Project Report

Cooperative Q-Learning for Rejection of Persistent Adversarial Inputs in Networked Linear Quadratic Systems

By Murtaza Rangwala

ECE6504: Autonomous Coordination

Overview

This paper addresses the optimization of the distributed performance criteria of agents in a network who are synchronizing to a leader noted using cooperative Q-learning. Graphical games based on Game-Theory are generally used to model these multi-agent learning policies. However, most of these game based techniques, require complete modelling of the system, which is not feasible for very large networks. Moreover, solving ODEs in conventional optimal control policies for large networks of multi-agent systems, also suffers from a high computation cost. Q-learning, a popular reinforcement learning algorithm, is applied to this problem, in this paper. Q-learning, is a model-free learning algorithm, which has been applied successfully in this area. However, this learning policy, is well defined for discrete-time based systems, and has not been well established for continuous-time based systems.

The paper proposes a novel way of modelling a continuous time system, and moreover, guarantees convergence to a Nash Equilibrium for a network of agents, synchronizing to one leader. The paper also provides guarantees for an optimal solution, and closed loop stability of the system, while assuming constraints that there is a unique non-negative solution for the saddle point of the Nash Equilibrium.

Problem Formulation

A strongly connected networked-system, G, of N agents is considered, and modelled as follows:

$$\dot{x_i}(t) = A_i x_i(t) + B_i u_i(t) + D_i v_i(t),\ and\ x_i(0) = x_{i0},\ t \geq 0, \forall i \epsilon N$$

Each state, $x_i(t)$, is a measurable state vector available for feedback by each agent and the initial conditions are known for all agents. $u_i(t)$ is the control input, or the minimizing player, and $v_i(t)$ is adversarial input, or the maximizing player. The assumption of (A_i, B_i) is stabilizable is made throughout the paper. This is required for the existence of a linear feedback control, Theorem 9.7, Ref [1]. The dynamics of the leader is defined as, $\dot{x_L}(t) = A_L x_L(t)$. The agents should seek to cooperatively track the state of the leader node, with the above defined dynamics of the leader. The distributed performance of the system, is then defined as the neighbourhood tracking error of each agent, considering the error of the agents neighbour and its own with the leader.

$$e_i = x_i(t) - x_L(t)$$

$$e_i = \sum_{j \epsilon N_i} a_{ij}(x_i - x_j) + g_i(x_i - x_L),\ \forall i \epsilon N$$

For the system to be asymptotically tracking the leader, the neighbourhood error should be, $e_i \to 0$.

The dynamics of the error is then given by,

$$\dot{e_i} = Ae_i + (d_i + g_i)(B_i u_i + D_i v_i) - \sum_{j \epsilon N_i} a_{ij}(B_j u_j - D_j v_j),\ \forall i \epsilon N$$

$$d_i = \sum_{j \epsilon N_i} a_{ij}, defined\ as\ the\ weighted\ degree\ of\ the\ node\ i$$

The cost function for neighbourhood tracking error, is then defined for each agent. Moreover, the factor of 0.5, is chosen arbitrarily, for a cleaner solution for the following theorems. The weights can be user-defined for each system that you have, and do not affect the optimality and convergence guarantees.

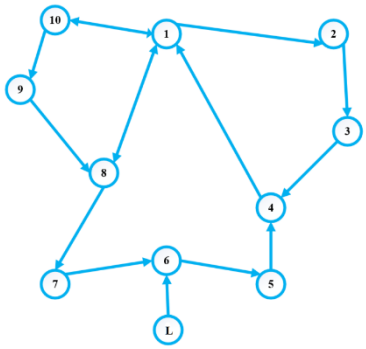

Fig. 1. A networked system $\mathcal{G}$ with 10 agents and a leader node L pinned to agent 6.

The cost function is then defined as,

$$J_i\left(e_i(0), \mathrm{u}_i, \mathrm{u}_{N_i}, \mathrm{v}_i, \mathrm{v}_{N_i}\right) = \frac{1}{2}\int_0^\infty \left({e_i}^T \mathrm{H}_i e_i + ({\mathrm{u}_i}^T R_{ii} \mathrm{u}_i - {\gamma_{ii}}^2 {v_i}^T v_i) + \sum_{j \epsilon N_i} \left({\mathrm{u}_j}^T R_{ij} \mathrm{u}_j - {\gamma_{ij}}^2 {v_j}^T v_j \right) \right) dt$$

The problem is generally formulated as an infinite-time horizon linear quadratic soft-constrained differential game. The assumption that, (A_i, B_i) *are stabilizable*, and $\left(\sqrt{H_i}, A\right)$ *are detectable* also guarantees, to ensure a unique non-negative solution, Theorem 4.8 and Theorem 9.7, Ref 1. The detectability of $\left(\sqrt{H_i}, A\right)$, guarantees that stabilizing control is generated naturally through the solution process, Theorem 9.7.

For a strongly connected graph, with our initial choice of the control input, and adversarial input, we have the graphical Nash equilibrium,

$$J_i\left(e_i(0), {\mathrm{u}_i}^*, {\mathrm{u}_{N_i}}^*, {\mathrm{v}_i}^*, {\mathrm{v}_{N_i}}^*\right) = \min_{\mathrm{u}_i} J_i\left(e_i(0), \mathrm{u}_i, {\mathrm{u}_{N_i}}^*, {\mathrm{v}_i}^*, {\mathrm{v}_{N_i}}^*\right) = \max_{\mathrm{v}_i} J_i\left(e_i(0), {\mathrm{u}_i}^*, {\mathrm{u}_{N_i}}^*, \mathrm{v}_i, {\mathrm{v}_{N_i}}^*\right),$$

from equation 2.11a.

