What Information Theory says about Bounded Rational Best Response

David H. Wolpert[1]

NASA Ames Research Center, Moffett Field, CA, 94035, USA
dhw@email.arc.nasa.gov

Summary. Probability Collectives (PC) provides the information-theoretic extension of conventional full-rationality game theory to bounded rational games. Here an explicit solution to the equations giving the bounded rationality equilibrium of a game is presented. Then PC is used to investigate games in which the players use bounded rational best-response strategies. Next it is shown that in the continuum-time limit, bounded rational best response games result in a variant of the replicator dynamics of evolutionary game theory. It is then shown that for team (shared-payoff) games, this variant of replicator dynamics is identical to Newton-Raphson iterative optimization of the shared utility function.

1 Introduction

Recent work has used information theory [9, 12] to provide a principled extension of noncooperate conventional game theory to accommodate bounded rationality [25, 27]. Intuitively, this extension starts with the observation that in the real world ascertaining a game's equilibrium is an exercise in statistical inference: one is given (or assumes) partial information about the behavior of the players, and from that infers (!) what the joint mixed strategy is likely to be. There are many ways to do such statistical inference. The one investigated in [27] is based on information theory's version of Occam's razor: Predict the joint mixed strategy that has as little extra information as possible beyond the provided partial knowledge while being consistent with that knowledge. This version of Occam's razor is known as the Maximum entropy (Maxent) principle [9, 12]. It tells us that the mixed strategy of a game's equilibrium, $q(x \in X) = \prod_i q_i(x_i)$, is the solution to a coupled set of Lagrangian functions that are specified by the game structure and the provided partial knowledge.

Sec. 2 reviews how information theory can be used to derive bounded rational noncooperative game theory. Some simple examples of the bounded rational equilibrium solutions of games are then presented. Sec. 3 analyzes scenarios in which the players use bounded rational versions of best response strategies. Particular attention is paid to team games, in which the players share the same utility function. The analysis for this case provide insight into how to optimize the sequence of moves by the players, as far as their shared utility is concerned. This can be viewed as a formal way to optimize the organization chart of a corporation.

Best response strategies, even bounded rational ones, are poor models of real-world computational players that use Reinforcement Learning (RL) [20]. Sec. 4 considers iterated games in which players use a (bounded rational) variant of best response, a variant that is more realistic for computational players, and arguably for human players as well. In this variant the conditional expected utilities used by player i to update her strategy, expected payoff given move x_i, is a decaying average of recent conditional expected utilities. This decay biases the player to dampen large and sudden changes in her strategy. This variant is then explored for the case of team games. The continuum limit of the dynamics of such games is shown to be variant of the replicator dynamics. It is shown such continuum-limit bounded rational best response is identical to Newton-Raphson iterative optimization of the shared utility function of such games.

The formalism presented in this paper is a special case of the field of Probability Collectives (PC), a case in which the joint distribution over the variables of interest is a product distribution. This special case is known as Product Distribution (PD) theory [25, 27, 29, 28, 26, 7]. PC has many applications beyond those considered in this paper, e.g., distributed optimization and control [16, 15, 2, 29]. Finally, see [16] for relations to other work in game theory, optimization, statistical physics, and reinforcement learning.

2 Bounded Rational Noncooperative Game Theory

In this section we motivate PD theory as the information-theoretic formulation of bounded rational game theory. We use the integral sign ($\int$) with the associated measure implicit, i.e., it indicates sums if appropriate, Lebesgue integrals over $\mathbb{R}^n$ if appropriate, etc. In addition, the subscript (i) is used to indicate all index values other than i. Finally, we use $\mathcal{P}$ to indicate the set of all probability distributions over a vector space, and $\mathcal{Q}$ to indicate the subset of $\mathcal{P}$ consisting of all product distributions (i.e., the associated Cartesian product of unit simplices).

In noncooperative game theory one has a set of N **players**. Each player i has its own set of allowed **pure strategies**. A **mixed strategy** is a distribution $q_i(x_i)$ over player i's possible pure strategies. Each player i also has a **private utility** function g_i that maps the pure strategies adopted by all N of the players into the real numbers. So given mixed strategies of all the players, the expected utility of player i is $E(g_i) = \int dx \ \prod_j q_j(x_j) g_i(x)$.

In a **Nash equilibrium** every player adopts the mixed strategy that maximizes its expected utility, given the mixed strategies of the other players. More formally, $\forall i, q_i = \text{argmax}_{q'_i} \int dx \; q'_i \prod_{j \neq i} q_j(x_j) \; g_i(x)$. Perhaps the major objection that has been raised to the Nash equilibrium concept is its assumption of **full rationality** [10, 6, 18, 4]. This is the assumption that every player i can both calculate what the strategies $q_{j \neq i}$ will be and then calculate its associated optimal distribution. In other words, it is the assumption that every player will calculate the entire joint distribution $q(x) = \prod_j q_j(x_j)$.

In the real world, this assumption of full rationality almost never holds, whether the players are humans, animals, or computational agents [5, 17, 10, 3, 8, 1, 22, 14]. This is due to the cost of computation of that optimal distribution, if nothing else. This real-world **bounded rationality** is a major impediment to applying conventional game theory in the real world.

