

Notes on MLPS: A Model for Learning in Policy Space for Agents in Repeated Games

Steven O. Kimbrough
Draft working paper.

December 11, 2004

Note: This is a working paper and subject to potentially indefinite modification.

An Excel workbook, called 2by2markov.xls, accompanies this document. In that workbook, the sheet labeled 2x2Policies contains a simple model, for 2×2 stage games, played repeatedly by agents each with 2 policies in their consideration sets. In this document **Typewriter font** expressions correspond (unless otherwise noted) to labels and quantities in the workbook.

The simulation environment called Game22.jar, written by Ann Kuo under my direction, is available at <http://opim-sun.wharton.upenn.edu/~sok/workingmemory/Game22.zip>.

Contents

1	Context	3
2	Policy Space Learning Regimes for Repeated Play	3
3	The Basic MLPS Model	4
4	The Basic Model Algorithmically	5
4.1	Parameters, Formulas, and Other Values	5
5	Example: Stag Hunt	6
6	Further Illustrative Results with the Basic MLPS Model	10
6.1	Prisoner's Dilemma	10
6.2	Game #57	11
6.3	Constant Sum Games	12
6.3.1	Game #11	12
6.4	Coordination Games	13
7	Comments and Points Arising Regarding the Basic MLPS Model	14
7.1	Mathematical Properties: Markov Model	14
7.2	Interpretation of the Model in the Context of Games	15
8	Simulation Results for Relaxation of MLPS Model Assumptions	15
9	Larger Policy Spaces	16
10	Discussion	16

1 Context

Two players are to play a stage game repeatedly. Except through play, there can be no communication between the two players. Each player settles on a common (or at least similar) type of meta-strategy in approaching the super-game: from a *consideration set* of strategies (the elements of which we will call *policies*), attempt to learn which strategy works best by

1. sampling and playing policies from its consideration set,
2. recording the performances in actual play of the policies selected, and
3. favoring in subsequent play the better-performing policies.

Two (or n) agents playing in this manner may be said to engage in *Learning in Policy Space* (LPS). The purpose of this note is to elaborate, explore, and comment on a simple mathematical model—called the MLPS model—for learning in policy space (for play in repeated games).

The concluding section of the paper discusses briefly the significance of MLPS in the larger context of the theory of rationality. I want to emphasize, however, that the main aim and thrust of this note is an exercise in exploring the MLPS model. Larger issues are mostly deferred.

2 Policy Space Learning Regimes for Repeated Play

Consider the following informally specified, stylized *learning regime* for an agent confronted with an indefinitely repeated game. The agent begins by constructing a *consideration set*, \mathcal{P} , of *policies*. Learning proceeds by selecting a policy, $p \in \mathcal{P}$, taking into account its current (at t) attractiveness score, A_p^t . The agent then employs the policy in play for a number of rounds, and updates its evaluation score. At this point, policy selection recurs and the process is repeated.

We shall call learning regimes of this sort *policy space learning regimes*. Policies may be thought of as partial strategies in the game-theoretic sense: policies tell an agent how to play in the current round, and if adopted for the entire game would constitute strategies. Policies may rely on memory, on the history of the game; may be randomized in the sense that they employ a random number generator in choosing how to play; and in general policies may compute, reason, and learn to an arbitrary degree so long as they do not rely on unavailable information, do not require impossible computations, and do invariably direct a unique possible play in the game by the agent (within a specified time and resource limit).

In what follows we present and explore a Markov model for policy space learning. The final section discusses and interprets the results, and provides motivation—based on psychological, biological, and computational factors—for the plausibility of policy space learning regimes.

3 The Basic MLPS Model

1. Two players, row, R , and column, C , play a supergame, G .
2. The supergame, G , consists of an indefinitely long (possibly infinite) series of games, g_1, g_2, \dots
3. Each game, g_i , consists of n_e epochs.
4. Each epoch consists of l_e rounds of play.
5. Each round of play consists of one play of a 2×2 game, called the *stage* game, for example a Prisoner's Dilemma game, with the following strategy set and payoff structure for R and C :

	C_1	C_2
R_1	(R, R)	(S, T)
R_2	(T, S)	(P, P)

Note: As usual, $T > R > P > S \geq 0$ and $2R > T + S$.

6. Each player has a consideration set, S , consisting of two policies for play in each game. The two strategies are TIT-FOR-TAT (TFT), and ALWAYS DEFECT, (ALLD), i.e., always play R_2 or C_2 if you are respectively Row or Column.
7. At the end of each game, g_{t-1} , each player, p , picks a focal policy, f_p^t , from its consideration set, \mathcal{P} , for play in game g_t . The players choose independently, using the *fitness-proportional* choice rule.

Let $\text{EV}(p, i, j, k)$ be the expected value returned to player p for policy i , when p has focal policy j and $-p$ (the counter-player) has focal policy k . Then

$$\Pr(f_p^{t+1} = i | f_p^t = j, f_{-p}^t = k) = \text{EV}(p, i, j, k) / \sum_i \text{EV}(p, i, j, k) \quad (1)$$

fitness-
proportional
choice rule

That is, the probability that a player chooses a policy for focus in the next game is the proportion of value it returned, compared to all the player's policies, during the previous game.

8. During each game, each player p plays its focal strategy in $(1 - \varepsilon_p) \times 100$ percent of the rounds and its non-focal strategy in $\varepsilon_p \times 100$ percent of the rounds. (Assume for the simple, basic model that $\varepsilon_p = \varepsilon_{-p}$.) An oracle arranges the play so that in $(1 - \varepsilon)^2 \times 100$ percent of the rounds both players play their focal strategies, in $(1 - \varepsilon)\varepsilon \times 100$ percent of the rounds Row plays its focal strategy and Column plays its non-focal strategy, in $(1 - \varepsilon)\varepsilon \times 100$ percent of the rounds Column plays its focal strategy and Row plays its non-focal strategy, and in $\varepsilon^2 \times 100$ percent of the rounds both players play their non-focal strategies.