Using the definition of the value function, from Chapter 4, Ref 1, we get,

$$V_i^*\left(e_i(t)\right) = \min_{\mathrm{u}_i} \max_{\mathrm{v}_i} \frac{1}{2}\int_t^\infty \left({e_i}^T \mathrm{H}_i e_i + ({\mathrm{u}_i}^T R_{ii} \mathrm{u}_i - {\gamma_{ii}}^2 {v_i}^T v_i) + \sum_{j \epsilon N_i} \left({\mathrm{u}_j}^T R_{ij} \mathrm{u}_j - {\gamma_{ij}}^2 {v_j}^T v_j \right) \right) dt$$

where the weighted norms are considered for the feedback, from time t.

Note, that the system matrices A_i, B_i, D_i are not considered for the value function.

Existence of Nash Equilibrium

Through Theorem 1, proposed in the paper, the authors prove the existence of a Nash Equilibrium given a set of constraints that are satisfied.

The Hamilton-Jacobi-Bellman equation, for the value function that is defined above, using equation 4.60, and Theorems 4.28 and 4.29, is given as:

$$H_i\left(e_i, \mathrm{u}_i, \mathrm{u}_{N_i}, \mathrm{v}_i, \mathrm{v}_{N_i}, \frac{\partial V_i^*}{\partial e_i}\right) = \frac{\partial {V_i^*}^T}{\partial e_i}\left(A e_i + (d_i + g_i)(B_i u_i + D_i v_i) - \sum_{j \epsilon N_i} a_{ij}\left(B_j u_j - D_j v_j\right) \right) + \frac{1}{2}\left({e_i}^T \mathrm{H}_i e_i + ({\mathrm{u}_i}^T R_{ii} \mathrm{u}_i - {\gamma_{ii}}^2 {v_i}^T v_i) + \sum_{j \epsilon N_i} \left({\mathrm{u}_j}^T R_{ij} \mathrm{u}_j - {\gamma_{ij}}^2 {v_j}^T v_j \right) \right)$$

Using Theorems 3.2, 3.12, and 3.13, we get the optimal control and worst case adversarial input as

$$u_i^*(e_i) = \arg\min_{u_i} \mathcal{H}_i(e_i, u_i, u_{\mathcal{N}_i}, v_i, v_{\mathcal{N}_i}, \frac{\partial V_i^*}{\partial e_i}$$
$$= -(d_i + g_i)R_{ii}^{-1}B_i^T \frac{\partial V_i^*}{\partial e_i}, \ \forall e_i,$$

$$v_i^*(e_i) = \arg\max_{v_i} \mathcal{H}_i(e_i, u_i, u_{\mathcal{N}_i}, v_i, v_{\mathcal{N}_i}, \frac{\partial V_i^*}{\partial e_i}$$
$$= \frac{(d_i + g_i)}{\gamma_{ii}^2} D_i^T \frac{\partial V_i^*}{\partial e_i}, \ \forall e_i.$$

The saddle point solution, will satisfy the Hamilton-Jacobi-Bellman equation –

$$H_i\left(e_i, \mathrm{u_i}^*, \mathrm{u_{N_i}}^*, \mathrm{v_i}^*, \mathrm{v_{N_i}}^*, \frac{\partial V_i^*}{\partial e_i}\right) = 0$$

Then we have a unique non-negative solution, P_i, that solves the above equation, where the value function can be represented as,

$$V_i^*(e_i) = \frac{1}{2} {e_i}^T \mathrm{P}_i e^i$$

The above Hamiltonian can then be rewritten as,

$$H_i\left(e_i, \mathrm{u_i}^*, \mathrm{u_{N_i}}^*, \mathrm{v_i}^*, \mathrm{v_{N_i}}^*, {e_i}^T \mathrm{P_i}\right) = 0$$

$$u_i^*(e_i) = -(d_i + g_i)R_{ii}^{-1}B_i^T P_i e_i, \ \forall e_i,$$

$$v_i^*(e_i) = \frac{(d_i + g_i)}{\gamma_{ii}^2} D_i^T P_i e_i, \ \forall e_i.$$

Solving for the above equation, requires a complete knowledge of the system and is also computationally expensive.

Theorem1: Let the pairs(A, B_i) are stabilizable and $\left(\sqrt{H_i}, A\right)$ detectable. Suppose that the graph G is strongly connected and that there exists unique symmetric positive definite matrices $\mathrm{P_i}$, that satisfy the coupled Riccati inequalities,

$$H_i\left(e_i, \mathrm{u_i}^*, \mathrm{u_{N_i}}^*, \mathrm{v_i}^*, \mathrm{v_{N_i}}^*, {e_i}^T \mathrm{P_i}\right) = 0$$

with $V_i^*(0) = 0, \ \forall i \in \mathcal{N}$,

$$B_i R_{ii}^{-1} B_i^T \succ \frac{1}{\gamma_{ii}^2} D_i D_i^T, \ \forall i \in \mathcal{N}$$
$$B_j R_{jj}^{-T} R_{ij} R_{jj}^{-1} B_j^T \succ \frac{\gamma_{ij}^2}{\gamma_{jj}^4} D_j D_j^T, \ \forall i, j \in \mathcal{N},$$

with optimal control input,

$$\bar{u}_i(e_i) = -(d_i + g_i)R_{ii}^{-1}B_i^T P_i e_i, \forall e_i,$$

and worst case adversarial input,

$$\bar{v}_i(e_i) = \frac{(d_i + g_i)}{\gamma_{ii}^2} D_i^T P_i e_i, \forall e_i.$$

Then all the agents synchronize asymptotically to the state of the leader, i.e. the equilibrium point of the closed loop system.

