one simply includes the projection $\mathbf{\Phi}$ on both sides of the analogue of equation (P).

Observation 3.2. *If F is a potential game with potential function $f : X \to \mathbf{R}$, then any C^1 extension $\tilde{f} : \mathbf{R}^n_+ \to \mathbf{R}$ of f satisfies*

$$(8) \qquad \mathbf{\Phi}\nabla \tilde{f}(x) = \mathbf{\Phi} F(x) \quad \textit{for all } x \in X.$$

Conversely, if the population game F admits a function $\tilde{f}$ satisfying condition (8)*, then F is a potential game, and the restriction $f = \tilde{f}\big|_X$ is a potential function for F.*

This observation is immediate from the relevant definitions. In particular, if $\tilde{f}$ and f agree on X, then for all $x \in X$ the gradient vectors $\nabla\tilde{f}(x)$ and $\nabla f(x)$ define identical linear operators on TX, implying that $\mathbf{\Phi}\nabla\tilde{f}(x) = \mathbf{\Phi}\nabla f(x)$. But since $\mathbf{\Phi}\nabla f(x) = \nabla f(x)$ by definition, it follows that $\mathbf{\Phi}\nabla\tilde{f}(x) = \nabla f(x)$; this equality and definition (P) yield the result.

Observation 3.2 can be restated in a way that uses simple differences rather than the orthogonal projection $\mathbf{\Phi}$.[3] Note that condition (8) is equivalent to the dual condition

$$z'\nabla\tilde{f}(x) = z'F(x) \quad \text{for all } z \in TX \text{ and } x \in X.$$

[3] I thank an anonymous referee for suggesting condition (9).

Since the set of vectors of the form $e^p_j - e^p_i$ span the subspace TX, this condition is itself equivalent to

$$\frac{\partial \tilde{f}}{\partial x^p_j}(x) - \frac{\partial \tilde{f}}{\partial x^p_i}(x) = F^p_j(x) - F^p_i(x) \text{ for all } p \in \mathcal{P}, i, j \in S^p \text{ and } x \in X. \tag{9}$$

We therefore have

Proposition 3.3. *Observation 3.2 remains true if condition* (8) *is replaced by condition* (9).

For an application of this result, see Section 3.5 below.

We conclude with a representation of potential games and potential functions that uses multipopulation versions of conditions due to Hofbauer and Sigmund (1988, p. 243). Let $\mathcal{X}^p = \{\chi^p \in \mathbf{R}^{n^p-1}_+ : \sum_{i=1}^{n^p-1} \chi^p_i \leq m^p\}$ denote the projection of $X^p \subset \mathbf{R}^{n^p}$ onto $\mathbf{R}^{n^p-1}$, and let $\mathcal{X} = \prod_{p\in\mathcal{P}} \mathcal{X}^p$. Then $\psi(\chi^p) = (\chi^p_1, \ldots, \chi^p_{n^p-1}, m^p - \sum_{i=1}^{n^p-1} \chi^p_i)$ is a bijection from $\mathcal{X}^p$ to X^p, and $\psi(\chi) = (\psi^1(\chi^1), \ldots, \psi^p(\chi^p))$ is a bijection from $\mathcal{X}$ to X.

Proposition 3.4. *Let $F : X \to \mathbf{R}^n$ be a population game, and define the function $G : \mathcal{X} \to \mathbf{R}^{n-p}$ by*

$$G^p_i(\chi) = F^p_i(\psi(\chi)) - F^p_{n^p}(\psi(\chi)) \text{ for all } p \in \mathcal{P}, i < n^p \text{ and } \chi \in \mathcal{X}. \tag{10}$$

Then F is a potential game with potential function $f : X \to \mathbf{R}$ if and only if there exists a function $g : \mathcal{X} \to \mathbf{R}$ such that

$$\frac{\partial g}{\partial \chi^p_i}(\chi) = G^p_i(\chi) \text{ for all } p \in \mathcal{P}, i < n^p \text{ and } \chi \in \mathcal{X}. \tag{11}$$

When these statements are true, f and g are related by $g(\chi) = f(\psi(\chi)) + c$.