Alternatively, at the beginning of each epoch, $j = 1, 2, \dots, l_e$ in the i^{th} game, $e_{i,j}$, each player independently uses the ε -greedy rule to pick a policy for play throughout the epoch. With probability $(1 - \varepsilon)$ a player picks its focal strategy, p_e^p , and with probability ε a player picks its non-focal strategy. The length of the game, the number of epochs, l_e , is sufficiently high that the players achieve their expected returns with negligible error.

9. Throughout each game, each player records the rewards it receives when playing a given policy. At the conclusion of each game, each player, p calculates the attractiveness of each of the policies in its consideration set, i , as follows: $A_t^p = (\text{total reward received while playing policy } i) / (\text{the number of rounds played in which policy } i \text{ was in effect}) = \text{EV}(p, i, j, k)$.
10. Each player then chooses a focal strategy for the next game, using fitness-proportional choice on its A_t^p values.

4 The Basic Model Algorithmically

(Remember: This is a two-player model, Row and Column, but I'll try to abstract.)

4.1 Parameters, Formulas, and Other Values

Players The set of players.

$sp(i, j, k)$ Payoff in the stage game to player i if player 1 plays j and player 2 plays k .

\mathcal{P}^i Player i 's consideration set of policies. In a two-player stage game in strategic form Row is player 1 and Column is player 2. More generally, let *Players* be the set of players. Then $i = 1, 2, \dots, |\text{Players}|$. (We assume an enumeration of the players.)

Let `numPlayers` = $|\text{Players}|$.

$|\mathcal{P}^i|$ The number of policies in player i 's consideration set.

Let `numPolicies(i)` = $|\mathcal{P}^i|$.

p_j^i Policy j (such that $j = 1, 2, \dots, |\mathcal{P}^i|$) in player i 's consideration set ($i = 1, 2, \dots, |\text{Players}|$). (We assume an enumeration of the consideration set.)

Let `policy(i, j)` = p_j^i .

l_e The number of games played in an epoch.

Let `numGamesPerEpoch` = l_e .

ε_i Player i 's ε ; probability in a given epoch that player i does not play its focal strategy.

Let `epsilon(i)` = ε_i .

$r(i, g, h, l_e)$ Payoff (return) to player i in an epoch of l_e games, when player 1 uses policy g and player 2 plays policy h .

Let `return(i, g, h, numGamesPerEpoch)` = $r(i, g, h, l_e)$.

$\text{EV}(i, j, g, h)$ Expected value returned to player i for policy j , when i has focal policy g and $-i$ (counter-player) has focal policy h .

Let `ev(i, j, g, h)` = $\text{EV}(i, j, g, h)$.

f_p^t The focal policy of player p at time t .

Let `f(p, t)` = f_p^t .

$\Pr(f_p^t = i)$ The probability that at period t , player p has as its focal policy, policy $i \in \mathcal{P}^p$.

Now we have two key equations:

$$\Pr(f_p^{t+1} = i | f_p^t = j, f_{-p}^t = k) = \text{EV}(p, i, j, k) / \sum_i \text{EV}(p, i, j, k) \quad (2)$$

And since the players pick their focal strategies independently:

$$\Pr(f_p^{t+1} = i, f_{-p}^{t+1} = h | f_p^t = j, f_{-p}^t = k) = \Pr(f_p^{t+1} = i | f_p^t = j, f_{-p}^t = k) \cdot \Pr(f_{-p}^{t+1} = h | f_p^t = j, f_{-p}^t = k) \quad (3)$$

Let $\text{Prob1}(i, j, k) = \Pr(f_1^{t+1} = i | f_1^t = j, f_2^t = k)$ and $\text{Prob2}(i, j, k) = \Pr(f_2^{t+1} = i | f_1^t = j, f_2^t = k)$.

5 Example: Stag Hunt

The purpose of this section is to illustrate the basic MLPS model using Stag Hunt as an example stage game.

	Hunt Stag (C1)	Hunt Hare (C2)
Hunt Stag (R1)	(R, R) $(4, 4)$	(S, T) $(0, 3)$
Hunt Hare (R2)	(T, S) $(3, 0)$	(P, P) $(1, 1)$

Figure 1: Stag Hunt stage game, with example payoffs. R = reward for cooperation. T = temptation. P = penalty for mutual lack of cooperation. S = sucker's payoff.

We shall assume that each player, Row and Column, has two policies in its consideration set: TIT FOR TAT (TFT) and ALWAYS DEFECT (ALLD). TIT FOR TAT hunts stag (cooperates) on the first round of play in each epoch in which the policy is in force. After that, it mimics the counter-player's play on the previous round. The policy of ALWAYS DEFECT hunts hare on every round of play in each epoch in which it is in effect. There are consequently four states in the process, corresponding to the points in $\mathcal{P}_R \times \mathcal{P}_C$. The states, and their places in the enumeration we shall use, are as follows:

1. Row has TFT as its focal policy for the current game and Column has TFT as its focal policy for the current game. Call this state 1, alternatively, state (1,1).
2. Row has TFT as its focal policy for the current game and Column has ALLD as its focal policy for the current game. Call this state 2, alternatively, state (1,2).
3. Row has ALLD as its focal policy for the current game and Column has TFT as its focal policy for the current game. Call this state 3, alternatively, state (2,1).

