Reinforcement Learning Policy Iterations for Nash Differential Games

Kanchu Kiran
kanchu.kiran@rutgers.edu

***Abstract*— Incorporating the foundational methodologies introduced by Li and Gajic [1], this study presents an in-depth examination of reinforcement learning within the framework of Nash differential games. This project integrates reinforcement learning with Nash differential games, applying policy iteration algorithms to identify optimal strategies within complex, dynamic environments. Leveraging Lyapunov iterations for solving Coupled Algebraic Riccati equations, it offers a computational framework for achieving stabilizing solutions in multi-agent scenarios. By iterating towards equilibrium strategies reflective of both cooperative and competitive interactions, this work embodies a synthesis of game theory, control theory, and applied mathematics. It presents a feasible, useful, and novel contribution to the strategic analysis in fields like economics and strategic planning, demonstrating the practical applicability of theoretical constructs through MATLAB simulations.**

I. Project Description

This project explores Nash differential games, focusing on the dynamics of multiple players whose decisions impact each other over time, highlighting the complexity of non-zero sum games found in economic and strategic scenarios. Utilizing policy iteration algorithms for strategy optimization, the research employs Lyapunov iterations for solving coupled algebraic Riccati equations, enhancing the identification of stabilizing solutions and Nash equilibria. The study, grounded in system dynamics and control actions, leverages MATLAB for critical analysis, aiming at advancing the understanding of Nash differential games through policy and Lyapunov iterations, with significant implications for strategic decision-making in various domains. The methodology incorporates Lyapunov iterations for computational efficiency and robustness, particularly beneficial for handling the high-dimensional state spaces in Nash games. These iterations aid in solving coupled algebraic Riccati equations, essential for identifying stabilizing solutions and achieving Nash equilibria, following the approach by Li and Gajic (1995)[1]. Our model is based on system dynamics and cost function parameters, using matrices to represent state transitions and control actions, with the aim of optimizing each player's cost function. The MATLAB implementation explores stabilizability and detectability conditions, guiding towards a feasible and stable Nash equilibrium. This research intersects game theory, control theory, and applied mathematics, offering insights into strategic decision-making and dynamic interactions across various fields.

The project has four stages: Problem Formulation, Implementation, Simulation Results and Conclusion.

II. Problem Formulation

A. System Dynamics

Consider a controlled linear dynamic system corresponding to the Nash differential game strategies is given by the following state equation:

$$\dot{x} = Ax + B_1u_1 + B_2u_2, \quad x(t_0) = x_0 \tag{1}$$

where $x \in R^n$ is the state vector, $u_1 \in R^{m_1}$ and $u_2 \in R^{m_2}$ are control inputs A, B_1, and B_2 are constant matrices of appropriate dimensions.

B. Performance Criterion

In this Nash differential game, the performance criterion for each player is defined by individual and coupled cost functions. With each control agent a quadratic type function is associated defined as:

$$J_1(u_1, u_2, x_0) = \frac{1}{2}\int_{t_0}^{\infty} (x^TQ_1x + u_1^TR_{11}u_1 + u_2^TR_{12}u_2)\, dt \tag{2}$$

$$J_2(u_1, u_2, x_0) = \frac{1}{2}\int_{t_0}^{\infty} (x^TQ_2x + u_1^TR_{21}u_1 + u_2^TR_{22}u_2)\, dt \tag{3}$$

The weighting matrices are symmetric and defined as follows:

$$Q_i \geq 0, \quad R_{ii} > 0, \quad R_{ij} \geq 0, \quad i = 1, 2; \quad j = 1, 2; \quad i \neq j \tag{4}$$

where Q_i are positive semidefinite, and R_{ii} are positive definite matrices. The optimal solution to the given problem leads to the so-called Nash optimal strategies u_1^* and u_2^* satisfying:

$$J_1(u_1^*, u_2^*) \leq J_1(u_1, u_2^*), \quad J_2(u_1^*, u_2^*) \leq J_2(u_1^*, u_2) \tag{5}$$

These equations describe the long-term costs incurred by each player, incorporating both the state variables and control actions. The matrices Q_1, Q_2 represent the cost of the states, while R_{11}, R_{22} are the control costs, and R_{12}, R_{21} quantify the interaction effects between the players' strategies.
The goal of each player in the Nash game is to minimize their respective cost function, reflecting a balance between individual objectives and the impact of the other playerâs decisions.