2.1 Review of the Minimum Information Principle

Shannon was the first person to realize that based on any of several separate sets of very simple desiderata, there is a unique real-valued quantification of the amount of syntactic information in a distribution $P(y)$. He showed that this amount of information is the negative of the Shannon entropy of that distribution, $S(P) = -\int dy \; P(y) \ln[\frac{P(y)}{\mu(y)}]$. So for example, the distribution with minimal information is the one that doesn't distinguish at all between the various y, i.e., the uniform distribution. Conversely, the most informative distribution is the one that specifies a single possible y. Note that for a product distribution, entropy is additive, i.e., $S(\prod_i q_i(y_i)) = \sum_i S(q_i)$.

Say we given some incomplete prior knowledge about a distribution $P(y)$. How should one estimate $P(y)$ based on that prior knowledge? Shannon's result tells us how to do that in the most conservative way: have your estimate of $P(y)$ contain the minimal amount of extra information beyond that already contained in the prior knowledge about $P(y)$. Intuitively, this can be viewed as a version of Occam's razor: introduce as little extra information beyond that you are provided in your inferring of P. This minimum information approach is called the maxent principle. It has proven extremely powerful in domains ranging from signal processing to supervised learning [12]. In particular, it is has been successfully used in many statistics applications, including econometrics [13]. It has even provided what many consider the cleanest derivation of the foundations of statistical physics [11].

2.2 Maxent Lagrangians

Much of the work on equilibrium concepts in game theory adopts the perspective of an external observer of a game. We are told something concerning the game, e.g., its cost functions, information sets, etc., and from that wish to predict what joint strategy will be followed by real-world players of the game. Say that in addition to such information, we are told the expected utilities of the players. What is our best estimate of the distribution q that generated those expected cost values? By the maxent principle, it is the distribution with maximal entropy, subject to those expectation values.

To formalize this, for simplicity assume a finite number of players and of possible strategies for each player. To agree with the convention in fields other than game theory (e.g., optimization, statistical physics, etc.), from now on we implicitly flip the sign of each g_i so that the associated player i wants to minimize that function rather than maximize it. Intuitively, this flipped $g_i(x)$ is the "cost" to player i when the joint-strategy is x.

With this convention, given prior knowledge that the expected utilities of the players are given by the set of values $\{\epsilon_i\}$, the maxent estimate of the associated q is given by the minimizer of the Lagrangian

$$\mathscr{L}(q) \triangleq \sum_i \beta_i[E_q(g_i) - \epsilon_i] - S(q) \tag{1}$$

$$= \sum_i \beta_i[\int dx \prod_j q_j(x_j) g_i(x) - \epsilon_i] - S(q) \tag{2}$$

where the subscript on the expectation value indicates that it evaluated under distribution q. The $\{\beta_i\}$ are "inverse temperatures" implicitly set by the constraints on the expected utilities.

Solving, we get the coupled equations

$$q_i(x_i) \propto e^{-E_{q_{(i)}}(G|x_i)} \tag{3}$$

where the overall proportionality constant for each i is set by normalization, and $G \triangleq \sum_i \beta_i g_i$ [1]. In Eq. 3 the probability of player i choosing pure strategy x_i depends on the effect of that choice on the utilities of the other players. This reflects the fact that our prior knowledge concerns all the players equally.

If we wish to focus only on the behavior of player i, it is appropriate to modify our prior knowledge. First consider the case of maximal prior knowledge, in which we know the actual joint-strategy of the players, and therefore all of their expected costs. For this case, trivially, the maxent principle says we should "estimate" q as that joint-strategy (it being the q with maximal entropy that is consistent with our prior knowledge). The same conclusion holds if our prior knowledge also includes the expected cost of player i.

Modify this maximal set of prior knowledge by removing from it specification of player i's strategy. So our prior knowledge is the mixed strategies of all players other than i, together with player i's expected cost. We can incorporate prior knowledge of the other players' mixed strategies directly, without introducing Lagrange parameters. The resultant **maxent Lagrangian** is

$$\mathscr{L}_i(q_i) \triangleq \beta_i[\epsilon_i - E_q(g_i)] - S_i(q_i)$$

solved by a set of coupled **Boltzmann distributions**:

[1] The subscript $q_{(i)}$ on the expectation value indicates that it is evaluated according the distribution $\prod_{j \neq i} q_j$.

$$q_i(x_i) \propto e^{-\beta_i E_{q_{(i)}}(g_i|x_i)}. \tag{4}$$

Following Nash, we can use Brouwer's fixed point theorem to establish that for any non-negative values $\{\beta\}$, there must exist at least one product distribution given by the product of these Boltzmann distributions (one term in the product for each i).

The first term in $\mathscr{L}_i$ is minimized by a perfectly rational player. The second term is minimized by a perfectly *irrational* player, i.e., by a perfectly uniform mixed strategy q_i. So β_i in the maxent Lagrangian explicitly specifies the balance between the rational and irrational behavior of the player. In particular, for $\beta \to \infty$, by minimizing the Lagrangians we recover the Nash equilibria of the game. More formally, in that limit the set of q that simultaneously minimize the Lagrangians is the set of mixed strategy equilibria of the game, together with the set of delta functions about the pure Nash equilibria of the game. The same is true for Eq. 3.