4. Row has ALLD as its focal policy for the current game and Column has ALLD as its focal policy for the current game. Call this state 4, alternatively, state (2,2).

(Row player is coded as player 1; Column player as player 2.) Assuming the payoffs as in Table 1 and $l_e = 10$ rounds per epoch the returns to Row in each of the four states are as follows.

1. $\text{return}(1,1,1,10) = l_e R = 40$.
2. $\text{return}(1,1,2,10) = (l_e - 1)P + S = 9$.
3. $\text{return}(1,2,1,10) = T + (l_e - 1)P = 12$.
4. $\text{return}(1,2,2,10) = l_e P = 10$.
5. $\text{return}(2,1,1,10) = l_e R = 40$.
6. $\text{return}(2,1,2,10) = T + (l_e - 1)P = 12$.
7. $\text{return}(2,2,1,10) = (l_e - 1)P + S = 9$.
8. $\text{return}(2,2,2,10) = l_e P = 10$.

	$1 - \varepsilon$: TFT	ε : ALLD	Total
$1 - \varepsilon$: TFT	$(1 - \varepsilon)(l_e R)$ $(1 - \varepsilon)40$	$\varepsilon((l_e - 1)P + S)$ 9ε	$40 - 31\varepsilon$
ε : ALLD	$(1 - \varepsilon)(T + (l_e - 1)P)$ $12(1 - \varepsilon)$	$\varepsilon(l_e P)$ 10ε	$12 - 2\varepsilon$

Table 1: State 1: Payoffs to Row in Stag Hunt example when the system is in state 1 and $l_e = 10$.

	ε : TFT	$1 - \varepsilon$: ALLD	Total
$1 - \varepsilon$: TFT	$\varepsilon(l_e R)$ $\varepsilon 40$	$(1 - \varepsilon)((l_e - 1)P + S)$ $9(1 - \varepsilon)$	$9 + 31\varepsilon$
ε : ALLD	$\varepsilon(T + (l_e - 1)P)$ 12ε	$(1 - \varepsilon)(l_e P)$ $10(1 - \varepsilon)$	$10 + 2\varepsilon$

Table 2: State 2: Payoffs to Row in Stag Hunt example when the system is in state 2 and $l_e = 10$.

	$1 - \varepsilon$: TFT	ε : ALLD	Total
ε : TFT	$(1 - \varepsilon)(l_e R)$ $(1 - \varepsilon)40$	$\varepsilon((l_e - 1)P + S)$ 9ε	$40 - 31\varepsilon$
$1 - \varepsilon$: ALLD	$(1 - \varepsilon)(T + (l_e - 1)P)$ $12(1 - \varepsilon)$	$\varepsilon(l_e P)$ 10ε	$12 - 2\varepsilon$

Table 3: State 3: Payoffs to Row in Stag Hunt example when the system is in state 3 and $l_e = 10$.

	ε : TFT	$1 - \varepsilon$: ALLD	Total
ε : TFT	$\varepsilon(l_e R)$ $\varepsilon 40$	$(1 - \varepsilon)((l_e - 1)P + S)$ $9(1 - \varepsilon)$	$9 + 31\varepsilon$
$1 - \varepsilon$: ALLD	$\varepsilon(T + (l_e - 1)P)$ 12ε	$(1 - \varepsilon)(l_e P)$ $10(1 - \varepsilon)$	$10 + 2\varepsilon$

Table 4: State 4: Payoffs to Row in Stag Hunt example when the system is in state 4 and $l_e = 10$.

Recall that $\text{ev}(i, j, k, 1) =$ the expected value returned to player i during an epoch, when player i plays strategy j , player 1 has focal strategy k and player 2 has focal strategy 1. Thus from Tables 1–4 and from the symmetry of the Stag Hunt game we have the following.

1. $\text{ev}(1, 1, 1, 1) = 40 - 31\varepsilon_2 = 36.9$ at $\varepsilon_2 = 0.1$.
2. $\text{ev}(1, 2, 1, 1) = 12 - 2\varepsilon_2 = 11.8$ at $\varepsilon_2 = 0.1$.
3. $\text{ev}(1, 1, 1, 2) = 9 + 31\varepsilon_2 = 12.1$ at $\varepsilon_2 = 0.1$.
4. $\text{ev}(1, 2, 1, 2) = 10 + 2\varepsilon_2 = 10.2$ at $\varepsilon_2 = 0.1$.
5. $\text{ev}(1, 1, 2, 1) = \text{ev}(1, 1, 1, 1) = 36.9$ at $\varepsilon_2 = 0.1$, since 1's focal policy is immaterial for this calculation.
6. $\text{ev}(1, 2, 2, 1) = \text{ev}(1, 2, 1, 1) = 11.8$ at $\varepsilon_2 = 0.1$, since 1's focal policy is immaterial for this calculation.
7. $\text{ev}(1, 1, 2, 2) = \text{ev}(1, 1, 1, 2) = 12.1$ at $\varepsilon_2 = 0.1$, since 1's focal policy is immaterial for this calculation.
8. $\text{ev}(1, 2, 2, 2) = \text{ev}(1, 2, 1, 2) = 10.2$ at $\varepsilon_2 = 0.1$, since 1's focal policy is immaterial for this calculation.