$$\dot{V}_i = \frac{\partial V_i}{\partial e_i}\left(Ae_i + (d_i + g_i)(B_i\bar{u}_i + D_i\bar{v}_i) - \sum_{j\in\mathcal{N}_i} a_{ij}(B_j\bar{u}_j + D_j\bar{v}_j)\right)$$
$$= -\frac{1}{2}\left(e_i^T H_i e_i + (\bar{u}_i^T R_{ii}\bar{u}_i - \gamma_{ii}^2 \bar{v}_i^T \bar{v}_i)\right.$$
$$\left. + \sum_{j\in\mathcal{N}_i} (\bar{u}_j^T R_{ij}\bar{u}_j - \gamma_{ij}^2 \bar{v}_j^T \bar{v}_j)\right), \ \forall i \in \mathcal{N},$$

Using eqn. 2.17a as a reference from the textbook, we get an upper bound on the dynamics of the value function,

$$\dot{V}_i \leq -\frac{1}{2}\underline{\lambda}(\mathrm{H}_i)\|e_i\|^2$$

Using V_i as a Lyapunov function for every agent, the authors prove that,

$V_i^*(e_i) = 0$ and $\lim_{t\to\infty} V_i(e_i(t)) = 0$

With the above conditions, we can arbitrarily define control inputs and adversarial inputs, that satisfy,

$$\mathcal{J}_i(e_i(0); \bar{u}_i, \bar{u}_{\mathcal{N}_i}, \bar{v}_i, \bar{v}_{\mathcal{N}_i}) = \int_0^\infty \frac{1}{2}\left(e_i^T H_i e_i + (\bar{u}_i^T R_{ii}\bar{u}_i - \gamma_{ii}^2 \bar{v}_i^T \bar{v}_i)\right.$$
$$\left. + \sum_{j\in\mathcal{N}_i} (\bar{u}_j^T R_{ij}\bar{u}_j - \gamma_{ij}^2 \bar{v}_j^T \bar{v}_j)\right) dt$$
$$+ \frac{1}{2} e_i(0)^T P_i e_i(0) + \frac{1}{2}\int_0^\infty \dot{V}_i^* dt$$
$$= \int_0^\infty \frac{1}{2}\left(e_i^T H_i e_i + (\bar{u}_i^T R_{ii}\bar{u}_i - \gamma_{ii}^2 \bar{v}_i^T \bar{v}_i)\right.$$
$$\left. + \sum_{j\in\mathcal{N}_i} (\bar{u}_j^T R_{ij}\bar{u}_j - \gamma_{ij}^2 \bar{v}_j^T \bar{v}_j)\right) dt$$
$$+ \frac{1}{2} e_i(0)^T P_i e_i(0) + \frac{1}{2}\int_0^\infty \frac{\partial V_i^*}{\partial e_i}\left(Ae_i + (d_i + g_i)(B_i\bar{u}_i + D_i\bar{v}_i)\right.$$
$$\left. - \sum_{j\in\mathcal{N}_i} a_{ij}(B_j\bar{u}_j + D_j\bar{v}_j)\right) dt, \ \forall i \in \mathcal{N},$$

$$\mathcal{J}_i(e_i(0);\bar{u}_i,u^*_{\mathcal{N}_i},\bar{v}_i,v^*_{\mathcal{N}_i}) = \int_0^\infty \frac{1}{2}\Big((\bar{u}_i-u_i^*)^T R_{ii}(\bar{u}_i-u_i^*) - \gamma_{ii}^2(\bar{v}_i-v_i^*)^T(\bar{v}_i-v_i^*)\Big)dt + \frac{1}{2}e_i(0)^T P_i e_i(0),\ \forall i\in\mathcal{N}.$$

$$\mathcal{J}_i(e_i(0);u_i^*,u^*_{\mathcal{N}_i},\bar{v}_i,v^*_{\mathcal{N}_i}) = -\frac{1}{2}\gamma_{ii}^2\int_0^\infty(\bar{v}_i-v_i^*)^T(\bar{v}_i-v_i^*)dt + \frac{1}{2}e_i(0)^T P_i e_i(0),\forall i\in\mathcal{N}.$$

$$\mathcal{J}_i(e_i(0);\bar{u}_i,u^*_{\mathcal{N}_i},v_i^*,v^*_{\mathcal{N}_i}) = \frac{1}{2}\int_0^\infty(\bar{u}_i-u_i^*)^T R_{ii}(\bar{u}_i-u_i^*)dt + \frac{1}{2}e_i(0)^T P_i e_i(0),\ \forall i\in\mathcal{N}.$$

Setting all your inputs as your optimal values, gives

$$\mathcal{J}_i(e_i(0);u_i^*,u^*_{\mathcal{N}_i},v_i^*,v^*_{\mathcal{N}_i}) = \frac{1}{2}e_i(0)^T P_i e_i(0),\ \forall i\in\mathcal{N}.$$

This gives us the inequality,

$$J_i(e_i(0),\mathrm{u}_i{}^*,\mathrm{u}_{N_i}{}^*,\mathrm{v}_i{}^*,\mathrm{v}_{N_i}{}^*) = \min_{\mathrm{u}_i} J_i(e_i(0),\mathrm{u}_i,\mathrm{u}_{N_i}{}^*,\mathrm{v}_i{}^*,\mathrm{v}_{N_i}{}^*) = \max_{\mathrm{v}_i} J_i(e_i(0),\mathrm{u}_i{}^*,\mathrm{u}_{N_i}{}^*,\mathrm{v}_i,\mathrm{v}_{N_i}{}^*)$$

Q-Function

In order to formulate a Q-Function, the authors use the method of defining an Advantage function. This allow the matrix Q, to be state independent, easing computation.