Note also that independent of information-theoretic considerations, the Boltzmann distribution is a reasonable (highly abstracted) model of human behavior. Typically humans do some "exploration" as well as "exploitation", trying each move with probability that rises as the expected cost of the move falls. This is captured in the Boltzmann distribution mixed strategy.

One can formalize the concept of the rationality of a player in a way that applies to any distribution, not just a Boltzmann distribution. One does this with a **rationality operator** which maps a q and a g_i to a non-negative real value measuring the rationality of player i in adopting strategy q_i given private cost function g_i and strategies $q_{(i)}$ of the other players. For the solution in Eq. 4 and private cost g_i, the value of that operator is just β_i [27].

Eq. 3 is just a special case of Eq. 4, where all player's share the same private cost function, G. (Such games are known as **team games**.) This relationship reflects the fact that for this case, the difference between the maxent Lagrangian and the one in Eq. 2 is independent of q_i. Due to this relationship, our guarantee of the existence of a solution to the set of maxent Lagrangians implies the existence of a solution of the form Eq. 3. Typically players will be closer to minimizing their expected cost than maximizing it. For prior knowledge consistent with such a case, the β_i are all non-negative.

For each player i define $f_i(x, q_i(x_i)) \triangleq \beta_i g_i(x) + \ln[q_i(x_i)]$. Then we can write the maxent Lagrangian for player i as

$$\mathscr{L}_i(q) = \int dx\; q(x) f_i(x, q_i(x_i)). \tag{5}$$

Now in a bounded rational game every player sets its strategy to minimize its Lagrangian, given the strategies of the other players. In light of Eq. 5, this means that we can interpret each player in a bounded rational game as being perfectly rational for a cost function that incorporates its computational cost. To do so we simply need to expand the domain of "cost functions" to include (logarithms of) probability values as well as joint moves.

2.3 Examples of Bounded Rational Equilibria

It can be difficult to start with a set of cost functions and associated rationalities β_i and then solve for the associated bounded rational equilibrium q. Solving for q when prior knowledge consists of expected costs ϵ_i rather than rationalities can be even more tedious. (In that situation the β_i are not specified upfront but instead are Lagrange parameters that we must solve for.) However there is an alternative approach to constructing examples of games and their bounded rational equilibria that is quite simple. In this alternative one starts with a particular mixed strategy q and then solves for a game for which q is a bounded rational equilibrium, rather than the other way around.

To illustrate this, consider a 2-player single-stage game. Let each player have 3 possible moves, indicated by the numerals 0, 1, and 2. Say the (bounded rational) mixed strategy equilibrium is

$$\begin{aligned} &q_1(0) = 1/2,\ q_1(1) = 1/4,\ q_1(2) = 1/4; \\ &q_2(0) = 2/3,\ q_2(1) = 1/4,\ q_2(2) = 1/12\ . \end{aligned} \tag{6}$$

Now we know that at the equilibrium, $q_1(x_1) \propto e^{-\beta_1 E(g_1|x_1)}$, where β_1 is player 1's rationality, and g_1 is her cost function (the negative of her cost function). This means for example that

$$e^{-(\beta_1[E(g_1|x_1=0)-E(g_1|x_1=1)])} \;=\; \frac{q_1(0)}{q_1(1)} \;=\; 2, \;\; \text{i.e.,}$$

$$\beta_1[E(g_1 \mid x_1 = 0) - E(g_1 \mid x_1 = 1)] = -\ln(2). \tag{7}$$

A similar equation governs the remaining independent difference in expectation values for player 1. The analogous two equations for player 2 also hold.

Now define the vectors $\mathbf{g}_{i;j}(.) \triangleq g_i(x_i = j, .)$. So for example $\mathbf{g}_{1;0} = (g_1(x_1 = 0, x_2 = 0), g_1(x_1 = 0, x_2 = 1), g_1(x_1 = 0, x_2 = 2))$. Then we can express our equations compactly as four dot product equalities:

$$\begin{aligned} \beta_1(\mathbf{g}_{1;0} - \mathbf{g}_{1;1}) \cdot q_2 \;&=\; -\ln(2)\ ; \quad \beta_1(\mathbf{g}_{1;0} - \mathbf{g}_{1;2}) \cdot q_2 \;=\; -\ln(2)\ ; \\ \beta_2(\mathbf{g}_{2;0} - \mathbf{g}_{2;1}) \cdot q_1 \;&=\; -\ln(8/3)\ ; \quad \beta_2(\mathbf{g}_{2;0} - \mathbf{g}_{2;2}) \cdot q_1 \;=\; -\ln(8)\ . \end{aligned} \tag{8}$$

We can absorb each β_i into its associated g_i; all that matters is their product.

We can now plug in for the vectors q_1 and q_2 from Eq. 6 and simply write down a set of solutions for the four three-dimensional vectors $\mathbf{g}_{i,j}$. For these $\{g_i\}$ the bounded ratinal equilibrium is given by the q of Eq. 6. If desired, we can evaluate the associated expected values of the cost functions for the two players; our q is the bounded ratinal equilibrium for those expected costs.