Proof. Suppose first that F is a potential game with potential function f. Then defining $g : \mathcal{X} \to \mathbf{R}$ by $g(\chi) = f(\psi(\chi))$, and letting e^p_i denote the (i, p)th standard basis vector in $\mathbf{R}^n$, we find that

$$\frac{\partial g}{\partial \chi^p_i}(\chi) = \frac{\partial f}{\partial (e^p_i - e^p_{n^p})}(\psi(\chi)) = F^p_i(\psi(\chi)) - F^p_{n^p}(\psi(\chi)) = G^p_i(\chi),$$

so equation (11) holds.

Conversely, suppose that G and g satisfy equation (11). If we define $f : X \to \mathbf{R}$ by

$f(x) = g(\psi^{-1}(x))$, then since $\psi^{-1}(x) = (\ldots, x_1^p, \ldots, x_{n^p-1}^p, x_1^{p+1}, \ldots)$ it follows that

$$\frac{\partial f}{\partial(e_i^p - e_{n^p}^p)}(x) = \frac{\partial g}{\partial \chi_i^p}(\psi^{-1}(x)) = G_i^p(\psi^{-1}(x)) = F_i^p(x) - F_{n^p}^p(x).$$

As vectors of the form $e_i^p - e_{n^p}^p$ span TX, we conclude that F is a potential game with potential function f. ■

3.5 Example Revisited

For a first application of our new definition, we revisit random matching in normal form potential games (Example 3.2).[4]

Let U be a normal form potential game with potential function V and auxiliary functions W^p, and let $F : X \to \mathbf{R}^n$ be the corresponding population game:

$$U^p(s) = V(s) + W^p(s^{-p});$$
$$F_{s^p}^p(x) = \sum_{s^{-p} \in S^{-p}} (V(s) + W^p(s^{-p})) \prod_{r \neq p} x_{s^r}^r$$

This F is the restriction to X of the full potential game $\tilde{F} : \mathbf{R}_+^n \to \mathbf{R}^n$ from equation (6). If we define the function $\tilde{f} : \mathbf{R}_+^n \to \mathbf{R}$ as in equation (7),

$$\tilde{f}(x) = \sum_{s \in S} V(s) \prod_{r \in \mathcal{P}} x_{s^r}^r,$$

we find that

$$\frac{\partial \tilde{f}}{\partial x_{\hat{s}^p}^p}(x) - \frac{\partial \tilde{f}}{\partial x_{s^p}^p}(x) = \sum_{s^{-p} \in S^{-p}} \Big(V(\hat{s}^p, s^{-p}) - V(s^p, s^{-p}) \Big) \prod_{r \neq p} x_{s^r}^r$$
$$= F_{\hat{s}^p}^p(x) - F_{s^p}^p(x).$$

Therefore, by Proposition 3.3, F is a potential game with potential function $f = \tilde{f}\big|_X$.

Conversely, let U be a normal form game, and let $F : X \to \mathbf{R}$ be the corresponding population game as defined in equation (1). Since the payoffs to an agent in population p do not depend on the choices of other agents in population p, we can write $F^p(x) = F^p(x^{-p})$ whenever it is convenient to do so.

[4] I thank an anonymous referee for suggesting the arguments to follow, which improve significantly on my initial proof.

Suppose that F is a potential game with potential function $f : X \to \mathbf{R}$ as defined in equation (P). Fix a player p and two strategies $\hat{s}^p, s^p \in S^p$, and let e^p_i denote the (i, p)th standard basis vector in $\mathbf{R}^n$. Since $e^p_{\hat{s}^p} - e^p_{s^p} \in TX$, definition (P) implies that

$$(e^p_{\hat{s}^p} - e^p_{s^p})'\nabla f(x) = (e^p_{\hat{s}^p} - e^p_{s^p})'F(x) = F^p_{\hat{s}^p}(x) - F^p_{s^p}(x) = F^p_{\hat{s}^p}(x^{-p}) - F^p_{s^p}(x^{-p})$$

for all $x \in X$. Thus, letting ι^p_i denote the ith standard basis vector in $\mathbf{R}^{n^p}$, and setting $y^p(t) = t\,\iota^p_{\hat{s}^p} + (1-t)\,\iota^p_{s^p}$, we have that

$$f(\iota^p_{\hat{s}^p}, x^{-p}) - f(\iota^p_{s^p}, x^{-p}) = \int_0^1 \nabla f(y^p(t), x^{-p})'(e^p_{\hat{s}^p} - e^p_{s^p})\,\mathrm{d}t = F^p_{\hat{s}^p}(x^{-p}) - F^p_{s^p}(x^{-p}).$$