$$Q_i(e_i,\mathrm{u}_i,\mathrm{u}_{N_i},\mathrm{v}_i,\mathrm{v}_{N_i}) = V_i^*(e_i) + H_i\left(e_i,\mathrm{u}_i,\mathrm{u}_{N_i},\mathrm{v}_i,\mathrm{v}_{N_i},\frac{\partial V_i^*}{\partial e_i}\right)$$

$$= V_i^*(e_i) + \frac{\partial V_i^*}{\partial e_i}^T\Big(Ae_i + (d_i+g_i)(B_iu_i+D_iv_i) - \sum_{j\in\mathcal{N}_i}a_{ij}(B_ju_j+D_jv_j)\Big) + \frac{1}{2}\Big(e_i^T H_i e_i + (u_i^T R_{ii}u_i - \gamma_{ii}^2 v_i^T v_i) + \sum_{j\in\mathcal{N}_i}(u_j^T R_{ij}u_j - \gamma_{ij}^2 v_j^T v_j)\Big),\ \forall e_i,u_i,u_{\mathcal{N}_i},v_i,v_{\mathcal{N}_i},\ \forall i\in\mathcal{N},$$

$$Q_i(e_i,\mathrm{u}_i,\mathrm{u}_{N_i},\mathrm{v}_i,\mathrm{v}_{N_i}) = \frac{1}{2}U_i{}^T\begin{bmatrix} P_i+H_i+P_iA+A^TP_i & (d_i+g_i)P_iB_i & -row(a_{ij}P_iB_j) & (d_i+g_i)P_iD_i & -row(a_{ij}P_iD_j)_{j\epsilon N_i} \\ (d_i+g_i)B_i{}^TP_i & R_{ii} & 0 & 0 & 0 \\ -col(a_{ij}B_j{}^TP_i) & 0 & diag(R_{ij})_{j\epsilon N_i} & 0 & 0 \\ (d_i+g_i)D_i{}^TP_i & 0 & 0 & -\gamma_{ii}{}^2 & 0 \\ -col(a_{ij}D_j{}^TP_i) & 0 & 0 & 0 & -diag(\gamma_{ij}{}^2)_{j\epsilon N_i} \end{bmatrix}U_i$$

Where, $U_i = \begin{bmatrix} e_i^T & u_i^T & u_{\mathcal{N}_i}^T & v_i^T & v_{\mathcal{N}_i}^T \end{bmatrix}^T$

For the learning rate using the cooperative Q-Learning, we define,

Critic Approximator Weights

$$Q_i(e_i,\mathrm{u}_i{}^*,\mathrm{u}_{N_i}{}^*,\mathrm{v}_i{}^*,\mathrm{v}_{N_i}{}^*) = W_{ic}{}^T(U_i\otimes U_i)$$

$$W_{ic} := \tfrac{1}{2}\mathrm{vech}(\bar{Q}^i)$$

Control Actor Weights

$$\mathrm{u}_i{}^*(e_i) = \min_{\mathrm{u}_i} Q_i(e_i,\mathrm{u}_i,\mathrm{u}_{N_i},\mathrm{v}_i,\mathrm{v}_{N_i}) = -\left(Q_{\mathrm{u}_i\mathrm{u}_i}{}^i\right)^{-1}\left(Q_{\mathrm{u}_i\mathrm{e}_i}{}^i\right)e_i = W_{ia}{}^Te_i$$

Adversarial Actor Weights

$$\mathrm{v}_i{}^*(e_i) = \max_{v_i} Q_i(e_i,\mathrm{u}_i,\mathrm{u}_{N_i},\mathrm{v}_i,\mathrm{v}_{N_i}) = -\left(Q_{\mathrm{v}_i v_i}{}^i\right)^{-1}\left(Q_{v_i\mathrm{e}_i}{}^i\right)e_i = W_{id}{}^Te_i$$

Consider a system, which requires time 'T', to get excited to read the initial measurements. This system, has its value function defined though a Bellman equation, which writes the value of a decision problem at a certain point in time in terms of the payoff from some initial

choices and the value of the remaining decision problem that results from those initial choices.

$$V_i^*(e_i(t)) = V_i^*(e_i(t-T))$$

$$-\frac{1}{2}\int_{t-T}^{t}\left(e_i^T H_i e_i + (u_i^{*T} R_{ii} u_i^* - \gamma_{ii}^2 v_i^{*T} v_i^*)\right.$$
$$\left.+\sum_{j\in\mathcal{N}_i}(u_j^{*T} R_{ij} u_j^* - \gamma_{ij}^2 v_j^{*T} v_j^*)\right)d\tau,\ \forall i \in \mathcal{N},$$

Notice that we can rewrite, the above function as an optimal Q function,

$$Q_i^*(e_i(t), u_i^*(t), u_{\mathcal{N}_i}^*(t), v_i^*(t), v_{\mathcal{N}_i}^*(t)) =$$
$$Q^*(e_i(t-T), u_i^*(t-T), u_{\mathcal{N}_i}^*(t-T), v_i^*(t-T), v_{\mathcal{N}_i}^*(t-T))$$
$$-\frac{1}{2}\int_{t-T}^{t}\left(e_i^T H_i e_i + (u_i^{*T} R_{ii} u_i^* - \gamma_{ii}^2 v_i^{*T} v_i^*)\right.$$
$$\left.+\sum_{j\in\mathcal{N}_i}(u_j^{*T} R_{ij} u_j^* - \gamma_{ij}^2 v_j^{*T} v_j^*)\right)d\tau,\ \forall i \in \mathcal{N}.$$

Now the estimates of the weights for the critic, and the two actors, need to converge to an optimal value. In order to do that, the weights are propagated using gradient descent. An error function of the Q function, and the two actors are calculated for defining the gradient descent. This is nothing but the difference between the LHS and RHS of your above equation.