Note that the variables in the first pair of equalities in Eq. 8 are independent of those in the second pair. In other words, whereas the Boltzmann equations giving q for a specified set of g_i are a set of coupled equations, the equations giving the g_i for a specified q are not coupled. Note also that our equations for the $\mathbf{g}_{i;j}$ are (extremely) underconstrained. This illustrates how compressive the mapping from the g_i to the associated equilibrium q is. Bear in mind though that that mapping is also multi-valued in general; in general a single set of cost functions can have more than one equilibrium, just like it can have more than one Nash equilibrium.

The generalization of this example to arbitrary numbers of players with arbitrary move spaces is immediate. As before, indicate the moves of every player by an associated set of integer numerals starting at 0. Recall that the subscript (i) on a vector indicate all components but the i'th one. Also absorb the rationalities β_i into the associated g_i.

Now specify q and the vectors $g_i(x_i = 0, .)$ (one vector for each i) to be anything whatsoever. Then for all players i, the only associated constraint on the i'th cost function concerns certain projections of the vectors $g_i(x_i > 0, .)$ (one projection for each value $x_i > 0$). Concretely, $\forall i, x_i > 0$,

$$\int dx'_{(i)}\; g_i(x_i, x'_{(i)}) \prod_{j \neq i} q_j(x'_j) = -\ln(\frac{q_i(0)}{q_i(x_i)}) + \int dx'_{(i)} g_i(0, x'_{(i)}) \prod_{j \neq i} q_j(x'_j),$$

$$\text{i.e., } \forall i, x_i > 0, \mathbf{g}_i(x_i, .) \cdot q_{(i)} = -\ln(\frac{q_i(0)}{q_i(x_i)}) + \mathbf{g}_i(0, .) \cdot q_{(i)}. \tag{9}$$

All the terms on the right-hand side are specified, as well as the $q_{(i)}$ term on the left-hand side. Any $\mathbf{g}_i(x_i, .)$ that obeys the associated equation has the specified q as a bounded rational equilibrium.

See [27] for discussion of alternative interpretations of this information-theoretic formulation of bounded rationality. That reference also discusses kinds of prior knowledge that do not result in the Maxent Lagrangian, in particular knowledge based on finite data sets (Bayesian inference). A scalar-valued quantification of the rationality of a player is also presented there.

3 Bounded Rational Versions of Best Response

One crude way to try to find the q given by Eq. 4 would be an iterative process akin to the best-response scheme of game theory [10]. Given any current distribution q, in this scheme all agents i simultaneously replace their current distributions. In this replacement each agent i replaces q_i with the distribution given in Eq. 4 *based on the current* $q_{(i)}$. This scheme is the basis of the use of Brouwer's fixed point theorem to prove that a solution to Eq. 4 exists. Accordingly, it is called **parallel Brouwer updating**. (This scheme goes by many names in the literature, from Boltzmann learning in the RL community to block relaxation in the optimization community.)

Sometimes conditional expected costs for each agent can be calculated explicitly at each iteration. More generally, they must be estimated. This can be done via Monte Carlo sampling, iterated across a block of time. During that block the agents all repeatedly and jointly IID sample their (unchanging) probability distributions to generate joint moves, and the associated cost values are recorded. These are then use to estimate all the conditional expected costs, which then determine the parallel Brouwer update.

This is what is done in RL schemes in which each agent maintains an empirical estimate of its cost for each of its moves, and chooses its actual move stochastically, by sampling a Boltzmann distribution of those estimates. (See [25] for ways to get accurate MC estimates more efficiently than in this simple scheme.)

One alternative to parallel Brouwer updating is **serial** Brouwer updating, where we only update one q_i at a time. This is analogous to a Stackelberg game, in that one agent makes its move and then the other(s) respond [4, 6]. In a team game, any serial Brouwer updating must reduce the common Lagrangian, in contrast to the case with parallel Brouwer updating.

There are many versions of serial updating. In **cyclic** serial Brouwer updating, one cycles through the i in order. In **random** serial Brouwer updating, one cycles through them in a random fashion. In **greedy** serial Brouwer updating, instead of cycling through all i, at each iteration we choose what single player to update based on the associated drop in the common Lagrangian. Those drops can be evaluated without calculating the associated Boltzmann distributions. To see how, use N_i to indicate the normalization constant of Eq. 4. Then define the **Lagrangian gap** at q for player i as $\ln[N_i] + \int dx_i q_i(x_i) E_{q_{(i)}}(g_i \mid x_i) + \int dx_i q_i(x_i) \ln[q_i(x_i)]$. This is how much $\mathscr{L}$ is reduced if only q_i undergoes the Brouwer update [2].

Another obvious variant of these schemes is mixed serial/parallel Brouwer updating, in which one subset of the players moves in synchrony, followed by another subset, and so on. Such updating in a team game can be viewed as a simple model of the organization chart of the players. For example, this is the case when the players are a corporation, with G being a common cost function based on the corporation's performance.

Say we observe the functioning of such an organization over time, and view those observations as Monte Carlo sampling of its behavior. Then we can use those samples to statistically estimate how best to do serial/parallel Brouwer updating, for the purpose of minimizing the shared cost function G. This can be viewed as a way to optimize the organization chart coupling the players.