Now define $V : S \to \mathbf{R}$ by $V(s) = f(\iota^1_{s^1}, \ldots, \iota^p_{s^p})$. Fix a strategy $s^q \in S^q$ for each player $q \neq p$; then setting $y^q = \iota^q_{s^q}$ for each such q, we find that

$$\begin{aligned} U^p(\hat{s}^p, s^{-p}) - U^p(s^p, s^{-p}) &= F^p_{\hat{s}^p}(y^{-p}) - F^p_{s^p}(y^{-p}) \\ &= f(\iota^p_{\hat{s}^p}, y^{-p}) - f(\iota^p_{s^p}, y^{-p}) \\ &= V(\hat{s}^p, s^{-p}) - V(s^p, s^{-p}). \end{aligned}$$

Therefore, V is a potential function for U as defined in equation (4).

In summary, we have established

Proposition 3.5. *Let U be a normal form game, and let $F : X \to \mathbf{R}^n$ be the population game generated by random matching in U. Then F is a potential game if and only if U is a potential game.*

4. Characterizations of Large Population Potential Games

4.1 Creating Full Potential Games from Potential Games

What is the relationship between full potential games and potential games? Under the former definition, condition (FP) requires that payoffs be completely determined by the potential function, which is defined on $\mathbf{R}^n_+$. Under the latter, condition (P) asks only that relative payoffs be determined by the potential function, now defined just on X.

To understand the relationship between the two definitions, take a potential game $F : X \to \mathbf{R}^n$ with potential function $f : X \to \mathbf{R}$ as given, and extend f to a full potential function $\tilde{f} : \mathbf{R}^n_+ \to \mathbf{R}$. Theorem 4.1 shows that the link between the full potential game $\tilde{F} \equiv \nabla \tilde{f}$ and the original game F depends on how the extension $\tilde{f}$ is chosen.

Theorem 4.1. *Let $F : X \to \mathbf{R}^n$ be a potential game with potential function $f : X \to \mathbf{R}$. Let $\tilde{f} : \mathbf{R}^n_+ \to \mathbf{R}$ be any C^1 extension of f, and define the full potential game $\tilde{F} : \mathbf{R}^n_+ \to \mathbf{R}^n$ by $\tilde{F}(x) = \nabla \tilde{f}(x)$. Then*

(*i*) *The population games F and $\tilde{F}\big|_X$ have the same relative payoffs: $\mathbf{\Phi} F(x) = \mathbf{\Phi}\tilde{F}(x)$ for all $x \in X$.*

(*ii*) *One can choose the extension $\tilde{f}$ in such a way that F and $\tilde{F}\big|_X$ are identical.*

Part (*i*) of the theorem shows that the full potential game $\tilde{F}$ generated from an arbitrary extension of the potential function f exhibits the same relative payoffs as F on their common domain X. It follows that F and $\tilde{F}$ have the same best response correspondences and Nash equilibria, but may exhibit different average payoff levels. Part (*ii*) of the theorem shows that by choosing the extension $\tilde{f}$ appropriately, we can make $\tilde{F}$ and F identical on X. To accomplish this, we construct the extension $\tilde{f}$ in such a way (equation (12) below) that its derivatives at states in X evaluated in directions orthogonal to TX encode information about average payoffs from the original game F.

In conclusion, Theorem 4.1(*ii*) demonstrates that if population masses are fixed, so that the relevant set of social states is X, then definition (FP), while more difficult to check, does not entail a loss of generality relative to definition (P).

Proof of Theorem 4.1. Part (*i*) follows from the fact that $\mathbf{\Phi}\tilde{F}(x) = \mathbf{\Phi}\nabla\tilde{f}(x) = \nabla f(x) = \mathbf{\Phi}F(x)$ for all $x \in X$; compare the discussion following Observation 3.2.