$$E_i = \widehat{W_{ic}}(U_i \otimes U_i) + \frac{1}{2}\int_{t-T}^{t}\left(e_i{}^T \mathrm{H}_i e_i + \left(\hat{u}_i^T R_{ii}\hat{u}_i - \gamma_{ii}{}^2\hat{v}_i^T\hat{v}_i\right) + \sum_{j\in N_i}\left(\hat{u}_j^T R_{ij}\hat{u}_j - \gamma_{ij}{}^2\hat{v}_j^T\hat{v}_j\right)\right)dt - \widehat{W_{ic}}(U_i(t-T)\otimes U_i(t-T))$$

$$E_{ia} = \widehat{W}_{ia}^T e_i + \left(\hat{Q}^i_{u_i u_i}\right)^{-1}\left(\hat{Q}^i_{u_i e_i}\right)e_i$$

$$K_{ia} = \frac{1}{2}\|E_{ia}\|^2$$

$$\dot{\widehat{W}}_{ia} = -\alpha_{ia}\frac{dK_{ia}}{d\widehat{W_{ia}}}$$

$$E_{id} = \widehat{W}_{id}^T e_i + \left(\hat{Q}^i_{v_i v_i}\right)^{-1}\left(\hat{Q}^i_{v_i e_i}\right)e_i$$

$$K_{id} = \frac{1}{2}\|E_{id}\|^2$$

$$\dot{\widehat{W}}_{id} = -\alpha_{id}\frac{dK_{id}}{d\widehat{W_{id}}}$$

We defined the squared-norm errors for each error function, and the gradients are calculated using the least squared method.

$$K_{ic} = \frac{1}{2}\|E_{ic}\|^2$$

$$\dot{\widehat{W}}_{ic} = -\alpha_{ic}\frac{dK_{ic}}{d\widehat{W_{ic}}}$$

Simplifying for the critic estimate gradient, we get,

Algorithm 1: Proposed Q-learning

1: **procedure**
2: Start with initial conditions for every agent $x_i(0)$, and random initial weights $\hat{W}_{ic}(0), \hat{W}_{ia}(0), \hat{W}_{id}(0)$ for the critic, actor and worst case adversarial approximator for each agent.
3: Propagate $t, x_i(t)$ and $x_j(t)$ in the neighborhood to compute e_i from (2).
4: Propagate $\hat{W}_{ic}(t), \hat{W}_{ia}(t), \hat{W}_{id}(t)$ $\triangleright$ $\dot{\hat{W}}_{ic}$ as in (39), $\dot{\hat{W}}_{ia}$ as in (40) and $\dot{\hat{W}}_{id}$ as in (41).
5: Compute the Q function $\hat{Q}_i$ from (31), the optimal control $\hat{u}_i$ from (32) and the worst case adversarial input $\hat{v}_i$ from (33) for each agent.
6: **end procedure**

$$\dot{\hat{W}}_{ic} = -\alpha_{ic}\frac{\partial K_{i1}}{\partial \hat{W}_{ic}}$$

$$= -\alpha_{ic}\frac{\left(U_i(t)\otimes U_i(t) - U_i(t-T)\otimes U_i(t-T)\right)}{\left(1+\left(U_i(t)\otimes U_i(t) - U_i(t-T)\otimes U_i(t-T)\right)^T\left(U_i(t)\otimes U_i(t) - U_i(t-T)\otimes U_i(t-T)\right)\right)^2}E_i,\ \forall i\in\mathcal{N},$$

Note, the gradient is normalized here. The derivation of the gradient, only gives you the numerator. A normalization is added to the gradient.

$$\dot{\tilde{W}}_{ia} = -\alpha_{ia}e_i e_i^T\tilde{W}_{ia} - \alpha_{ia}e_i e_i^T\tilde{Q}^i_{\mathrm{e_i u_i}}(\hat{Q}^i_{\mathrm{u_i u_i}})^{-1},\ \forall i\in\mathcal{N},$$

$$\dot{\tilde{W}}_{id} = -\alpha_{id}e_i e_i^T\tilde{W}_{id} - \alpha_{id}e_i e_i^T\tilde{Q}^i_{\mathrm{e_i v_i}}(\hat{Q}^i_{\mathrm{v_i v_i}})^{-1},\ \forall i\in\mathcal{N},$$

Since we know the values for $\hat{Q}^i_{\mathrm{v}_i v_i}$ and $\hat{Q}^i_{u_i u_i}$ from the look up table, we can simplify the above equation as:

$$\dot{\tilde{W}}_{ia} = -\alpha_{ia}e_i e_i^T\tilde{W}_{ia} - \alpha_{ia}e_i e_i^T\tilde{Q}^i_{\mathrm{e_i u_i}}R_{ii}^{-1},\ \forall i\in\mathcal{N},$$

$$\dot{\tilde{W}}_{id} = -\alpha_{d}e_i e_i^T\tilde{W}_{id} - \frac{\alpha_{id}}{\gamma_{ii}^2}e_i e_i^T\tilde{Q}^i_{\mathrm{e_i v_i}},\ \forall i\in\mathcal{N}.$$

Lemma 2 of the paper proves that the error dynamics of the critic, with fixed control inputs, then the critic weights will have an exponentially stable equilibrium point that satisfies,

$$\left\|\widehat{W}_{ic}(t)\right\| \le \left\|\widehat{W}_{ic}(t_0)\right\|\rho_{i1}e^{-\rho_{i2}(t-t_0)},\ \forall t > t_0 \geq 0, \rho_{i1}, \rho_{i2}\epsilon\mathbb{R}^+$$