Similarly, for Column we have:

1. $\text{ev}(2, 1, 1, 1) = 40 - 31\varepsilon_2 = 36.9$ at $\varepsilon_1 = 0.1$.
2. $\text{ev}(2, 2, 1, 1) = 12 - 2\varepsilon_2 = 11.8$ at $\varepsilon_1 = 0.1$.
3. $\text{ev}(2, 1, 1, 2) = 40 - 31\varepsilon_2 = 36.9$ at $\varepsilon_1 = 0.1$.
4. $\text{ev}(2, 2, 1, 2) = 12 - 2\varepsilon_2 = 11.8$ at $\varepsilon_1 = 0.1$.
5. $\text{ev}(2, 1, 2, 1) = 9 + 31\varepsilon_2 = 12.1$ at $\varepsilon_1 = 0.1$.
6. $\text{ev}(2, 2, 2, 1) = 10 + 2\varepsilon_2 = 10.2$ at $\varepsilon_2 = 0.1$.
7. $\text{ev}(2, 1, 2, 2) = 9 + 31\varepsilon_2 = 12.1$ at $\varepsilon_1 = 0.1$.
8. $\text{ev}(2, 2, 2, 2) = 10 + 2\varepsilon_2 = 10.2$ at $\varepsilon_2 = 0.1$.

Now we calculate the $\Pr(f_p^{t+1} = i | f_1^t = j, f_2^t = k)$ values for $p \in \{1, 2\}$. First, for player 1:

1. $\Pr(f_1^{t+1} = 1 | f_1^t = 1, f_2^t = 1) = \text{ev}(1, 1, 1, 1) / (\text{ev}(1, 1, 1, 1) + \text{ev}(1, 2, 1, 1)) = 36.9 / (36.9 + 11.8) = 0.757700205 = \text{Prob1}(1, 1, 1)$.
2. $\Pr(f_1^{t+1} = 2 | f_1^t = 1, f_2^t = 1) = \text{ev}(1, 2, 1, 1) / (\text{ev}(1, 1, 1, 1) + \text{ev}(1, 2, 1, 1)) = 11.8 / (36.9 + 11.8) = 0.242299795 = \text{Prob1}(2, 1, 1) = 1 - \Pr(f_1^{t+1} = 1 | f_1^t = 1, f_2^t = 1)$.

3. $\Pr(f_1^{t+1} = 1 | f_1^t = 1, f_2^t = 2) = \text{ev}(1, 1, 1, 2) / (\text{ev}(1, 1, 1, 2) + \text{ev}(1, 2, 1, 2)) = 12.1 / (12.1 + 10.2) = 0.542600897 = \text{Prob1}(1, 1, 2)$.
4. $\Pr(f_1^{t+1} = 2 | f_1^t = 1, f_2^t = 2) = \text{ev}(1, 2, 1, 2) / (\text{ev}(1, 1, 1, 2) + \text{ev}(1, 2, 1, 2)) = 10.2 / (12.1 + 10.2) = 0.457399103 = \text{Prob1}(2, 1, 2) = 1 - \Pr(f_1^{t+1} = 1 | f_1^t = 1, f_2^t = 2)$.
5. $\Pr(f_1^{t+1} = 1 | f_1^t = 2, f_2^t = 1) = \text{ev}(1, 1, 2, 1) / (\text{ev}(1, 1, 2, 1) + \text{ev}(1, 2, 2, 1)) = 36.9 / (36.9 + 11.8) = 0.757700205 = \text{Prob1}(1, 2, 1)$.
6. $\Pr(f_1^{t+1} = 2 | f_1^t = 2, f_2^t = 1) = \text{ev}(1, 2, 1, 2) / (\text{ev}(1, 1, 1, 2) + \text{ev}(1, 2, 1, 2)) = 11.8 / (36.9 + 11.8) = 0.242299795 = \text{Prob1}(2, 2, 1) = 1 - \Pr(f_1^{t+1} = 1 | f_1^t = 2, f_2^t = 1)$.
7. $\Pr(f_1^{t+1} = 1 | f_1^t = 2, f_2^t = 2) = \text{ev}(1, 1, 2, 2) / (\text{ev}(1, 1, 2, 2) + \text{ev}(1, 2, 2, 2)) = 12.1 / (12.1 + 10.2) = 0.542600897 = \text{Prob1}(1, 2, 2)$.
8. $\Pr(f_1^{t+1} = 2 | f_1^t = 2, f_2^t = 2) = \text{ev}(1, 2, 2, 2) / (\text{ev}(1, 1, 2, 2) + \text{ev}(1, 2, 2, 2)) = 10.2 / (12.1 + 10.2) = 0.457399103 = \text{Prob1}(2, 2, 2) = 1 - \Pr(f_1^{t+1} = 1 | f_1^t = 2, f_2^t = 2)$.

Now for player 2:

1. $\Pr(f_2^{t+1} = 1 | f_1^t = 1, f_2^t = 1) = \text{ev}(2, 1, 1, 1) / (\text{ev}(2, 1, 1, 1) + \text{ev}(2, 2, 1, 1)) = 36.9 / (36.9 + 11.8) = 0.757700205 = \text{Prob2}(1, 1, 1)$.
2. $\Pr(f_2^{t+1} = 2 | f_1^t = 1, f_2^t = 1) = \text{ev}(2, 2, 1, 1) / (\text{ev}(2, 1, 1, 1) + \text{ev}(2, 2, 1, 1)) = 11.8 / (36.9 + 11.8) = 0.242299795 = \text{Prob2}(2, 1, 1) = 1 - \Pr(f_2^{t+1} = 1 | f_1^t = 1, f_2^t = 1)$.
3. $\Pr(f_2^{t+1} = 1 | f_1^t = 1, f_2^t = 2) = \text{ev}(2, 1, 1, 2) / (\text{ev}(2, 1, 1, 2) + \text{ev}(2, 2, 1, 2)) = 36.9 / (36.9 + 11.8) = 0.757700205 = \text{Prob2}(1, 1, 2)$.
4. $\Pr(f_2^{t+1} = 2 | f_1^t = 1, f_2^t = 2) = \text{ev}(2, 2, 1, 2) / (\text{ev}(2, 1, 1, 2) + \text{ev}(2, 2, 1, 2)) = 11.8 / (36.9 + 11.8) = 0.242299795 = \text{Prob2}(2, 1, 2) = 1 - \Pr(f_2^{t+1} = 1 | f_1^t = 1, f_2^t = 2)$.
5. $\Pr(f_2^{t+1} = 1 | f_1^t = 2, f_2^t = 1) = \text{ev}(2, 1, 2, 1) / (\text{ev}(2, 1, 2, 1) + \text{ev}(2, 2, 2, 1)) = 12.1 / (12.1 + 10.2) = 0.542600897 = \text{Prob2}(1, 2, 1)$.
6. $\Pr(f_2^{t+1} = 2 | f_1^t = 2, f_2^t = 1) = \text{ev}(2, 2, 2, 1) / (\text{ev}(2, 1, 2, 1) + \text{ev}(2, 2, 2, 1)) = 10.2 / (12.1 + 10.2) = 0.457399103 = \text{Prob2}(2, 2, 1) = 1 - \Pr(f_2^{t+1} = 1 | f_1^t = 2, f_2^t = 1)$.
7. $\Pr(f_2^{t+1} = 1 | f_1^t = 2, f_2^t = 2) = \text{ev}(2, 1, 2, 2) / (\text{ev}(2, 1, 2, 2) + \text{ev}(2, 2, 2, 2)) = 12.1 / (12.1 + 10.2) = 0.542600897 = \text{Prob2}(1, 2, 2)$.
8. $\Pr(f_2^{t+1} = 2 | f_1^t = 2, f_2^t = 2) = \text{ev}(2, 2, 2, 2) / (\text{ev}(2, 1, 2, 2) + \text{ev}(2, 2, 2, 2)) = 10.2 / (12.1 + 10.2) = 0.457399103 = \text{Prob2}(2, 2, 2) = 1 - \Pr(f_2^{t+1} = 1 | f_1^t = 2, f_2^t = 2)$.

Extracting the values from Table 5 produces a state transition matrix, $\mathbf{P} = \{p_{i,j}\}$, for a Markov chain. At convergence: (0.4779 0.2134 0.2134 0.0953). So 90%+ of the time at least one agent is playing TFT. Note the expected take for Row per epoch by state:

1. $(1 - \varepsilon)(40 - 31\varepsilon) + \varepsilon(12 - 2\varepsilon) = 34.39$

	s(1)=(1,1)	s(2)=(1,2)	s(3)=(2,1)	s(4)=(2,2)
s(1)	Prob1(1,1,1)* Prob2(1,1,1) $0.7577 \cdot 0.7577$ = 0.5741	Prob1(1,1,1)* Prob2(2,1,1) $0.7577 \cdot 0.2423$ = 0.1836	Prob1(2,1,1)* Prob2(1,1,1) $0.2423 \cdot 0.7577$ = 0.1836	Prob1(2,1,1)* Prob2(2,1,1) $0.2423 \cdot 0.2423$ = 0.0587
s(2)	Prob1(1,1,2)* Prob2(1,1,2) $0.5426 \cdot 0.7577$ = 0.4111	Prob1(1,1,2)* Prob2(2,1,2) $0.5426 \cdot 0.2423$ = 0.1315	Prob1(2,1,2)* Prob2(1,1,2) $0.4574 \cdot 0.7577$ = 0.3466	Prob1(2,1,2)* Prob2(2,1,2) $0.4574 \cdot 0.2423$ = 0.1108
s(3)	Prob1(1,2,1)* Prob2(1,2,1) $0.7577 \cdot 0.5426$ = 0.4111	Prob1(1,2,1)* Prob2(2,2,1) $0.7577 \cdot 0.4574$ = 0.3466	Prob1(2,2,1)* Prob2(1,2,1) $0.2423 \cdot 0.5426$ = 0.1315	Prob1(2,2,1)* Prob2(2,2,1) $0.2423 \cdot 0.4574$ = 0.1108
s(4)	Prob1(1,2,2)* Prob2(1,2,2) $0.5426 \cdot 0.5426$ = 0.2944	Prob1(1,2,2)* Prob2(2,2,2) $0.5426 \cdot 0.4574$ = 0.2482	Prob1(2,2,2)* Prob2(1,2,2) $0.4574 \cdot 0.5426$ = 0.2482	Prob1(2,2,2)* Prob2(2,2,2) $0.4574 \cdot 0.4574$ = 0.2092

Table 5: Stag Hunt transition matrix data assuming fitness proportional policy selection by both players, based on previous Tables 1–4. Numeric example for $\varepsilon = 0.1 = \varepsilon_1 = \varepsilon_2$.

2. $(1 - \varepsilon)(9 + 31\varepsilon) + \varepsilon(10 + 2\varepsilon) = 11.91$
3. $\varepsilon(40 - 31\varepsilon) + (1 - \varepsilon)(12 - 2\varepsilon) = 14.31$
4. $\varepsilon(9 + 31\varepsilon) + (1 - \varepsilon)(10 + 2\varepsilon) = 10.39$

Further in expectation, Row (and Column, too) gets

$$(0.4779 \ 0.2134 \ 0.2134 \ 0.0953) \cdot (34.39 \ 11.91 \ 14.31 \ 10.39)' = 23.02$$

(per epoch of length $l_e = 10$, or 2.302 per round of play), much better than the 10.39 both would get if they played ALLD with ε -greedy exploration. Note that even the latter is larger than the return, 10 per epoch or 1 per round, of settling on the risk-dominant outcome of mutually hunting hare. There is a third, mixed, equilibrium of the one-shot Stag Hunt game. For this example it occurs at $((\frac{1}{2}R1, \frac{1}{2}R2), (\frac{1}{2}C1, \frac{1}{2}C2))$. At this equilibrium each player can expect a return of 2 from a round of play.

