Sparse Linear Injection Attack on Multi-Agent Consensus Control Systems

Kam Fai Elvis Tsang, Mengyu Huang, Karl Henrik Johansson, Ling Shi

***Abstract*— This paper investigates the problem of false data injection attack on the communication channels in a multi-agent system executing a consensus protocol. We formulate a non-convex optimisation problem for an attack strategy with minimal one-step attack energy to guarantee instability of the consensus system. We propose an algorithm based on ADMM to solve the problem efficiently in case standard solvers are not available. Numerical simulations are provided to illustrate the effectiveness of the attack strategy.**

***Index Terms*— Multi-agent systems, networked control systems, integrity attack, optimization**

I. Introduction

MULTI-AGENT systems have gained much attention in both academic and industrial communities thanks to its vast potential in various areas, including logistic management, distributed computing and robotics, to name but a few. The consensus problem refers to the objective for a set of agents to reach a mutually agreeable state, i.e., consensus. This is particularly useful in applications such as multi-vehicular networks, formation control and distributed optimisation.

The control protocol for the multi-agent consensus problem has been well studied in the past decade [1–5]. Olfati-Saber *et al.* [1, 2] introduced a consensus protocol for both discrete and continuous time multi-agent control systems, which have been the foundations for much work in this area. While the protocol can achieve consensus exponentially, it requires continuous communication among agents and is hence impractical in many applications. To resolve this, event-triggered communication protocols have been proposed under which the agents only exchange information when a certain event-based threshold is exceeded. Dimarogonas *et al.* [6] proposed a centralised event-triggered mechanism and a self-triggered counterpart that does not require continuous tracking of neighbour information. Yi *et al.* [7] proposed a dynamic event-triggering law that further reduces the communications with a dynamic threshold function.

Kam Fai Elvis Tsang, Mengyu Huang and Ling Shi are with the Department of Electronic and Computer Engineering, Hong Kong University of Science and Technology, Hong Kong. Email: {kftsang, mhuangak, eesling}@ust.hk

Karl Henrik Johansson is with the Digital Futures and School of Electrical Engineering and Computer Science. KTH Royal Institute of Technology, SE-100 44 Stockholm, Sweden. Email: kallej@kth.se

The work by Kam Fai Elvis Tsang, Mengyu Huang and Ling Shi is supported by a Hong Kong RGC General Research Fund 16210619.

While much research effort on multi-agent systems focused on control and event-triggering protocol, fewer works considered security-related problems such as jamming and false data injection attacks. Kikuchi *et al.* [8] considered malicious jamming attack and showed that the system can still reach consensus under the proposed algorithm. Xu *et al.* [9] considered a similar problem and proposed a self-triggered protocol able to handle unreliable networks. Sundaram and Hadjicostis [10] considered the multi-agent consensus problem with malicious agents. LeBlanc *et al.* [11] considered a similar problem but with restriction of using only local information. The literature on injection attacks on multi-agent system is limited. Ma *et al.* [12] considered injection attacks to the communication channels but limited to zero-mean random attacks. Most emphasis in the literature has been on the perspective of the agents instead of the adversaries. It is therefore of interest to investigate false data injection attacks on the communication channels in multi-agent systems with specific adversarial models capturing capabilities and resources.

In this paper, we consider an injection attack problem for a discrete-time multi-agent consensus system. We first analyse the stability of the consensus error dynamics and derive condition to deprive the system of convergence. We show that this condition can be formulated as an optimisation problem. The main contribution of this paper are twofold:

1) We propose an optimisation-based approach to design an injection attack strategy to a multi-agent system executing a consensus control protocol and
2) We transform the non-convex optimisation problem to a convex problem with strongly convex objective that can be efficiently solved with alternating direction method of multipliers (ADMM).

The remainder of the paper is organised as follows. We introduce the mathematical preliminaries in Section II. We formulate the attack problem as an non-convex optimisation problem to minimise the one-step attack energy under attack constraints and conditions for non-convergence of the consensus protocol in Section III. We present an algorithm to transform and relax the optimisation problem to a convex one that can be solved efficiently with guaranteed feasibility in Section IV. Section V provides numerical simulations to illustrate the effectiveness of the proposed algorithm and presents some insights into the results. Finally, Section VI concludes the paper and discusses some future directions.

II. Preliminaries

A. Notations

We denote an $n \times n$ identity matrix by I_n and a vector with all entries being 1 by $\mathbf{1}_n$. For any two matrices X, Y, the operation $X \otimes Y$ represents the Kronecker product. The function $\|x\|_0$ is the cardinality operator to count the number of zero entries in x. For any function $g(x)$, the notation $g(x)^+$ denotes $\max\{0, g(x)\}$ element-wise. For any matrix M, $\lambda_i(M)$ denotes the i-th largest eigenvalue and $\rho(M) = \max_i |\lambda_i(M)|$ is the spectral radius.