4 Parallel Brouwer with Data-Aging is Nearest Newton

This section considers a variant of best-response that is more realistic (more accurately models RL-based computational players that are actually used in

[2]Proof sketch: Write entropy after the update as a sum of non-i entropies (unchanged by the update) plus i's new entropy. Then expand i's new entropy. This gives the value of the new Lagrangian as -$\ln[N_i]$. Then do the subtraction.

machine learning, and arguably more accurately models human players as well). In this variant the expected cost used by each player to update her strategy is a decaying average of recent expected utilities; this decay reflects a conservative preference for dampening large changes in strategy.

Such a bias is used (implicitly or otherwise) in most multi-player RL algorithms. For example, in the COIN framework each agent i collects a data set of pairs of what value its private cost function has at timestep t together with the move it made then. It then estimates its cost for move x_i as a weighted average of all the cost values in its data set for that move. The weights are exponentially decaying functions of how long ago the associated observation was made. This **data-aging** is crucial to reflect the non-stationarity of agent i's environment, i.e., that the other agents are changing their strategies with time. Arguably, humans use similar modifications to best response. Indeed, in idealized learning rules like ficticious play, such dampening is crucial.[3]

4.1 The Dynamics of Brouwer Updating

Consider a multi-stage game where at the end of iteration t, each player i updates her distribution $q_i(., t)$ to

$$q_i(x_i, t) = \frac{e^{-\Phi_i(x_i,t)}}{\int dx'_i e^{-\Phi_i(x'_i,t)}} \tag{10}$$

where $\Phi_i(x_i, t)$ can implicitly depend on distributions for times $t' \neq t$. This is a generalization of parallel Brouwer updating, where the function being exponentiated can be Q values (as in Q-learning [24]), single-instant reward values, distorted versions of these (e.g., to incorporate data-aging), etc. This generalization encompasses both uncountably and countably many values of t.

As an example, for single-instant rewards (i.e., conventional parallel Brouwer), $\Phi_i(x_i, t)$ is player i's estimate of (β_i times) her conditional expected cost for taking move x_i at time $t - 1$. If that estimate were exact, this would mean

$$\Phi_i(x_i, t) = \beta E(g_i \mid x_i) \;=\; \beta \int dx_{(i)} q_{(i)}(x_{(i)}, t-1) g_i(x_i, x_{(i)}). \tag{11}$$

As another example, for Q-learning, one player is Nature and for all t her distribution is a delta function. In this case $\Phi_i(x_i, t)$ is the Q-value for player

[3]Brouwer updating with data-aging has minimal memory requirements on the agents. Say agent i has just made a particular move, getting cost r, and that the most recent previous time it made that time was T iterations ago. Then the new estimated cost for that move, E', is related to the previous one, E, by $E' = \frac{r+k^T Ea}{1+k^T a}$, where k is a constant less than 1, and a is initially set to 1, while itself also being updated according to a += k^T. So agent i only needs to keep a running tally of E, a, and T for each of its possible moves to use data-aging, rather than a tally of all historical time-cost pairs.

i taking action x_i, when the state of Nature is as specified by the associated delta function in $q(., t-1)$.

Note that no Monte Carlo sampling is being done here, as there is in most real RL; this is an abstracted version of such RL. Alternatively, the analysis here becomes exact when Φ_i is evaluated closed form, or (if Φ_i is an empirical expectation value) when there are enough Monte Carlo samples so that empirical averages effectively give exact values of expected quantities.

At this point we have to say something about how Φ_i evolves with time. Consider the case where Φ_i is an "estimate" of the current value of some function ϕ_i, formed by exponential aging of the previous (exactly known) ϕ_i values. In our case, assuming there have been an infinite number of preceding timesteps, at each of which we have the function ϕ_i, this is the same as geometric data-aging:

$$\Phi_i(x_i, t) = \alpha\phi_i(x_i, q(t-1)) + (1-\alpha)\Phi_i(x_i, t-1). \tag{12}$$

For example, in parallel Brouwer updating, $\phi_i(x_i, t) = \beta E(g_i \mid x_i, q_{(i)}(t))$, while $\Phi_i(x_i, t)$ is a geometric average of the previous values $\{\phi_i(x_i, t' < t)\}$. [4]

4.2 The Continuum-Time Limit

To go to the continuum-time limit, let t be a real variable, and replace the temporal delay value of 1 in Eq. 12 with δ, and replace α with $\alpha\delta$ (we'll eventually take $\delta \to 0$). In the $\delta \to 0$ limit, assuming q is a continuous function of t, Eq. 12 becomes

$$\frac{d\Phi_i(x_i, q)}{dt} = \alpha[\phi_i(x_i, q) - \Phi_i(x_i, q)]. \tag{13}$$

where from now on, for clarity the t argument in Φ_i is no longer explicitly written, whereas the $q(t)$ variable is. In addition, since Eq. 10 already allows for uncountable t, we can differentiate it with respect to t to get

$$\frac{dq(x_i, t)}{dt} = -q_i(x_i, t)[\frac{d\Phi_i(x_i, t)}{dt} - \int dx'_i q_i(x'_i, t)\frac{d\Phi_i(x'_i, t)}{dt}]. \tag{14}$$