To prove part (*ii*), we first extend f and F from the state space X to its affine hull $\text{aff}(X) = \{y = (y^1, \ldots, y^p) \in \mathbf{R}^n : \sum_{i \in S^p} y^p_i = m^p\}$. Let $\hat{f} : \text{aff}(X) \to \mathbf{R}$ be a C^1 extension of $f : X \to \mathbf{R}$, and let $\hat{g}^p : \text{aff}(X) \to \mathbf{R}$ be a continuous extension of population p's average payoff function, $\frac{1}{n^p}\mathbf{1}'F^p : X \to \mathbf{R}$. (The existence of these extensions follows from the Whitney Extension Theorem—see Krantz and Parks (1999).) Then define $\hat{G} : \text{aff}(X) \to \mathbf{R}^n$ by $\hat{G}^p(x) = \mathbf{1}\hat{g}^p(x)$, so that $F(x) = \mathbf{\Phi}F(x) + (I - \mathbf{\Phi})F(x) = \nabla\hat{f}(x) + \hat{G}(x)$ for all $x \in X$. If after this we define $\hat{F} : \text{aff}(X) \to \mathbf{R}^n$ by $\hat{F}(x) = \nabla\hat{f}(x) + \hat{G}(x)$, then $\hat{F}$ is a continuous extension of F, and $\nabla\hat{f}(x) = \mathbf{\Phi}\hat{F}(x)$ for all $x \in \text{aff}(X)$.

With this groundwork complete, we can extend f to all of $\mathbf{R}^n_+$ via

$$\tilde{f}(y) = \hat{f}(\xi(y)) + (y - \xi(y))'\hat{F}(\xi(y)), \tag{12}$$

where $\xi(y) = \mathbf{\Phi}y + z^{\perp}_{TX}$ is the closest point to y in $\text{aff}(X)$. (Here, $z^{\perp}_{TX}$ is the orthogonal translation vector that sends TX to $\text{aff}(X)$: namely, $(z^{\perp}_{TX})^p = \frac{m^p}{n^p}\mathbf{1}$.) Since $\tilde{F} \equiv \nabla\tilde{f}$, the chain and product rules imply that

$$\tilde{F}(y)' = \nabla\tilde{f}(y)'$$

$$
\begin{aligned}
&= \nabla f(\xi(y))'\mathbf{\Phi} + (y-\xi(y))'DF(\xi(y))\mathbf{\Phi} + F(\xi(y))'(I-\mathbf{\Phi})\\
&= (\mathbf{\Phi}F(\xi(y)))'\mathbf{\Phi} + (y-\xi(y))'DF(\xi(y))\mathbf{\Phi} + F(\xi(y))' - F(\xi(y))'\mathbf{\Phi}\\
&= F(\xi(y))'\mathbf{\Phi}\mathbf{\Phi} + (y-\xi(y))'DF(\xi(y))\mathbf{\Phi} + F(\xi(y))' - F(\xi(y))'\mathbf{\Phi}\\
&= (y-\xi(y))'DF(\xi(y))\mathbf{\Phi} + F(\xi(y))'.
\end{aligned}
$$

In particular, if $x \in X$, then $\xi(x) = x$, and so $\tilde{F}(x) = F(x)$. This completes the proof of the theorem. ■

Example 4.2. Random matching in symmetric normal form potential games. If a single population of agents are randomly matched to play the symmetric two-player game $U = (U^1, U^2) = (A, A')$, the resulting population game $F : X \to \mathbf{R}^n$ has payoff function $F(x) = Ax$.

If U is a normal form potential game, then any potential function V of U is a symmetric matrix. In fact, Sandholm (2008) shows that U is a potential game if and only if $\Phi A\Phi$ is symmetric, and that in this case $V = \Phi A\Phi + \Phi A(I-\Phi) + (I-\Phi)A'\Phi$ is a potential function for U. If we define $f : X \to \mathbf{R}$ by $f(x) = \frac{1}{2}x'Vx$, then $\nabla f(x) = \Phi Vx = \Phi Ax = \Phi F(x)$ for all $x \in X$, so F is a potential game with potential function f. Moreover, if we define $\tilde{f} : \mathbf{R}^n_+ \to \mathbf{R}$ by

$$
\tilde{f}(x) = \tfrac{1}{2}\xi(x)'V\xi(x) + (x-\xi(x))'A\xi(x), \quad \text{where} \quad \xi(x) = x + \tfrac{1}{2}(1-\mathbf{1}'x)\mathbf{1}, \tag{13}
$$

then the proof of Theorem 4.1 shows that $\nabla\tilde{f}(x) = F(x)$ for all $x \in X$, so that $\tilde{f}$ encodes all information about payoffs in F, not just information about relative payoffs.