C. Hamiltonian

The Hamiltonians H_1 and H_2 in the Nash game optimization problem are defined as:

$$\begin{aligned} H_1(x(t), p_1(t), u_1(t), u_2(t)) &= \frac{1}{2}x(t)^T Q_1 x(t) \\ &+ u_1(t)^T R_{11} u_1(t) + u_2(t)^T R_{12} u_2(t) \\ &+ p_1(t)^T \left(Ax(t) + B_1 u_1(t) + B_2 u_2(t)\right) \end{aligned} \tag{6}$$

$$\begin{aligned} H_2(x(t), p_2(t), u_1(t), u_2(t)) &= \frac{1}{2}x(t)^T Q_2 x(t) \\ &+ u_1(t)^T R_{21} u_1(t) + u_2(t)^T R_{22} u_2(t) + \\ &p_2(t)^T \left(Ax(t) + B_1 u_1(t) + B_2 u_2(t)\right) \end{aligned} \tag{7}$$

These equations describe the Hamiltonian functions for each player in the Nash differential game. The Hamiltonian for each player is a composite function that includes the state $x(t)$, the co-state $p_i(t)$, and the control inputs $u_1(t)$ and $u_2(t)$. The state and control terms are weighted by the matrices $Q_1, Q_2, R_{11}, R_{12}, R_{21}$, and R_{22}, while the dynamics of the system are captured by the matrices A, B_1, and B_2. These Hamiltonians are central to determining the optimal control strategies for the players in the game.

D. Necessary Conditions

The necessary conditions for optimum in the Nash differential game are given by the following set of equations:

$$\frac{dx(t)}{dt} = Ax(t) + B_1 u_1(t) + B_2 u_2(t),\ x(t_0) = x_0,\ i = 1, 2 \tag{8}$$

$$\begin{aligned} &\frac{dp_1(t)}{dt} = -Q_1 x(t) - A^T p_1(t) + \left(\frac{\partial H_1}{\partial u_2}\right)^T, \\ &p_1(t_f) = \frac{\partial \gamma_1(x(t_f))}{\partial x}, \end{aligned} \tag{9}$$

$$\begin{aligned} &\frac{dp_2(t)}{dt} = -Q_2 x(t) - A^T p_2(t) + \left(\frac{\partial H_2}{\partial u_1}\right)^T, \\ &p_2(t_f) = \frac{\partial \gamma_2(x(t_f))}{\partial x} \end{aligned} \tag{10}$$

where the Hamiltonians H_1 and H_2 are given by:

$$\frac{\partial H_1}{\partial u_1} = R_{11} u_1 + B_1^T p_1 = 0,\ u_1^{opt}(t) = -R_{11}^{-1} B_1^T p_1(t), \tag{11}$$

$$\frac{\partial H_2}{\partial u_2} = R_{22} u_2 + B_2^T p_2 = 0,\ u_2^{opt}(t) = -R_{22}^{-1} B_2^T p_2(t), \tag{12}$$

These conditions outline that the optimal control strategies in feedback form are linear functions of the state variables, characterized by the gain matrices $K_1(t)$ and $K_2(t)$. Furthermore, the conditions specify the dynamics of the co-state variables $p_1(t)$ and $p_2(t)$. Additionally, the derivatives of the Hamiltonians with respect to the control inputs are related to the system matrices as follows:

$$\frac{\partial H_1}{\partial u_1} = R_{11} u_1 + B_1^T p_1, \quad \frac{\partial H_2}{\partial u_2} = R_{22} u_2 + B_2^T p_2 \tag{13}$$

The state and co-state equations when both players use feedback controls become:

$$\frac{dx(t)}{dt} = (A - S_1 K_1(t) - S_2 K_2(t))x(t), \tag{14}$$

$$\begin{aligned} \frac{dp_1(t)}{dt} &= -(Q_1 + K_1(t) Z_1 K_1(t))x(t) \\ &\quad - (A - S_1 K_1(t))^T p_1(t), \end{aligned} \tag{15}$$

$$\begin{aligned} \frac{dp_2(t)}{dt} &= -(Q_2 + K_2(t) Z_2 K_2(t))x(t) \\ &\quad - (A - S_2 K_2(t))^T p_2(t), \end{aligned} \tag{16}$$

where the matrices S_1, S_2, Z_1, and Z_2 are defined by:

$$\begin{aligned} S_1 &= B_1 R_{11}^{-1} B_1^T, & S_2 &= B_2 R_{22}^{-1} B_2^T, \\ Z_1 &= B_1 R_{11}^{-1} R_{12}^{-1} B_1^T, & Z_2 &= B_2 R_{22}^{-1} R_{21}^{-1} B_2^T \end{aligned} \tag{17}$$

These equations illustrate the structure of the Nash equilibrium in terms of feedback controls and demonstrate the interdependence of the agents' strategies through the system matrices.