If we knew the dynamics of ϕ_i, we could solve Eq. 13 via integrating factors, in the usual way. Instead, here we'll plug Eq. 13's formula for $\frac{d\Phi_i}{dt}$ into Eq. 14. Then use Eq. 10 to write $\Phi_i(x_i, q) =$ constant $-\ln(q_i(x_i))$. The result is

$$\begin{aligned}\frac{dq_i(x_i)}{dt} &= \alpha q_i(x_i)\ [\phi_i(x_i, q) + \ln(q_i(x_i))]\ - \\ &\int dx'_i\ \alpha q_i(x'_i)[\phi_i(x'_i, q_i) + \ln(q_i(x'_i))].\end{aligned} \tag{15}$$

[4]To see that Eq. 12 is exponential data-aging of ϕ_i, write the value of $\Phi_i(x_i, t)$ under such aging as a linear combination of two terms, one of which involves the value $\Phi_i(x_i, t-1)$. Then set the aging exponent γ = -ln$(1-\alpha)$, to get Eq. 12.

4.3 Relation with Nearest Newton Descent and Replicator Dynamics

As mentioned previously, there are many ways to find equilibria, and in particular many distributed algorithms for doing so. This is especially so in team games, where finding such equilibria reduces to descending a single overarching Lagrangian. One natural idea for descent in such games is to use the Newton-Raphson descent algorithm. However that algorithm cannot be applied directly to search across q in a distributed fashion, due to the need to invert matrices coupling the agents. As an alternative, one can consider what new distribution p the Newton algorithm would step to if there was no restriction that p be a product distribution. One can then ask what product distribution is closest to p, according to Kullback-Leibler distance [9]. It turns out that one *can* solve for that optimal product distribution. The associated update rule is called the **Nearest Newton** algorithm [29].

It turns out that when one writes down the Nearest Newton update rule, it says to replace each component $q_i(x_i, t)$ with the exact quantity appearing on the right-hand side of Eq. 15, where α is the stepsize of the update, and $\phi_i(x_i, t) = \beta E(G \mid x_i, q_{(i)}(., t))$, as in parallel Brouwer updating for a team game [5]. In other words, in team games, the continuum limit of having each player using (bounded rational) best response is identical to the continuum limit of the Newton-Raphson algorithm for descending the Lagrangian, with the data-aging parameter α giving the stepsize.

Eq. 15 arises in yet other contexts as well. Say Φ_i is conditional expected reward for agent i. In other words, $\phi_i(x_i, t-1) = \beta E(g_i \mid x_i, q(., t-1))$, exactly as in Brouwer updating. Under this choice, the ln term in Eq. 15 vanishes in the $\beta \to \infty$ limit. This makes the righthand side of Eq. 15 involve only the difference between player i's expected cost and the average expected cost of all players:

$$\begin{aligned} \frac{dq_i(x_i)}{dt} &= \alpha q_i(x_i)\ [\phi_i(x_i, q)\ -\ \int dx'_i\ q_i(x_i)\phi_i(x'_i, q)] \\ &= \alpha q_i(x_i)[E(g_i \mid x_i, q(., t)) - E(g_i \mid q(., t))]. \end{aligned} \tag{16}$$

Now take all agents to share the same move space, g_i to be the same function G for all agents, and similarly $q_i(., t)$ to be independent of i, so that it can be written as $Q(., t)$. (Note that if g_i is independent of i, and in addition $q_i(., t)$ is independent of i at one moment t, then $q_i(., t')$ is independent of i for all t'.) Under this choice, Eq. 16 reduces to

$$\frac{dQ(x)}{dt} = \alpha Q(x)[E(G \mid x, Q(., t)) - E(G \mid Q(., t))]. \tag{17}$$

[5] More generally, Nearest Newton uses this update rule with $\phi_i(x_i, t) = \beta E(g_i \mid x_i, q_{(i)}(., t))$ where each $g_i(x) = G(x) - D(x_{(i)})$ for some function D. See [29].

This is one common form of the replicator dynamics equation of evolutionary game theory [21, 23].[6] In other words, Eq. 16 has been been explored as an abstraction of the dynamics of populations of organisms playing non-cooperative games with one another. In that abstraction, one has an infinite population of agents where a fraction $Q(x,t)$ of them would play move x if asked to play a game at time t. One randomly chooses N of the agents and has them play such a game with each other. The payoff to player i when the joint move is x is given by (the symmetric payoff matrix) $G(x)$. Then the proportional increase of the fraction of players making move x is given by the average payoff for those players, minus the average payoff for all players.

This 3-way connection suggests using some of the techniques for solving replicator dynamics to expedite either parallel Brouwer or Nearest Newton.

4.4 Convergence and Equilibria

By Eq. 15, at equilibrium, for each i, $q_i(x_i)[\phi_i(x_i,q) + \ln(q_i(x_i))]$ must be independent of i. One way this can occur is if this product equals 0. Now $q_i(x_i)$ can never be 0, by Eq. 10. Therefore for this product to equal 0, it must be that $q_i(x_i) \propto e^{-\phi_i(x_i,q)}$. Intuitively, this is exactly what we want, according to Eq. 10 and our interpretation of $\Phi_i(x_i,q)$ as an estimate of $\phi_i(x_i,q)$. Note that this solution means that $\phi_i(x_i,q) = \Phi_i(x_i,q)$, so that (by Eq. 13) $\Phi_i(x_i,q)$ has also reached an equilibrium.