This is proved by defining a Lyapunov function

$$\mathcal{L}_i = \frac{1}{2\alpha_{ic}}\left\|\tilde{W}_{ic}\right\|^2,\ t\geqslant 0,\ \forall i\in\mathcal{N}.$$

$$\dot{\mathcal{L}}_i = \frac{1}{\alpha_{ic}}\tilde{W}_{ic}^T\dot{\tilde{W}}_{ic},\ \forall i\in\mathcal{N}$$

After substituting the value of your dynamics of the estimate function of critic, defined at the top of this page, we can rewrite the equation to,

$$\dot{\mathcal{L}}_i = -\tilde{W}_{ic}^T\Delta_i\Delta_i^T\tilde{W}_{ic}, \forall i\in\mathcal{N}.$$

Where,

$$\Delta_i(t) = \frac{\left(U_i(t)\otimes U_i(t) - U_i(t-T)\otimes U_i(t-T)\right)}{1+\left(U_i(t)\otimes U_i(t) - U_i(t-T)\otimes U_i(t-T)\right)^T\left(U_i(t)\otimes U_i(t) - U_i(t-T)\otimes U_i(t-T)\right)}$$

which is persistently exciting over the complete time interval.

The solution to linear time-varying system, with a state transition matrix defined as $\frac{\partial\Phi_i(t,t_0)}{\partial t} := -\alpha_{ic}\Delta_i(t)\Delta_i^T(t)\Phi_i(t,t_0)$

$$\tilde{W}_{ic}(t) = \Phi_i(t,t_0)\tilde{W}_{ic}(t_0),\ \forall t, t_0 > 0,\ \forall i\in\mathcal{N}$$

The state-transition matrix is exponentially stable, if we have a persistently exciting $\Delta_i(t)$.

And for some arbitrary coefficients, we would satisfy,

$$\|\Phi_i(t,t_0)\| = \rho_{i1}e^{-\rho_{i2}(t-t_0)},\ \forall t, t_0 > 0, \forall i\in\mathcal{N}.$$

For which we fulfil the criteria,

$$\left\|\tilde{W}_{ic}(t)\right\| \leqslant \left\|\tilde{W}_{ic}(t_0)\right\|\rho_{i1}e^{-\rho_{i2}(t-t_0)}, \forall t, t_0 > 0, \forall i\in\mathcal{N}$$

The next step of the paper provides guarantee on the closed loop stability of the origin of $\left[e_i{}^T\quad \widehat{W}_{ic}^T\quad \widehat{W}_{N_i a}^T\quad \widehat{W}_{ia}^T\quad \widehat{W}_{N_i d}^T\quad \widehat{W}_{ia}^T\right]$.

Theorem 2

The theorem states that considering the neighbourhood tracking error defined as part of this paper, and optimal control and worst case adversarial inputs, the tuning laws generated before and assuming that the tuning law for the critic is faster than the control and adversarial actor, and given the following conditions,

$$\alpha_{ia} > 1;\ \alpha_{id} > 1,$$

$$\alpha_{ia}\bar{\lambda}(R_{ii}^{-1}) + \frac{\alpha_{id}}{\gamma_{ii}^2} < \frac{2}{\delta_i}\Big(\underline{\lambda}\big(H_i + Q^i_{\mathrm{e_i u_i}} R_{ii}^{-1}(Q^i_{\mathrm{e_i u_i}})^T\big) - \frac{1}{2}\bar{\lambda}\big(Q^i_{\mathrm{e_i u_i}}(Q^i_{\mathrm{e_i u_i}})^T + \frac{2}{\gamma_{ii}^2} Q^i_{\mathrm{e_i v_i}}(Q^i_{\mathrm{e_i v_i}})^T\big)\Big),$$

$$\alpha_{ja}\bar{\lambda}(R_{jj}^{-1}) + \frac{\alpha_{jd}}{\gamma_{jj}^2} < \frac{2}{\delta_j}\Big(\underline{\lambda}\big(Q^j_{\mathrm{e_j u_j}} R_{jj}^{-1} R_{ij} R_{jj}^{-1}(Q^j_{\mathrm{e_j u_j}})^T - \frac{1}{2}\bar{\lambda}\big(Q^i_{\mathrm{e_i u_j}}(Q^i_{\mathrm{e_i u_j}})^T + \frac{\gamma_{ij}^2}{\gamma_{jj}^4} Q^i_{\mathrm{e_i v_j}}(Q^i_{\mathrm{e_i v_j}})^T\big)\Big),\ \forall j \in \mathcal{N}_i,$$

A Lyapunov function, is chosen such that,

$$\mathcal{V} = \sum_{i=1}^{N}\Big(V_i^*(e_i) + \frac{1}{2}\left\|\tilde{W}_{ic}\right\|^2 + \frac{1}{2}\operatorname{trace}\{\tilde{W}_{ia}^T\tilde{W}_{ia}\} + \frac{1}{2}\sum_{j\in\mathcal{N}_i}\operatorname{trace}\{\tilde{W}_{ja}^T\tilde{W}_{ja}\} + \frac{1}{2}\operatorname{trace}\{\tilde{W}_{id}^T\tilde{W}_{id}\} + \frac{1}{2}\sum_{j\in\mathcal{N}_i}\operatorname{trace}\{\tilde{W}_{jd}^T\tilde{W}_{jd}\}\Big),$$

Similar, to the dynamics of V, derived for above, we can get,

$$\dot{\mathcal{V}} = \sum_{i=1}^{N}\Big(\frac{\partial V_i^*(e_i)}{\partial e_i}^T\Big(Ae_i + (d_i+g_i)(B_i\hat{u}_i + D_i\hat{v}_i) - \sum_{j\in\mathcal{N}} a_{ij}(B_j\hat{u}_j + D_j\hat{v}_j)\Big) + \tilde{W}_{ic}^T\dot{\tilde{W}}_{ic} + \operatorname{trace}\{\tilde{W}_{ia}^T\dot{\tilde{W}}_{ia}\} + \sum_{j\in\mathcal{N}_i}\operatorname{trace}\{\tilde{W}_{ja}^T\dot{\tilde{W}}_{ja}\} + \operatorname{trace}\{\tilde{W}_{id}^T\dot{\tilde{W}}_{id}\} + \sum_{j\in\mathcal{N}_i}\operatorname{trace}\{\tilde{W}_{jd}^T\dot{\tilde{W}}_{jd}\}\Big),$$