For a simple example, suppose that

$$
A = \begin{pmatrix} 4 & 0 \\ 3 & 2 \end{pmatrix}, \quad \text{so that} \quad V = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}
$$

is a potential function for $U = (A, A')$. Then $f(x) = \frac{1}{2}x'Vx = \frac{1}{2}\left((x_1)^2 + 2(x_2)^2\right)$ satisfies $\nabla f(x) = \Phi F(x) = \Phi Ax$ for all $x \in X$, while applying equation (13) shows that

$$
\tilde{f}(x) = \frac{1}{8}\Big(13(x_1)^2 - 7(x_2)^2 - 6x_1x_2 + 6x_1 + 30x_2\Big)
$$

satisfies $\nabla\tilde{f}(x) = F(x) = Ax$ for all $x \in X$. §

4.2 Differential Characterizations of Potential Games

Like full potential games, potential games can be characterized by a symmetry condition on the payoff derivatives $DF(x)$.

Theorem 4.3. *Suppose that the population game $F : X \to \mathbf{R}^n$ is C^1. Then F is a potential game if and only if it satisfies* externality symmetry*:*

(ES) $\quad$ *$DF(x)$ is symmetric with respect to $TX \times TX$ for all $x \in X$.*

Condition (ES) requires that at each state $x \in X$, the derivative $DF(x)$ be a symmetric bilinear form on $TX \times TX$. This condition can be stated more explicitly as

$$z' DF(x) \hat{z} = \hat{z}' DF(x) z \text{ for all } z, \hat{z} \in TX \text{ and } x \in X. \tag{14}$$

Proof of Theorem 4.3. To begin, suppose that F is a potential game with potential function $f : X \to \mathbf{R}$ as in condition (P). This means that for all $x \in X$ we have $\nabla f(x) = \mathbf{\Phi} F(x)$, or, equivalently, that for all $x \in X$, $\nabla f(x) = F(x)$ as linear forms on TX. Taking the derivative of each side of this identity with respect to directions in TX, we find that $\nabla^2 f(x) = DF(x)$ as bilinear forms on $TX \times TX$. But since $\nabla^2 f(x)$ is a symmetric bilinear form on $TX \times TX$ (by virtue of being a second derivative—see Lang (1997, Theorem 17.5.3)), $DF(x)$ is as well.

Next, suppose that F satisfies condition (ES). Define $\xi : \mathbf{R}^n \to \text{aff}(X)$ by $\xi(y) = \mathbf{\Phi} y + z^{\perp}_{TX}$, as in the proof of Theorem 4.1. Let $\overline{X} = \{y \in \mathbf{R}^n : \xi(y) \in X\}$, and define $\overline{F} : \overline{X} \to \mathbf{R}^n$ by

$$\overline{F}(y) = \mathbf{\Phi} F(\xi(y)).$$

Then differentiating yields

$$D\overline{F}(y) = D\Big(\mathbf{\Phi} F(\xi(y))\Big) = \mathbf{\Phi} DF(\xi(y)) \mathbf{\Phi}. \tag{15}$$

By the standard integrability condition, $\overline{F}$ admits a potential function if and only if $D\overline{F}(y)$ is symmetric for all $y \in \overline{X}$. Equation (15) tells us that the latter statement is true if and only if F satisfies condition (ES).