E. Nash Strategies and Coupled Algebraic Riccati Equations

In game theory, Nash strategies represent a set of strategies where each player's strategy is optimal given the other players' strategies. In the context of linear-quadratic differential games, Nash strategies can be obtained through the synthesis of feedback control laws that rely on solving coupled algebraic Riccati equations (CAREs). The closed-loop Nash strategy for each player is defined by:

$$u_i^* = -R_{ii}^{-1} B_i^T K_i x, \quad i = 1, 2 \tag{18}$$

where K_i satisfies the coupled algebraic Riccati equations:

$$\begin{aligned} &K_1 A + A^T K_1 + Q_1 - K_1 S_1 K_1 - K_2 S_2 K_1 - K_1 S_2 K_2 \\ &+ K_2 Z_2 K_2 = N_1(K_1, K_2) = 0 \end{aligned} \tag{19}$$

$$\begin{aligned} &K_2 A + A^T K_2 + Q_2 - K_2 S_2 K_2 - K_2 S_1 K_1 - K_1 S_1 K_2 \\ &+ K_1 Z_1 K_1 = N_2(K_1, K_2) = 0 \end{aligned} \tag{20}$$

where, $S_i = B_i R_{ii}^{-1} B_i^T,\ Z_i = B_j R_{ij}^{-1} R_{ji} R_{jj}^{-1} B_j^T$ and $i, j = 1, 2,\ i \neq j$

The existence of solutions to these CAREs and the subsequent derivation of Nash strategies are contingent upon the system's parameters satisfying certain detectability and stabilizability conditions. These conditions ensure that the control laws derived are not only mathematically sound but also practically implementable.

F. Lyapunov Iterations

Lyapunov iterations are instrumental in computing the solutions to the coupled algebraic Riccati equations, which are pivotal for determining the closed-loop Nash strategies in linear-quadratic differential games.

1) Initialization: The initialization of the Lyapunov iterations is based on the assumption that the system under consideration satisfies certain stabilizability-detectability conditions. These conditions ensure that there exists a unique positive definite solution to an auxiliary algebraic Riccati equation, which provides the initial stabilizing solutions.

a) Hurwitz Condition (Stabilizable and Detectable): Either the triple $(A, B_1, \sqrt{Q_1})$ or $(A, B_2, \sqrt{Q_2})$ is stabilizable-detectable. To guarantee the existence of a stabilizing solution, the system described by the matrices A, B_i, and the weighting matrix Q_i must satisfy the Hurwitz condition, which necessitates that the system is both stabilizable and detectable. Mathematically, this is captured by the following set of auxiliary Riccati equations:

$$K_i^{(0)} A + A^T K_i^{(0)} + Q_i - K_i^{(0)} S_i K_i^{(0)} = 0, \quad i = 1, 2 \tag{21}$$

These equations serve as the initial conditions for the iterative process, where $K_i^{(0)}$ are the initial stabilizing solutions, and S_i are the corresponding solutions to the Lyapunov equations:

$$S_i = B_i R_{ii}^{-1} B_i^T, \quad i = 1, 2 \tag{22}$$

The matrices R_{ii} are derived from the cost functionals associated with each control agent in the Nash game, signifying the control effort penalties.

2) Iterative Procedure: The Lyapunov iterations for solving these equations are performed under the stabilizability-detectability assumption. The iterations are defined as:

$$\begin{aligned}&(A - S_1 K_1^{(i)} - S_2 K_2^{(i)})^T K_1^{(i+1)} \\ &+ K_1^{(i+1)} (A - S_1 K_1^{(i)} - S_2 K_2^{(i)}) = -Q_1^{(i)},\end{aligned} \tag{23}$$

$$\begin{aligned}&(A - S_1 K_1^{(i)} - S_2 K_2^{(i)})^T K_2^{(i+1)} \\ &+ K_2^{(i+1)} (A - S_1 K_1^{(i)} - S_2 K_2^{(i)}) = -Q_2^{(i)}\end{aligned} \tag{24}$$

for $i, j = 1, 2$, $i \neq j$, where $Q_i^{(n)}$ are updated iteratively. The algorithm starts with initial conditions obtained from auxiliary algebraic Riccati equations and iterates towards the stabilizing solution. The iterative process is designed to converge to the unique positive definite stabilizing solution of the coupled algebraic Riccati equations, ensuring the existence and uniqueness of the solution.