When our equilibrium has $q_i(x_i)[\phi_i(x_i,q) + \ln(q_i(x_i))] = A \neq 0$, we have

$$q_i(x_i) \propto e^{-q_i(x_i)\phi_i(x_i,q)}. \tag{18}$$

In light of Eq. 10, this means that $\Phi_i(x_i,q) \neq \phi_i(x_i,q)$. So by Eq. 13, $\Phi_i(x_i,q)$ hasn't reached an equilibrium in this case:

$$\frac{d\Phi_i(x_i,q)}{dt} = \alpha\phi_i(x_i,q)[1 - q_i(x_i)]. \tag{19}$$

If both $q_i(x_i)$ and $\phi_i(x_i,q)$ were frozen at this point, this solution for $\Phi_i(x_i,q)$ would not obey Eq. 12. So either $q_i(x_i)$ and/or $\phi_i(x_i,q)$ cannot be frozen. In fact, if $\phi_i(x_i,q)$ varies with time, then we know by Eq. 15 that $q_i(x_i)$ varies as well. So in either case $q_i(x_i)$ must vary, i.e., this is not a proper equilibrium.

Although the dynamics has the desired fixed point, it may take a long time to converge there. There are several ways to analyze that: One is to examine the second derivatives (with respect to time) of the q_i and/or the Φ_i. Another is to examine the time-dependence of the residual error,

$$r_i^{ge}(x_i,t) \triangleq \frac{e^{-\Phi_i(x_i,t)}}{\int dx'_i e^{-\Phi_i(x'_i,t)}} - \frac{e^{-\phi_i(x_i,t)}}{\int dx'_i e^{-\phi_i(x'_i,t)}}. \tag{20}$$

[6] If the stepsize α is an appropriately increasing function of $E(G)$ other versions of that dynamics arise. In particular, α is inversely proportional to $E(G)$ in some versions of the replicator dynamics.

The next subsection includes a convergence analysis involving residual errors, but for pedagogical purposes, it considers a different variant of Brouwer from the ones considered so far.

4.5 Other Variants of Brouwer Updating

Data-aging can be viewed as moving only part-way from the current Φ_i to what it should be (i.e. to ϕ_i). An alternative is to dispense with the Φ_i and ϕ_i altogether, and instead step part-way from the current q to what it should be, i.e., partially move to the (bounded rational) best response mixed strategy. Formally, this means replacing Eq. 10 so that the update is not implicit, in how $\Phi_i(x_i, t)$ depends on the past value of $q(t-1)$ (Eq. 12), but explicit:

$$q_i(x_i, t) = q_i(x_i, t-1) + \alpha[h_i(x_i, q_{(i)}(t-1)) - q_i(x_i, t-1)] \tag{21}$$

where $h_i(x_i, q_{(i)}(t))$ is the Boltzmann distribution over x_i giving what $q_i(x_i, t)$ should be to minimize the Lagrangian for fixed (not necessarily optimal) $q_{(i)}$, and we implicitly have small stepsize α.

The only fixed point of this updating rule is where $q_i = h_i \ \forall i$. So just like with continuum-limit parallel Brouwer, we have the correct equilibrium. To investigate how fast the update rule of Eq. 21 arrives at that equilibrium, write its error at time t as the residual

$$\begin{aligned} r_i^{st}(x_i, t) &= q_i(x_i, t) - h_i(x_i, q_{(i)}(t)) \\ &= q_i(x_i, t-1)[1-\alpha] \ + \ \alpha h_i(x_i, q_{(i)}(t-1)) \ - \ h_i(x_i, q_{(i)}(t)) \\ &= q_i(x_i, t-1)[1-\alpha] \ + \ \alpha h_i(x_i, q_{(i)}(t-1)) \\ &\quad - \ h_i[x_i, q_{(i)}(t-1) \ + \ \alpha[h_{(i)}(q(t-1)) - q_{(i)}(t-1)]] \end{aligned} \tag{22}$$

where we have assumed that all all players other than i are updating themselves in the same way that i does (i.e., via Eq. 21), and $h_{(i)}(q(t-1))$ means the vector of the values of all $h_{j \neq i}(x_j)$ evaluated for $q(t-1)$.

With obvious notation, rewrite Eq. 22 as

$$\begin{aligned} r_i^{st}(x_i, t) = q_i(x_i, t-1)[1-\alpha] \\ + \ \alpha h_i(x_i, q_{(i)}(t-1)) \\ - \ h_i[x_i, \ q_{(i)}(t-1) - \alpha r_{(i)}(t-1)]. \end{aligned} \tag{23}$$

Now use the fact that α is small to expand the last h_i term on the righthand side to first order in its second (vector-valued) argument, getting the result

$$r_i^{st}(x_i, t) \approx r_i(x_i, t)[1-\alpha] \ + \ \alpha \nabla h_i \cdot r_{(i)}(t-1) \tag{24}$$

where the gradient of h_i is with respect to the vector components of its second argument. Accordingly, if $r_i^{st}(x_i)$ starts much larger than the other residuals, it will be pushed down to their values. Conversely, if it starts much smaller than them, it will rise.