The above equations can be separated into four different terms,

$$\dot{\mathcal{V}} = \sum_{i=1}^{N}(T_{i1} + T_{i2} + T_{i3} + T_{i4}),$$

$$T_{i1} := \frac{\partial V_i^*(e_i)}{\partial e_i}^T\Big(Ae_i - (d_i+g_i)B_i\tilde{W}_{ia}^Te_i - (d_i+g_i)D_i\tilde{W}_{id}^Te_i + (d_i+g_i)B_iu_i^* + (d_i+g_i)D_iv_i^* + \sum_{j\in\mathcal{N}} a_{ij}(B_j\tilde{W}_{ja}^Te_j - B_ju_j^*) + \sum_{j\in\mathcal{N}} a_{ij}(D_j\tilde{W}_{jd}^Te_j - D_jv_j^*)\Big),$$

which is very similar to the Hamilton-Jacobi equations. Since, we know that the Hamiltonian is zero, at the optimal value, we subtract "zero" vis-à-vis, the optimal Hamiltonian from the above equation.

$$T_{i1} = -\frac{1}{2}\sum_{j\in\mathcal{N}_i}(d_j+g_j)^2e_j^TP_j(B_jR_{jj}^{-T}R_{ij}R_{jj}^{-1}B_j^T - \frac{\gamma_{ij}^2}{\gamma_{jj}^4}D_jD_j^T)P_je_j - \frac{1}{2}(d_i+g_i)^2e_i^TP_i(B_iR_{ii}B_i^T - \frac{1}{\gamma_{ii}^2}D_iD_i^T)P_ie_i - \frac{1}{2}e_i^TH_ie_i + \frac{\partial V_i^*(e_i)}{\partial e_i}^T\Big(\sum_{j\in\mathcal{N}} a_{ij}(B_j\tilde{W}_{ja}^T + D_j\tilde{W}_{jd}^T)e_j - (d_i+g_i)(B_i\tilde{W}_{ia}^T + D_i\tilde{W}_{id}^T)e_i\Big). \quad (62)$$

$$T_{i3} := -\alpha_{ia}\operatorname{trace}\{\tilde{W}_{ia}^Te_ie_i^T\tilde{W}_{ia} + \tilde{W}_{ia}^Te_ie_i^T\tilde{Q}^i_{\mathrm{e_i u_i}}R_{ii}^{-1}\}$$

$$T_{i4} := -\alpha_{id}\operatorname{trace}\{\tilde{W}_{id}^Te_ie_i^T\tilde{W}_{id} + \tilde{W}_{ia}^Te_ie_i^T\tilde{Q}^i_{\mathrm{e_i v_i}}\gamma_{ii}^{-2}\} - \sum_{j\in\mathcal{N}_i}\alpha_{jd}\operatorname{trace}\{\tilde{W}_{jd}^Te_je_j^T\tilde{W}_{jd} + \tilde{W}_{jd}^Te_je_j^T\tilde{Q}^j_{\mathrm{e_j v_j}}\gamma_{jj}^{-2}\}.$$

The term can be upper bounded after using Young's inequality,

$$T_{i1} \leqslant -\Big(\underline{\lambda}\big(H_i + Q^i_{\mathrm{e_i u_i}}R_{ii}^{-1}(Q^i_{\mathrm{e_i u_i}})^T\big) - \frac{1}{2}\bar{\lambda}\big(Q^i_{\mathrm{e_i u_i}}(Q^i_{\mathrm{e_i u_i}})^T + \frac{2}{\gamma_{ii}^2}Q^i_{\mathrm{e_i v_i}}(Q^i_{\mathrm{e_i v_i}})^T\big)\Big)\|e_i\|^2 - \sum_{j\in\mathcal{N}_i}\Big(\underline{\lambda}\big(Q^j_{\mathrm{e_j u_j}}R_{jj}^{-1}R_{ij}R_{jj}^{-1}(Q^j_{\mathrm{e_j u_j}})^T\big) - \frac{1}{2}\bar{\lambda}\big(Q^i_{\mathrm{e_i u_j}}(Q^i_{\mathrm{e_i u_j}})^T + \frac{\gamma_{ij}^2}{\gamma_{jj}^4}Q^i_{\mathrm{e_i v_j}}(Q^i_{\mathrm{e_i v_j}})^T\big)\Big)\|e_j\|^2 + \frac{1}{2}\left\|e_i^T\tilde{W}_{ia}\right\|^2 + \frac{1}{2}\sum_{j\in\mathcal{N}_i}\left\|e_j^T\tilde{W}_{ja}\right\|^2 + \frac{1}{2}\left\|e_i^T\tilde{W}_{id}\right\|^2 + \frac{1}{2}\sum_{j\in\mathcal{N}_i}\left\|e_j^T\tilde{W}_{jd}\right\|^2, \quad (63)$$