G. Optimal Performance Criterion

Lyapunov equations:

$$\begin{aligned}&(A - S_1 K_1 - S_2 K_2)^T V_1 + V_1 (A - S_1 K_1 - S_2 K_2) + Q_1 \\ &+ K_1^T S_1 K_1 + K_2^T Z_2 K_2 = 0\end{aligned} \tag{25}$$

$$\begin{aligned}&(A - S_1 K_1 - S_2 K_2)^T V_2 + V_2 (A - S_1 K_1 - S_2 K_2) + Q_2 \\ &+ K_2^T S_2 K_2 + K_1^T Z_1 K_1 = 0\end{aligned} \tag{26}$$

Moreover, by setting $V_1 = K_1$ and $V_2 = K_2$, we see that V_1 and V_2 satisfy the coupled algebraic Riccati equations (31), so that the optimal performance criteria are given by:

$$J_1^{opt}(x(t_0)) = \frac{1}{2} x_0^T V_1 x_0 = \frac{1}{2} x_0^T K_1 x_0 \tag{27}$$

$$J_2^{opt}(x(t_0)) = \frac{1}{2} x_0^T V_2 x_0 = \frac{1}{2} x_0^T K_2 x_0 \tag{28}$$

1) Convergence: The convergence of the Lyapunov iterations is evaluated by examining the stability of the updated solutions and the associated cost functionals. Convergence is indicated by the successive solutions $K_i^{(k)}$ approaching a limit that satisfies the coupled algebraic Riccati equations.

a) Condition for Convergence: The iterative solutions $K_i^{(k)}$ are said to converge if they satisfy the coupled algebraic Riccati equations as k approaches infinity:

$$\begin{aligned}&(A - S_1 K_1^{(\infty)} - S_2 K_2^{(\infty)})^T K_1^{(\infty)} \\ &+ K_1^{(\infty)} (A - S_1 K_1^{(\infty)} - S_2 K_2^{(\infty)}) \\ &+ (Q_1 + K_1^{(\infty)} S_1 K_1^{(\infty)} + K_2^{(\infty)} Z_2 K_1^{(\infty)}) = 0,\end{aligned} \tag{29}$$

$$\begin{aligned}&(A - S_1 K_1^{(\infty)} - S_2 K_2^{(\infty)})^T K_2^{(\infty)} \\ &+ K_2^{(\infty)} (A - S_1 K_1^{(\infty)} - S_2 K_2^{(\infty)}) \\ &+ (Q_2 + K_1^{(\infty)} Z_1 K_1^{(\infty)} + K_2^{(\infty)} S_2 K_2^{(\infty)}) = 0.\end{aligned} \tag{30}$$

This indicates that the limit points $K_i^{(\infty)}$ of the sequences $K_i^{(k)}$ represent the sought solutions of the Nash strategies.

III. Implementation

The implementation focuses on applying the policy iteration algorithm to a non-zero sum Nash differential game.

A. System Initialization

The MATLAB code initializes the system matrices and cost function parameters as follows:

```
A = [-0.0366, 0.0271, 0.0188, -0.4555;
     0.0482, -1.0100, 0.0024, -4.0208;
     0.1002, 0.2855, -0.7070, 1.3229;
     0, 0, 1.0000, 0];
B1 = [0.4422; 3.0447; -5.52; 0];
B2 = [0.1761; -7.5922; 4.99; 0];

Q1 = diag([3.5, 2, 4, 5]);
Q2 = diag([1.5, 6, 3, 1]);
R11 = 1;
R12 = 0.25;
R21 = 0.6;
R22 = 2;
The initial state vector x0 is set to [0;
    0; 0; 1].
```

B. Initialization and stabilizable-detectable check

The policy iteration begins with the computation of matrices $K1$ and $K2$ using the Algebraic Riccati Equation (ARE):

```
K1 = are(A, S1, Q1);
K2 = are(A - S1 * K1, S2, Q2 + K1 * S1 *
    K1);
```

The triple $(A, B_1, \sqrt{Q_1})$ or $(A, B_2, \sqrt{Q_2})$ is stabilizable-detectable. This implies weaker conditions than being controllable-observable.