6 Further Illustrative Results with the Basic MLPS Model

6.1 Prisoner's Dilemma

Under our standard conditions— $l_e = 10$ and $\varepsilon = \varepsilon_1 = \varepsilon_2 = 0.1$ —the limiting distribution of strategies is

$$\vec{\beta} = (0.374192023 \ 0.237520346 \ 0.237520346 \ 0.150767284)$$

	Cooperate	Defect
Cooperate	(R, R) $(3, 3)$	(S, T) $(0, 5)$
Defect	(T, S) $(5, 0)$	(P, P) $(1, 1)$

Figure 2: Canonical Prisoner’s Dilemma stage game, with example payoffs. R = reward for cooperation. T = temptation. P = penalty for mutual lack of cooperation. S = sucker’s payoff.

and the expected return per round of play for each player is 1.767317655, well above the value of 1 to be expected if both play their dominant single-shot strategy.

If we increase the value of mutual cooperation from 3 to 4 we get

$$\vec{\beta} = (0.44398415 \ 0.222337206 \ 0.222337206 \ 0.111341437)$$

and the expected return per round of play for each player is 2.271972563. With a larger reward for cooperation we get more cooperation and both players take home more value.

If we keep Row’s reward for mutual cooperation, R^{Row} at 4 but set Column’s to 3 we get

$$\vec{\beta} = (0.406648096 \ 0.250505999 \ 0.212153641 \ 0.130692263)$$

with Row taking on average 2.180738598 per round of play and Column taking 1.823748293.

6.2 Game #57

	Safe	Risky
Safe	$(r_{1,1}, c_{1,1})$ $(2, 3)$	$(r_{1,2}, c_{1,2})$ $(4, 2)$
Risky	$(r_{2,1}, c_{2,1})$ $(1, 1)$	$(r_{2,2}, c_{2,2})$ $(3, 4)$

Figure 3: Canonical Game #57 stage game, with example payoffs.

Game #57 (see [RGG76]) is like Prisoner’s Dilemma in that the unique Nash equilibrium—play (Safe, Safe) for a reward, here, of $(2, 3)$ —is not Pareto optimal. In this game there are two Pareto optimal outcomes: (Safe, Risky) and (Risky, Risky). Unlike Prisoner’s Dilemma, Game #57 is asymmetric. What will happen in repeated play? Under the basic MLPS model ($l_e = 10, \varepsilon = 0.1$, consideration sets of TfT (=1) and ALLD (=2)) the limiting state distribution is:

State:	1=(1,1)	2=(1,2)	3=(2,1)	4=(2,2)
Probability:	0.215978	0.251182	0.246343	0.286497

The per round expected payoffs are 2.75 for Row and 3.65 for Column. Both players do better than they would at the (one-shot) Nash equilibrium.

6.3 Constant Sum Games

6.3.1 Game #11

	C1	C2
R1	$(r_{1,1}, c_{1,1})$ (2, 3)	$(r_{1,2}, c_{1,2})$ (4, 1)
R2	$(r_{2,1}, c_{2,1})$ (1, 4)	$(r_{2,2}, c_{2,2})$ (3, 2)

Figure 4: Game #11, a constant sum stage game, with example payoffs.

Game #11 (see [RGG76], there called 11-ORDINAL) is constant-sum with a single Nash equilibrium (R1, C2). Because the game is constant-sum, every outcome is Pareto optimal.

What will happen in repeated play? Under the basic MLPS model ($l_e = 10, \varepsilon = 0.1$, consideration sets of TFT (=1) and ALLD (=2)) the limiting state distribution is:

State:	1=(1,1)	2=(1,2)	3=(2,1)	4=(2,2)
Probability:	0.260306	0.199272	0.306097	0.234326

The per round expected payoffs are 2.703 for Row and 2.297 for Column. In consequence of the game's being constant sum it is impossible for both players to do better than they would at the Nash equilibrium. According to our MLPS model, Row has a stronger position in the game than does Column, in direct conflict with the (one-shot) Nash equilibrium. This an artifact of the limited consideration sets these agents are using. For example, if we exchange the columns we get what should be an equivalent game:

	C2	C1
R1	$(r_{1,1}, c_{1,1})$ (4, 1)	$(r_{1,2}, c_{1,2})$ (2, 3)
R2	$(r_{2,1}, c_{2,1})$ (3, 2)	$(r_{2,2}, c_{2,2})$ (1, 4)

Figure 5: Game #11, a constant sum stage game, with example payoffs.

Notice that the one-shot Nash equilibrium remains (R1, C1). The limiting state distribution now is:

State:	1=(1,1)	2=(1,2)	3=(2,1)	4=(2,2)
Probability:	0.208329	0.427997	0.119065	0.244610

The per round expected payoffs are 1.728 for Row and 3.272 for Column. Here, at least, and not surprisingly, the MLPS model is sensitive to the contents of the consideration sets.

6.4 Coordination Games

	C1	C2
R1	$(r_{1,1}, c_{1,1})$ (1, 1)	$(r_{1,2}, c_{1,2})$ (0, 0)
R2	$(r_{2,1}, c_{2,1})$ (0, 0)	$(r_{2,2}, c_{2,2})$ (1, 1)

Figure 6: Balanced coordination stage game, with example payoffs.