Now let $\overline{f}$ be a potential function for $\overline{F}$, and let $f = \overline{f}\big|_X$. Then since $\xi(x) = x$ for all $x \in X$, we find that

$$\nabla f(x) = \mathbf{\Phi} \nabla \overline{f}(x) = \mathbf{\Phi} \overline{F}(x) = \mathbf{\Phi}\Big(\mathbf{\Phi} F(\xi(x))\Big) = \mathbf{\Phi} F(x),$$

and so f is a potential function for F in the sense of condition (P). This completes the proof of the theorem. ■

Example 4.4. Two-strategy games. If F: $X \to \mathbf{R}^n$ is a *two-strategy game* (i.e., if $p = 1$, $m = 1$, and $n = 2$), the state space X is the simplex in $\mathbf{R}^2$, so the tangent space TX is spanned by the vector $d = e_2 - e_1$. Thus, if z and $\hat{z}$ are vectors in TX, then $z = kd$ and $\hat{z} = \hat{k}d$ for some real numbers k and $\hat{k}$. This implies that that $z'DF(x)\hat{z} = k\hat{k}\, d'DF(x)\, d = \hat{z}'DF(x)z$ for all $x \in X$, and so that F is a potential game.

Indeed, it is easy to verify that the function $f : X \to \mathbf{R}$ defined by $f(x_1, 1 - x_1) = \int_0^{x_1} (F_1(t, 1-t) - F_2(t, 1-t))\, \mathrm{d}t$ is a potential function for F. While one can arrive at this f by an educated guess, one can also construct f from F mechanically using Proposition 3.4: The function $G : \mathcal{X} = [0, 1] \to \mathbf{R}$ defined by $G(\chi) = G_1(\chi) = F_1(\chi, 1-\chi) - F_2(\chi, 1-\chi)$ has a one-dimensional domain, and so is integrable; if we let $g(\chi) = \int_0^{\chi} G(t)\, \mathrm{d}t$, then $f(x) = g(x_1)$ is the potential function for F introduced above. §

Let e_i^p denote the (i, p)th standard basis vector in $\mathbf{R}^n$. Since the equality in (14) is linear both in z and in $\hat{z}$, and since the set of vectors of the form $e_j^p - e_i^p$ span the subspace TX, condition (14) can be expressed equivalently as

$$\text{(16)} \qquad \frac{\partial(F_j^p - F_i^p)}{\partial(e_l^q - e_k^q)}(x) = \frac{\partial(F_l^q - F_k^q)}{\partial(e_j^p - e_i^p)}(x) \;\text{ for all } i, j \in S^p,\ k, l \in S^q,\ p, q \in \mathcal{P}, \text{ and } x \in X.$$

The left hand side of equation (16) captures the change in the payoff to strategy $j \in S^p$ relative to strategy $i \in S^p$ as agents switch from strategy $k \in S^q$ to strategy $l \in S^q$. This effect must equal the change in the payoff of l relative to k as agents switch from i to j, as expressed on the right hand side of (16). Evidently, this condition and its interpretation are akin to those of full externality symmetry (FES) (see equation (3) and the subsequent discussion); however, equation (16) only refers to *relative* payoffs and to *feasible* changes in the social state.

In the single population case, conditions (14) and (16) are equivalent to a condition introduced by Hofbauer (1985) (also see Hofbauer and Sigmund (1988, Theorem 24.4)). While the three conditions are mathematically equivalent, condition (16) seems to admit the simplest game theoretic interpretation.

Proposition 4.5. *Suppose that F is a C^1 single population game ($p = 1$). Then F is a potential game if and only if it satisfies* triangular integrability*:*

$$\text{(17)} \qquad \frac{\partial F_i}{\partial(e_j - e_k)}(x) + \frac{\partial F_j}{\partial(e_k - e_i)}(x) + \frac{\partial F_k}{\partial(e_i - e_j)}(x) = 0 \;\textit{ for all } i, j, k \in S \textit{ and } x \in X.$$

Proof. In the single population context, condition (16) can be written as

$$(e_j - e_i)'DF(x)\,(e_l - e_k) = (e_l - e_k)'DF(x)\,(e_j - e_i) \;\text{ for all } i, j, k, l \in S \text{ and } x \in X.$$

Since for each fixed $j \in S$, the sets $\{(e_i - e_j)\}_{i \neq j}$ and $\{(e_j - e_k)\}_{k \neq j}$ each span TX, this condition is itself equivalent to the requirement that

$$(e_i - e_j)' DF(x)\,(e_j - e_k) = (e_j - e_k)' DF(x)\,(e_i - e_j) \text{ for all } i, j, k \in S \text{ and } x \in X.$$