At least one control agent must be capable of controlling and observing the unstable modes. Given that the context is a non-cooperative game, the joint effect of the agents is assumed to address the unstable modes, albeit this is considered somewhat idealistic.

To assert the first triple is stabilizable-detectable, the following conditions must be satisfied:

- The matrix $A - S_1 K_1^{(0)}$ must be stable, where S_1 and $K_1^{(0)}$ are matrices derived from the system parameters.
- The matrix $Q_2 + K_1^{(0)} Z_1 K_1^{(0)}$ should be positive semidefinite, indicating that the system does not respond negatively to state deviations.
- Finally, the matrix $A - S_1 K_1^{(0)} - S_2 K_2^{(0)}$ is also required to be stable, ensuring the overall system stability even after control adjustments.

These conditions are essential for the system to be not only stabilizable but also detectable, allowing for reliable control and observation of the system's dynamics.

C. Policy Iteration Loop

The main loop of the policy iteration algorithm involves updating the feedback matrices $K1$ and $K2$ and computing the cost functionals $J1$ and $J2$. The iteration continues until the convergence is reached:

```
for i = 2:num_iterations
    % Update K matrices
    % Store K matrices and compute cost
       functionals J
    % ...
end
```

IV. Simulation Results

A. State Trajectories

These trajectories illustrate the evolution of the system's state vector $x$as the policy iteration progresses, converging to the Nash equilibrium.

Each pair of figures shows the state trajectories corresponding to specific coordinates over the course of seven iterations. The convergence of these trajectories towards the same path indicates the algorithm's efficacy in stabilizing the system and finding an optimal solution for both players in the Nash differential game.

B. Control Trajectories

The control trajectories for each player in the Nash differential game are depicted in Figures 5 and 6. These plots display the control inputs u_1 and u_2 over time for seven iterations of the policy iteration algorithm.

In the case of Player 1, as depicted in Figure 5, the control input exhibits a rapid convergence towards the Nash equilibrium strategy after the initial iterations. The fluctuations in the control input diminish as the iterations progress, indicating a stabilization in the policy and approaching an optimal control strategy.

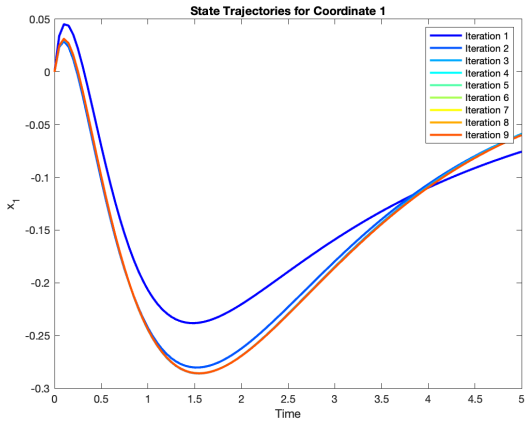

Fig. 1. State trajectories for Coordinate 1.

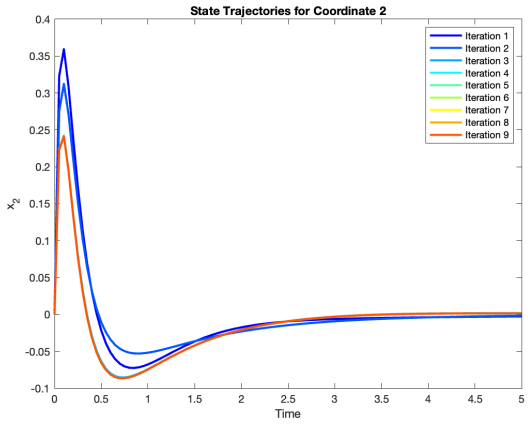

Fig. 2. State trajectories for Coordinate 2.

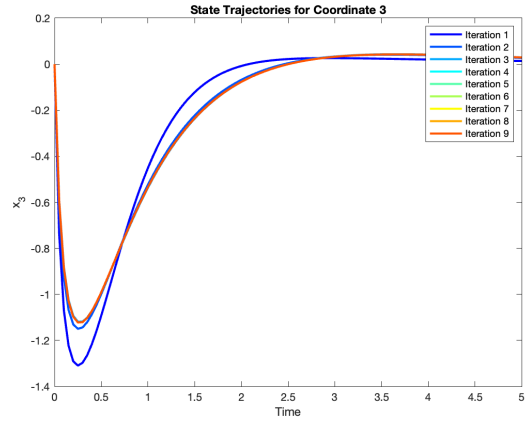

Fig. 3. State trajectories for Coordinate 3.