$$T_{i2} \leqslant -\frac{\alpha_{ic}}{4}\left\|\tilde{W}_{ic}\right\|^2.$$

$$T_{i3} = -\alpha_{ia}\operatorname{trace}\{\tilde{W}_{ia}^Te_ie_i^T\tilde{W}_{ia}\} - \alpha_{ia}\operatorname{trace}\{\tilde{W}_{ia}^Te_ie_i^T\tilde{Q}^i_{\mathrm{e_i u_i}}R_{ii}^{-1}\} - \sum_{j\in\mathcal{N}_i}\alpha_{ja}\operatorname{trace}\{\tilde{W}_{ja}^Te_je_j^T\tilde{W}_{ja} + \tilde{W}_{ja}^Te_je_j^T\tilde{Q}^j_{\mathrm{e_j u_j}}R_{jj}^{-1}\}$$
$$\leqslant -\frac{\alpha_{ia}}{2}\left\|e_i^T\tilde{W}_{ia}\right\|^2 + \frac{\alpha_{ia}\delta_i}{2}\bar{\lambda}(R_{ii}^{-1})\|e_i\|^2 - \sum_{j\in\mathcal{N}_i}\frac{\alpha_{ja}}{2}\left\|e_j^T\tilde{W}_{ja}\right\|^2 + \sum_{j\in\mathcal{N}_i}\frac{\alpha_{ja}\delta_j}{2}\bar{\lambda}(R_{jj}^{-1})\|e_j\|^2,$$

$$T_{i4} = -\alpha_{id}\operatorname{trace}\{\tilde{W}_{id}^Te_ie_i^T\tilde{W}_{id}\} - \alpha_{id}\operatorname{trace}\{\tilde{W}_{id}^Te_ie_i^T\tilde{Q}^i_{\mathrm{e_i v_i}}\gamma_{ii}^{-2} - \sum_{j\in\mathcal{N}_i}\alpha_{jd}\operatorname{trace}\{\tilde{W}_{jd}^Te_je_j^T\tilde{W}_{jd} + \tilde{W}_{jd}^Te_je_j^T\tilde{Q}^j_{\mathrm{e_j v_j}}\gamma_{jj}^{-2}\}$$
$$\leqslant -\frac{\alpha_{id}}{2}\left\|e_i^T\tilde{W}_{id}\right\|^2 + \frac{\alpha_{id}\delta_i}{2\gamma_{ii}^2}\|e_i\|^2 - \sum_{j\in\mathcal{N}_i}\frac{\alpha_{jd}}{2}\left\|e_j^T\tilde{W}_{jd}\right\|^2 + \sum_{j\in\mathcal{N}_i}\frac{\alpha_{jd}\delta_j}{2\gamma_{jj}^2}\|e_j\|^2.$$

$$T_{i2} := -\alpha_{ic}\tilde{W}_{ic}^{T}\frac{\Big(U_i(t)\otimes U_i(t)-U_i(t-T)\otimes U_i(t-T)\Big)\Big(U_i(t)\otimes U_i(t)-U_i(t-T)\otimes U_i(t-T)\Big)^{T}}{\Big(1+\Big(U_i(t)\otimes U_i(t)-U_i(t-T)\otimes U_i(t-T)\Big)^{T}\Big(U_i(t)\otimes U_i(t)-U_i(t-T)\otimes U_i(t-T)\Big)\Big)^{2}}\tilde{W}_{ic},$$

Since all of these terms are bounded, when you combine the terms,

$$\begin{aligned}\dot{\mathcal{L}} \leqslant \sum_{i=1}^{N}\Bigg(&-\Big(\underline{\lambda}(H_i+Q^i_{e_iu_i}R_{ii}^{-1}(Q^i_{e_iu_i})^T)\\ &-\frac{1}{2}\bar{\lambda}(Q^i_{e_iu_i}(Q^i_{e_iu_i})^T+\frac{2}{\gamma_{ii}^2}Q^i_{e_iv_i}(Q^i_{e_iv_i})^T)\\ &-\frac{\alpha_{ia}\delta_i}{2}\bar{\lambda}(R_{ii}^{-1})-\frac{\alpha_{id}\delta_i}{2\gamma_{ii}^2}\Big)\|e_i\|^2\\ &-\sum_{j\in\mathcal{N}_i}\Big(\underline{\lambda}(Q^j_{e_ju_j}R_{jj}^{-1}R_{ij}R_{jj}^{-1}(Q^j_{e_ju_j})^T)\\ &-\frac{1}{2}\bar{\lambda}(Q^i_{e_iu_j}(Q^i_{e_iu_j})^T+\frac{\gamma_{ij}^2}{\gamma_{jj}^4}Q^i_{e_iv_j}(Q^i_{e_iv_j})^T)\\ &-\frac{\alpha_{ja}\delta_j}{2}\bar{\lambda}(R_{jj}^{-1})-\frac{\alpha_{jd}\delta_j}{2\gamma_{jj}^2}\Big)\|e_j\|^2\\ &-\frac{\alpha_{ic}}{4}\left\|\tilde{W}_{ic}\right\|^2-\frac{1}{2}(\alpha_{ia}-1)\left\|e_i^T\tilde{W}_{ia}\right\|^2\\ &-\frac{1}{2}\sum_{j\in\mathcal{N}_i}(\alpha_{ja}-1)\left\|e_j^T\tilde{W}_{ja}\right\|^2\\ &-\frac{1}{2}(\alpha_{id}-1)\left\|e_i^T\tilde{W}_{id}\right\|^2-\frac{1}{2}\sum_{j\in\mathcal{N}_i}(\alpha_{jd}-1)\left\|e_j^T\tilde{W}_{jd}\right\|^2\Bigg),\end{aligned}$$

We see that the Lyapnov s bounded and negative, and hence the closed-loop stability is proved.

References

1. H_∞-Optimal Control and Related Minimax Design Problems, 2nd Edition, Basar, Pierre Bernhard

Commentary on the Derivations

I have understood the complete paper and attempted the derivation of each and every step. The only thing that I haven't understood is the application of the Young's inequality for finding the upper bound on the Ti1 in Theorem 2.