Expanding the above equality and collecting terms yields

$$\frac{\partial F_i}{\partial (e_j - e_k)}(x) + \frac{\partial F_j}{\partial (-e_j + e_k - e_i + e_j)}(x) + \frac{\partial F_k}{\partial (e_i - e_j)}(x) = 0 \text{ for all } i, j, k \in S \text{ and } x \in X.$$

This condition is clearly identical to condition (17). ■

The conditions stated above characterize potential games in terms of the derivatives of the payoff function F, whose domain is the set of social states X. Can we obtain a similar characterization that only requires differentiation of the usual sort—that is, differentiation of functions defined on the full-dimensional set $\mathbf{R}^n_+$? Corollary 4.6, a direct consequence of Theorem 4.3 and the relevant definitions, provides such a result.

Corollary 4.6. *Suppose that the population game $F : X \to \mathbf{R}^n$ is C^1, and let $\tilde{F} : \mathbf{R}^n_+ \to \mathbf{R}^n$ be any C^1 extension of F. Then F is a potential game if and only if $\mathbf{\Phi} D\tilde{F}(x) \mathbf{\Phi}$ is symmetric for all $x \in X$.*

4.3 Uses of Full Potential Functions

Despite the previous results, there are still situations in which the use of full potential functions as defined by condition (FP) is warranted. For one example, consider *population games with entry and exit*, in which the mass m^p of agents in population p not only can choose among the strategies in S^p, but also can select an outside option whose payoff is always zero. In such games, the relevant set of social states is $Y = \prod_{p \in \mathcal{P}} Y^p$, where $Y^p = \{x^p \in \mathbf{R}^{n^p}_+ : \sum_{i \in S^p} x^p_i \leq m^p\}$: that is, the number of *active* agents in population p varies between 0 and m^p, and all remaining agents choose the outside option.[5]

In this setting, we call F a *potential game* if there exists a function $f : Y \to \mathbf{R}$ satisfying condition (FP) on Y. It is not difficult to show that the basic results from Sandholm (2001) extend to potential games with entry and exit: the Nash equilibria of F are the states that satisfy the Kuhn-Tucker first-order conditions for maximizing f on Y, and evolutionary dynamics under which growth rates are positively correlated with payoffs converge to connected sets of rest points from all initial conditions.

[5]Of course, one could also model such an interaction as game with fixed population masses m^p, augmented strategy sets $S^p_0 = S^p \cup \{0\}$, and state space $Y_0 = \prod_{p \in \mathcal{P}} Y^p_0$, where $Y^p_0 = \{x^p \in \mathbf{R}^{n^p+1}_+ : \sum_{i \in S^p_0} x^p_i = m^p\}$.

The next example shows why condition (FP) is the appropriate definition of potential when entry and exit are allowed.

Example 4.7. A congestion model with (and without) entry and exit. Suppose that agents in a single unit-mass population select either an active strategy from the set $S = \{1, \ldots, n\}$ or the outside option, so that the set of population states is $Y = \{x \in \mathbf{R}^n_+ : \sum_{i \in S} x_i \leq 1\}$. Each active strategy $i \in S$ has a C^1 payoff function $\tilde{F}_i : Y \to \mathbf{R}$ of the form

$$\tilde{F}_i(x) = u_i(x_i) + v_i(x_T), \tag{18}$$

where we write $x_T \equiv \sum_{k \in S} x_k$. The payoff to strategy i is thus the sum of two terms, one depending on the mass of agents playing strategy i, and the other on the total mass of active agents. By definition, the payoff to the outside option is identically zero.

For the moment, assume that there is no entry or exit: that is, restrict the population state to the set $X_m = \{x \in \mathbf{R}^n_+ : \sum_{i \in S} x_i = m\} \subset Y$, where $m \in (0, 1]$ represents the (fixed) mass of active players. In this case, differences between the payoffs of active strategies completely determine incentives, and so are enough to characterize equilibria and evolutionary dynamics. These payoff differences are precisely what is captured by a potential function $f : X_m \to \mathbf{R}$ satisfying condition (P):[6]

$$\nabla f(x) = \Phi \tilde{F}(x) \text{ for all } x \in X_m.$$

It is easy to verify that the game defined in equation (18) satisfies externality symmetry (ES) throughout the state space Y, and so admits a potential function on X_m in the sense of condition (P). Indeed, such a function is given by

$$f(x) = \sum_{i \in S} \int_0^{x_i} u_i(a)\, \mathrm{d}a + \sum_{i \in S} x_i\, v_i(x_T) \quad \text{for all } x \in X_m. \tag{19}$$