There are two single-shot Nash equilibria in pure strategies, (R1, C1) and (R2, C2), and a mixed equilibrium at $((\frac{1}{2}R1, \frac{1}{2}R2), (\frac{1}{2}C1, \frac{1}{2}C2))$. Here both (R1, C1) and (R2, C2) are Pareto optimal. Under the basic MLPS model ($l_e = 10, \varepsilon = 0.1$, consideration sets of TFT (=1) and ALLD (=2)) the limiting state distribution is:

State:	1=(1,1)	2=(1,2)	3=(2,1)	4=(2,2)
Probability:	0.250000	0.250000	0.250000	0.250000

The per round expected payoffs are 0.950 for Row and 0.950 for Column. This is quite close to the expected return from the two pure strategy Nash equilibria (in the one-shot case) and much higher than the 0.5 expected from the mixed-strategy Nash equilibrium.

	C1	C2
R1	$(r_{1,1}, c_{1,1})$ (4, 4)	$(r_{1,2}, c_{1,2})$ (0, 0)
R2	$(r_{2,1}, c_{2,1})$ (0, 0)	$(r_{2,2}, c_{2,2})$ (1, 1)

Figure 7: Unbalanced coordination stage game, with example payoffs.

Here there are two single-shot Nash equilibria, (R1, C1) and (R2, C2), and a mixed equilibrium at $((\frac{1}{5}R1, \frac{4}{5}R2), (\frac{1}{5}C1, \frac{4}{5}C2))$. The single Pareto optimal outcome is (R1, R1). This game might be seen as a hyper Stag Hunt. Will the agents learn to cooperate? Under the basic MLPS model ($l_e = 10, \varepsilon = 0.1$, consideration sets of TFT (=1) and ALLD (=2)) the limiting state distribution is:

State:	1=(1,1)	2=(1,2)	3=(2,1)	4=(2,2)
Probability:	0.540909	0.194556	0.194556	0.069979

The per round expected payoffs are 2.379 for Row and 2.379 for Column. Note that if $\varepsilon = 0.01$ the payoffs become 2.447 each. In general, the limiting distribution is not sensitive to the values of ε_1 and ε_2 .

Note further that the expected payoff to a player at the mixed Nash equilibrium is 0.8, much lower. Notice the rather odd property here of the Nash equilibrium in mixed strategies that as

the outcomes become more and more unbalanced the shift in probability goes more and more to the *lower* return outcomes. Mathematically it is transparent why the Nash equilibrium has this property: In calculating its mixture of play, a player takes into account only the other player’s payoffs, not its own. This illustrates how the Nash equilibrium is inherently an adversarial, or at least pessimistic, concept. It assumes the worst and does not permit venture for reward. On the other hand, players governed by the MLPS model are able to take into account both their own and the counter-player’s payoffs and grope in the direction of a mutually-beneficial outcome.

7 Comments and Points Arising Regarding the Basic MLPS Model

The basic model given above—the Basic MLPS (Markov Learning in Policy Space) model—is straightforwardly generalized to games larger than 2×2 , and policy consideration sets larger than 2. Even so, much of what can be said here about the present restricted family of cases will apply to generalized versions of the model.

7.1 Mathematical Properties: Markov Model

The Markov model induced by the basic model, illustrated through Table 5, will in general¹ be ergodic (every state is reachable from each state) and aperiodic (from each state there is a positive probability of reaching any state in the next period). Specifically, in \mathbf{P} , $p_{i,j} > 0$. In short the resulting Markov chain is said to be *regular*, i.e., its states form an acyclic ergodic set. This observation occasions a number of comments:

1. The states of the super-game are determined by the starting state (which focal policies the players choose at the beginning of play) and \mathbf{P} . Thus, for example, if $\vec{\alpha}$ is the initial distribution of states of play (e.g., play starts in state 2, so $\vec{\alpha} = (0 \ 1 \ 0 \ 0)$), then the distribution of states at the end of game 1 (or rather at the start of play in game 2) is $\vec{\alpha}\mathbf{P}$. After n games the state of the system is described by $\vec{\alpha}\mathbf{P}^n$. As $n \rightarrow \infty$ the distribution of states of the system *converges uniquely* to $\vec{\beta}$, where $\vec{\beta}\mathbf{P} = \vec{\beta}$ and $\vec{\beta} \cdot \vec{e} = 1$ (where \vec{e} is a vector of 1s and is of length conforming to the requirements to hand).
2. Markov models of this sort—regular: ergodic and acyclic—in general:
 - are robust (small changes to \mathbf{P} result in small changes to $\vec{\beta}$), and
 - show rapid (geometric) rates of convergence.
3. $\vec{\beta}$, the convergent distribution states, will depend upon, will in part be a function of, the actual payoff amounts in the underlying stage game. For example if R the reward for cooperation in the Stag Hunt example (Table) is changed from 4 to 10 (for both players) the distribution of states at convergence changes from (0.4779 0.2134 0.2134 0.0953) to (0.7186 0.1291 0.1291 0.0232).

¹With the exception of transition matrices built via the modeling procedure from strategic form stage game representations in which for at least one player all outcomes yield a payoff of 0. We exclude this case.

7.2 Interpretation of the Model in the Context of Games

1. The unique limiting distribution of states of an MLPS model is to be contrasted with the Nash equilibrium as a solution concept for repeated games and replicator dynamics predictions of game dynamics.