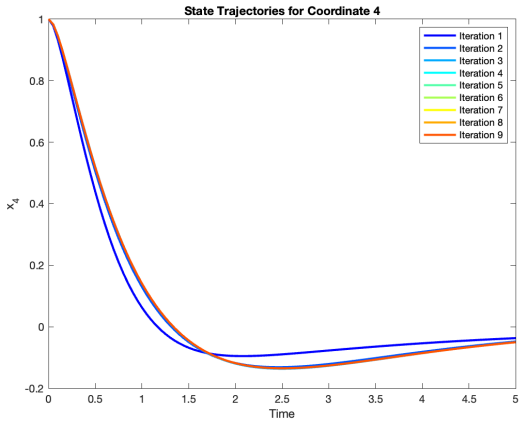

Fig. 4. State trajectories for Coordinate 4.

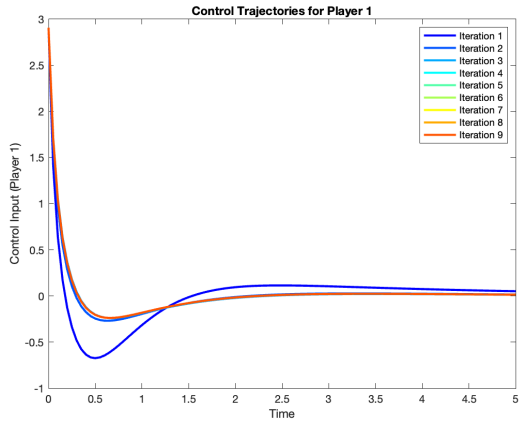

Fig. 5. Control trajectories for Player 1.

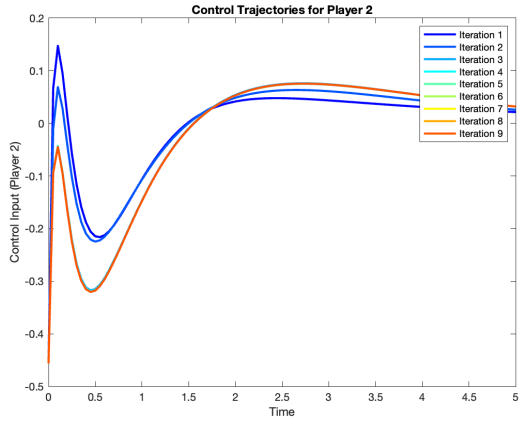

Fig. 6. Control trajectories for Player 2.

Similarly, the control input for Player 2, shown in Figure 6, also stabilizes over the iterations. The convergence is indicative of the players reaching a consensus on the best-response strategies, which is a characteristic of Nash equilibrium in a differential game setting. The diminishing variance in control inputs between successive iterations for both players signifies that the equilibrium point is not only reached but also maintained, validating the effectiveness of the policy iteration algorithm.

These trajectories corroborate the theoretical expectation that as the players adjust their strategies in response to each other's moves, they naturally gravitate towards a set of strategies that represent the Nash equilibrium, where no player has anything to gain by unilaterally changing their strategy.

C. Convergence and Performance Criterion

The convergence of the cost functionals J_1 and J_2 for both players is depicted in Figure 7. The graph tracks the cost associated with each player's strategy over successive iterations of the policy iteration algorithm.

The graph clearly shows that Player 1's cost functional J_1 experiences a significant drop between the first and the second iteration, followed by a plateau, which indicates a rapid convergence to a strategy that minimizes their cost. On the other hand, Player 2's cost functional J_2 starts at a lower value and remains relatively stable across the iterations, suggesting that Player 2's initial strategy was already near-optimal or that their optimal control strategy is less sensitive to the system dynamics and the actions of Player 1.