If agents are free to choose the outside option, then understanding incentives requires us to know not only the relative payoffs of different active strategies, but also the absolute payoffs of each strategy, so that they can be compared to the the outside option payoff of zero. This requires a full potential function $\tilde{f} : Y \to \mathbf{R}$, one that satisfies condition (FP):

$$\nabla \tilde{f}(x) = \tilde{F}(x) \text{ for all } x \in Y.$$

Even if we extend its domain to Y in the obvious way, the potential function f from

[6]If f were defined throughout Y, the appropriate analogue of condition (P) would require that $\Phi \nabla f(x) = \Phi \tilde{F}(x)$ at all $x \in Y$; see Observation 3.2.

equation (19) is not a full potential function for the game $\tilde{F}$ from equation (18):

$$\frac{\partial f}{\partial x_i}(x) = u_i(x_i) + v_i(x_T) + \sum_{k \in S} x_k\, v_k'(x_T) \neq \tilde{F}_i(x).$$

In fact, a full potential function for this game does not exist in general, since $\tilde{F}$ violates condition (FES):

$$\text{(20)} \qquad \frac{\partial \tilde{F}_i}{\partial x_j}(x) = v_i'(x_T) \neq v_j'(x_T) = \frac{\partial \tilde{F}_j}{\partial x_i}(x).$$

But $\tilde{F}$ does satisfy condition (FES) under additional assumptions. Suppose that $v_i \equiv v$ for all $i \in S$, so that the payoff to each active strategy depends on the number of active players in the same way. Then it is clear from equation (20) that condition (FES) is satisfied, and in this case

$$\tilde{f}(x) = \sum_{i \in S} \int_0^{x_i} u_i(a)\, \mathrm{d}a + \int_0^{x_T} v(a)\, \mathrm{d}a$$

defines a full potential function for $\tilde{F}$. §

Since condition (FP), unlike condition (P), ensures that potential provides information about average payoffs, it leaves open the possibility of using potential functions to investigate efficiency. Hofbauer and Sigmund (1988, p. 240-241) and Sandholm (2001, Section 5) show that if the full potential function $\tilde{f}$ of a full potential game $\tilde{F}$ is homogeneous of positive degree, then $\tilde{f}$ is proportional to aggregate payoffs in $\tilde{F}$;[7] thus, since evolution increases potential, it increases aggregate payoffs as well. Of course, homogeneity is defined in terms of a function's behavior along rays from the origin, and so is not meaningful for functions defined only on X. Consequently, even in games without outside options, the full force of definition (FP) is needed to prove efficiency results.

References

Akin, E. (1990). The differential geometry of population genetics and evolutionary games. In Lessard, S., editor, *Mathematical and Statistical Developments of Evolutionary Theory*, pages 1–93. Kluwer, Dordrecht.

[7] An example is provided by random matching in a common interest game—see Example 3.1.

Beckmann, M., McGuire, C. B., and Winsten, C. B. (1956). *Studies in the Economics of Transportation*. Yale University Press, New Haven.

Hofbauer, J. (1985). The selection mutation equation. *Journal of Mathematical Biology*, 23:41–53.

Hofbauer, J. and Sigmund, K. (1988). *Theory of Evolution and Dynamical Systems*. Cambridge University Press, Cambridge.

Krantz, S. G. and Parks, H. R. (1999). *The Geometry of Domains in Space*. Birkhäuser, Boston.

Lang, S. (1997). *Undergraduate Analysis*. Springer, New York, second edition.

Monderer, D. and Shapley, L. S. (1996). Potential games. *Games and Economic Behavior*, 14:124–143.

Sandholm, W. H. (2001). Potential games with continuous player sets. *Journal of Economic Theory*, 97:81–108.

Sandholm, W. H. (2008). Decompositions and potentials for normal form games. Unpublished manuscript, University of Wisconsin.