The Folk Theorem(s) of game theory tells us that in an indefinitely repeated game, nearly any series of outcomes can be supported by a Nash equilibrium. Thus, the Nash equilibrium becomes a useless predictor for repeated games.

In the replicator dynamics it is typical that there are multiple “basins of attraction.” Depending upon where the system starts, it could converge to any of many different configurations.

2. The limiting state distribution of an MLPS model does *not* in general correspond to a Nash equilibrium in mixed strategies for the stage game.

To illustrate, at convergence of the MLPS model ($l_e = 10, \varepsilon = 0.1$) for Prisoner’s Dilemma as in Figure 2, each player is playing ALWAYS DEFECT as its focal strategy with probability=0.39, but there is no mixed strategy equilibrium in the one-shot Prisoner’s Dilemma game.

3. Agents, players as modeled by the MLPS model, do not ‘know’ which state they are in. Each player picks its own focal policy at the beginning of a game. Neither player ‘knows’ what focal policy the other player will choose or has chosen. The system proceeds regularly from state to state with the players forever in ignorance of which states the system passes through, or even what the states are.
4. Players need know very little under an MLPS model. Specifically, no knowledge is required of the payoffs of the stage game, of any consideration set of policies other than one’s own, and so forth. Agents with only very minimal cognitive capabilities are required.

8 Simulation Results for Relaxation of MLPS Model Assumptions

/* More Java code needed.! see also Ann Kuo’s code. Discuss details with her. */

Simulation results indicate that the MLPS model is also robust to departures from its key assumptions. If, for example, players independently choose game and epoch lengths, perhaps with a random component, consider all eight memory-1 policies (coded as 000, 001, 010, 100, 111, 110, 101, 011) and keep a running account of the attractivenesses of the policies in their consideration sets, then the resulting behavior is similar to that under the MLPS model. The simulation program Game22.jar makes the calculations.

To illustrate, recall the Stag Hunt game in Figure 1, page 6. In that game simulated in Game22.jar each player robustly converges to a per round take of 3.7 (or more). In the canonical Prisoner’s Dilemma, Figure 2, page 11, the players robustly converge to about 2.6 (or more) each. In the canonical Game #57, Figure 3, page 11, the players converge robustly to about 2.9 for Row and 3.5 for Column. In Game #11, Figure 4, page 12, the players converge robustly to just over 2 for Row and just under 3 for Column. This, for once (the game is constant-sum) is very close to the single-shot Nash equilibrium. /* Is this provable in general? */

In the Balanced Coordination Game, Figure 6, page 13, the players converge robustly to average payoffs of nearly 0.9 each. In the Unbalanced Coordination Game, Figure 7, page 13, the player converge robustly to about 3.47 each.

9 Larger Policy Spaces

/* Java code needed for this. Examine behavior of, first, 2x2 games under, e.g., all 8 memory-1 strategies. */

10 Discussion

Points in summary:

1. An exploring rationality might be described as *maximum-seeking* rather than *maximum-taking* as in Rational Choice Theory. Exploring Rationality Theory does not assume an arrayed set of known choices of known values with known probabilities. It is in principle applicable to a broader range of circumstances than has heretofore been addressed with RCT.
2. The MLPS model is a model of an *exploring rationality* [Kim04], a rationality that countenances uncertainty in pursuit of reward. (Recall: a mixed-strategy Nash equilibrium may well be the risk dominant equilibrium.) Here, I use *risk* in contradistinction to uncertainty. Both are technical terms in the decision analysis literature.

One extreme possibility we know how to treat—namely, risk. In that case a probability distribution over the set of states is known—or, better yet, the decision maker deems it suitable to act as if it were known. [LR57, page 277]

3. A possible objection to ERT and the MLPS model in particular is that it sweeps under the rug assumptions about risk. In response:
 - (a) It is true that for every particular MLPS model there is *some* assumption regarding risk (probability distributions on outcomes) with which it is consistent. But for every outcome of reasoned deliberation there is some assumption regarding prejudice, hallucination, etc. that could explain the outcome.
 - (b) Any other model, including the Nash equilibrium, is subject to the same objection or observation. What makes the Nash assumptions sacrosanct? Why are these extreme positions on venturing in the face of uncertainty always correct?
4. The MLPS model has a number of attractive features. It is a model of exploring rationality. It gives reasonable results across a broad range of stage games in repeated play. While it makes strong, and ultimately implausible, assumptions, it is robust and may serve as a useful basis for refinement.
5. /* At some point I need to dig up Axelrod's list of desiderata: nice, forgiving, provocable, etc. Compare these to MLPS. What other desiderata are appropriate? */

Future work:

1. Qualitative predictions. MLPS models are responsive to—and can predict and explain—changes in stage game payoffs. For example, as the reward for mutual cooperation approaches the temptation to defect in Prisoner’s Dilemma, the (basic) MLPS model predicts that mutual cooperation will increase.

This is perhaps the top agenda item for empirical testing of the MLPS model.

2. Might it be possible to discover empirically whether subjects plausibly have consideration sets of policies and if so how they play them?
3. Systematic exploration of plausible consideration sets for stage games of import is also very high on the research agenda.

References

- [Kim04] Steven O. Kimbrough, *A note on exploring rationality in games*, Working paper, University of Pennsylvania, Philadelphia, PA, March 2004, Presented at SEP (Society of Exact Philosophy), spring 2004.
- [LR57] R. Duncan Luce and Howard Raiffa, *Games and decisions*, John Wiley, New York, NY, 1957, Reprinted by Dover Books, 1989.
- [RGG76] Anatol Rapoport, Melvin J. Guyer, and David G. Gordon, *The 2×2 game*, The University of Michigan Press, Ann Arbor, MI, 1976.