There are other ways one can reduce a stochastic game to a deterministic continuum-time process. In particular, this can be done in closed form for ficticious play games and some simple variants of it [19, 10].

Acknowledgements: I would like to thank Stefan Bieniawski, Bill Macready, George Judge, Chris Henze, and Ilan Kroo for helpful discussion.

References

1. Al-Najjar, N. I., and R. Smorodinsky, "Large nonanonymous repeated games", *Game and Economic Behavior* **37**, 26-39 (2001).
2. Antoine, N., S. Bieniawski, I. Kroo, and D. H. Wolpert, "Fleet assignment using collective intelligence", *Proceedings of 42nd Aerospace Sciences Meeting*, (2004), AIAA-2004-0622.
3. Arthur, W. B., "Complexity in economic theory: Inductive reasoning and bounded rationality", *American Economic Review* **84**, 2 (May 1994), 406–411.
4. Aumann, R.J., and S. Hart, *Handbook of Game Theory with Economic Applications*, North-Holland Press (1992).
5. Axelrod, R., *The Evolution of Cooperation*, Basic Books NY (1984).
6. Basar, T., and G.J. Olsder, *Dynamic Noncooperative Game Theory*, Siam Philadelphia, PA (1999), Second Edition.
7. Bieniawski, S., and D. H. Wolpert, "Adaptive, distributed control of constrained multi-agent systems", *Proceedings of AAMAS 04*, (2004).
8. Boutilier, C., Y. Shoham, and M. P. Wellman, "Editorial: Economic principles of multi-agent systems", *Artificial Intelligence Journal* **94** (1997), 1–6.
9. Cover, T., and J. Thomas, *Elements of Information Theory*, Wiley-Interscience New York (1991).
10. Fudenberg, D., and D. K. Levine, *The Theory of Learning in Games*, MIT Press Cambridge, MA (1998).
11. Jaynes, E. T., "Information theory and statistical mechanics", *Physical Review* **106** (1957), 620.
12. Jaynes, E. T., and G. Larry Bretthorst, *Probability Theory : The Logic of Science*, Cambridge University Press (2003).
13. Judge, G., D. Miller, and W. Cho, "An information theoretic approach to ecological estimation and inference", *Ecological Inference:New methodological Strategies* (King, Rosen, and Tanner eds.), Cambridge University Press (2004).
14. Kahneman, D., "A psychological perspective on economics", *American Economic Review (Proceedings)* **93:2** (2003), 162–168.
15. Lee, C. Fan, and D. H. Wolpert, "Product distribution theory for control of multi-agent systems", *Proceedings of AAMAS 04*, (2004).
16. Macready, William, and David H. Wolpert, "Distributed constrained optimization with semi-coordinate transformations", submitted (2004).
17. Neyman, A., "Bounded complexity justifies cooperation in the finitely repeated prisoner's dilemma", *Economics Letters* **19** (1985), 227–230.
18. Osborne, M., and A. Rubenstein, *A Course in Game Theory*, MIT Press Cambridge, MA (1994).

19. Shamma, J.S., and G. Arslan, "Dynamic fictitious play, dynamic gradient play, and distributed convergence to nash equilibria", submitted (2004).
20. Sutton, R. S., and A. G. Barto, *Reinforcement Learning: An Introduction*, MIT Press Cambridge, MA (1998).
21. Tuyls, K., D. Heytens, A. Nowe, and B. Manderick, "Extended replicator dynamics as a key to reinforcement learning in multi-agent systems", *Lecture Notes in Artificial Intelligence, LNAI, (ECML 2003)*, (2003).
22. Tversky, A., and D. Kahneman, "Advances in prospect theory: Cumulative representation of uncertainty", *Journal of Risk and Uncertainty* **5** (1992), 297–323.
23. Verbeeck, K., A. Nowe, and K. Tuyls, "Coordinated exploration in stochastic common interest games", *Proceedings of AAMAS-3. University of Wales, Aberystwyth*, (2003).
24. Watkins, C., and P. Dayan, "Q-learning", *Machine Learning* **8**, 3/4 (1992), 279–292.
25. Wolpert, D. H., "Factoring a canonical ensemble", cond-mat/0307630.
26. Wolpert, David H., "Finding bounded rational equilibria part 1: Iterative focusing", *Proceedings of the International Society of Dynamic Games Conference, 2004*, (2004), in press.
27. Wolpert, D. H., "Information theory — the bridge connecting bounded rational game theory and statistical physics", *Complex Engineering Systems* (A. M. D. Braha and Y. Bar-Yam eds.), (2004).
28. Wolpert, D. H., and S. Bieniawski, "Adaptive distributed control: beyond single-instant categorical variables", *Proceedings of MSRAS04* (A. S. et al ed.), Springer Verlag (2004).
29. Wolpert, D. H., and S. Bieniawski, "Distributed control by lagrangian steepest descent", *Proceedings of CDC 04*, (2004).