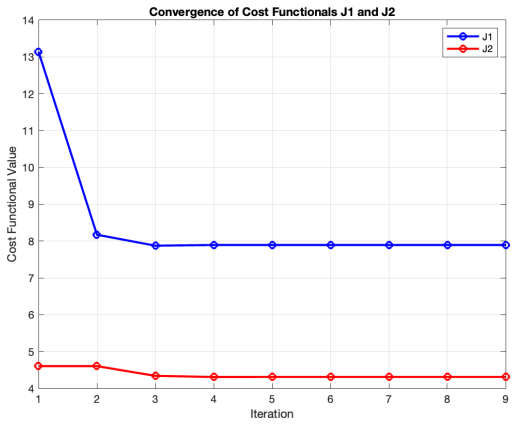

Fig. 7. Convergence of Cost Functionals J_1 and J_2.

The convergence of both cost functionals from the second iteration onwards implies that the Nash equilibrium was reached quickly and maintained throughout the remaining iterations and the control strategies derived from the Nash equilibrium are robust, leading to consistent performance even as the game progresses. This is indicative of the effectiveness of the policy iteration algorithm in finding stable strategies for both players where neither player can unilaterally improve their outcome, satisfying the Nash equilibrium condition in a non-zero sum game.

The stability of the cost functionals also suggests that the control strategies derived from the Nash equilibrium are robust, leading to consistent performance even as the game progresses.

This robustness is crucial in dynamic environments where the strategies must adapt to evolving conditions while maintaining optimality.

TABLE I
PERFORMANCE CRITERION VALUES AT EACH ITERATION

Iteration	Player 1 (J1)	Player 2 (J2)
1	13.142	4.6047
2	8.1727	4.6047
3	7.8751	4.3413
4	7.894	4.3095
5	7.8938	4.3123
6	7.8946	4.3123
7	7.8947	4.3124
8	7.8947	4.3124
9	7.8947	4.3124

The Table I illustrates the progression of cost functionals for both players as the policy iteration algorithm iteratively refines their strategies. The stability of these cost functionals from the second iteration onwards confirms the convergence to a Nash equilibrium. This equilibrium represents a situation where neither player can unilaterally improve their outcome, signifying the effectiveness of the policy iteration algorithm in finding stable and optimal strategies in the context of a non-zero sum differential game.

V. CONCLUSION

Drawing upon the meticulous exploration and implementation detailed in above sections, it becomes evident that the integration of reinforcement learning with Nash differential games through the application of policy iteration algorithms presents a significant advancement in the domain of strategic decision-making in complex dynamic environments. The research meticulously navigates the intricate landscape of coupled algebraic Riccati equations using Lyapunov iterations, thereby unveiling a novel computational methodology for deriving stabilizing solutions in multi-agent settings. This approach not only achieves convergence towards optimal Nash equilibria but also highlights the robustness and applicability of such strategies across various fields, ranging from economics to strategic planning.

The MATLAB simulations serve as a concrete testament to the theoretical framework's viability, showcasing the algorithm's effectiveness in stabilizing and optimizing control strategies within a non-zero sum differential game setting. By successfully addressing the challenges associated with the high-dimensional state spaces inherent in Nash games, this work demonstrates a significant leap forward in our ability to model and analyze strategic interactions in a mathematically rigorous and computationally efficient manner.

In conclusion, this project represents a confluence of game theory, control theory, and applied mathematics, providing a rich theoretical and practical foundation for future research in the strategic analysis of dynamic games. It propels the field forward by offering a sophisticated toolset for unraveling the complexities of multi-agent decision-making processes, thereby setting a new standard for excellence in research at the intersection of these disciplines.

REFERENCES

[1] Li, T. Y., & Gajic, Z. (1995). Lyapunov iterations for solving coupled algebraic Riccati equations of Nash differential games and the algebraic Riccati equation of zero-sum games. In *New Trends in Dynamic Games and Applications*, G. J. Olsder (Ed.), 333-351. Birkhauser.

[2] Vrabie, D., & Lewis, F. (2012). Integral Reinforcement Learning for Finding Online the Feedback Nash Equilibrium of Nonzero-Sum Differential Games. In *Advances in Reinforcement Learning*, A. Mellouk (Ed.), 313-330. IntechChina.

[3] Anderson, B. D. O., & Moore, J. B. (2007). *Optimal Control: Linear Quadratic Methods*. Dover Publications.

[4] Lewis, F. L., Vrabie, D., & Syrmos, V. L. (2012). *Optimal Control*. John Wiley & Sons.

[5] Sutton, R. S., & Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. MIT Press.

[6] Kleinman, D. L. (1968). On an iterative technique for Riccati equation computations. *IEEE Transactions on Automatic Control*, 13(1), 114-115.