B. Algebraic Graph Theory

We represent a multi-agent system with N agents by a weighted, undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where node $i \in \mathcal{V} = \{1, 2, \ldots, N\}$ represents agent i and the edge $(i, j) \in \mathcal{E}$ is a bidirectional communication link between agents i and j. The adjacency matrix $G = [G_{ij}]$ is used to characterise the graph with G_{ij} being the weights of the edge $(i, j) \in \mathcal{E}$ and $G_{ij} = 0$ if $(i, j) \notin \mathcal{E}$. In addition, we assume that there is no self-loop in the graph $\mathcal{G}$, or in other words, $G_{ii} = 0$ for all $i \in \mathcal{V}$. The Laplacian matrix $L = [L_{ij}]$ is defined as $L_{ii} = \sum_{m=1}^{N} G_{im}$ and $L_{ij} = -G_{ij}$ for $i \neq j$.

III. Problem Formulation

Consider the problem of injection attack on a multi-agent consensus control systems, where each agent is a scalar linear time-invariant (LTI) system:

$$x_{k+1}^i = Ax_k^i + u_k^i + w_k^i \tag{1}$$
$$y_k^i = Cx_k^i + v_k^i \tag{2}$$

where $x_k^i, u_k^i, y_k^i \in \mathbb{R}$ are the state, control input and measurement of agent i at time k respectively. The noises w_k^i and v_k^i are i.i.d. Gaussian random variables with covariances Q_i and R_i respectively. We assume the system is (A, C)-observable. At each time k, each agent obtains the minimum mean-square-error (MMSE) estimate of its own state, denoted $\hat{x}_k^i$, by a Kalman filter. Let $e_k^i = x_k^i - \hat{x}_k^i$. It is well known that $\mathbb{E}\left[e_k^i\right] = 0$ with $P_k^i = \mathbb{E}[e_k^i e_k^{i\,T}]$ given by the following recursion:

$$P_{k+1}^i = P_{k+1|k}^i - P_{k+1|k}^i C^T \left(CP_k^i C^T + R_i\right)^{-1} CP_{k+1|k}^i$$

where $P_{k+1|k}^i = AP_k^i A^T + Q_i$. We make the standing assumption that $\mathcal{G}$ is connected. We consider the following multi-agent consensus control protocol:

$$u_k^i = \epsilon \sum_{j=1}^{N} G_{ij}(\tilde{x}_k^j - \hat{x}_k^i) + (1 - A)\hat{x}_k^i \tag{3}$$

where $\tilde{x}_k^j$ is the MMSE estimate of x_k^j received by the neighbours of agent j. The parameter ϵ is supposed to fulfil the condition $\epsilon < (\max_i L_{ii})^{-1}$ which is sufficient for the system to reach consensus exponentially in the absence of disturbances [2]. Similar to the man-in-the-middle attack model [13–16], we consider the case where an attacker can gain access and has the ability to alter the data transmitted by any K agents at each k. Let z_k^j be the attack input on the broadcast data from agent j, then

$$\tilde{x}_k^j = \hat{x}_k^j + z_k^j \tag{4}$$
$$u_k^i = \epsilon \sum_{j=1}^{N} G_{ij}(\hat{x}_k^j - \hat{x}_k^i + z_k^j) + (1 - A)\hat{x}_k^i \tag{5}$$
$$= -\epsilon \sum_{j=1}^{N} L_{ij}\hat{x}_k^j + \epsilon \sum_{j=1}^{N} G_{ij} z_k^j + (1 - A)\hat{x}_k^i \tag{6}$$

We further let $x_k = \text{vec}([x_k^1 \quad \cdots \quad x_k^N])$ and similarly for $\hat{x}_k$, u_k, w_k, v_k, y_k, e_k, z_k. Inspired by the work of Guo *et al.* [14], we restrict the form of attack input to a linear false data injection attack strategy:

$$z_k = T_k \hat{x}_k \tag{7}$$

where $T_k \in \mathbb{R}^{N \times N}$ is the attack matrix to be designed and $T_k = 0$ corresponds to no attack. We then obtain the compact form of the system dynamics as follows:

$$\begin{aligned} x_{k+1} &= \hat{x}_k + (I_N \otimes A)(x_k - \hat{x}_k) - \epsilon L\hat{x}_k + \epsilon G T_k \hat{x}_k + w_k \\ &= (I_N - F_k)x_k + (F_k + (A - 1)I_N)e_k + w_k \end{aligned} \tag{8}$$

where $F_k = \epsilon(L - GT_k)$. The linear attack strategy makes it possible to incorporate the attack strategy in the internal dynamics of the agents, as shown in (8). This is conducive to deriving a necessary condition for the system to be internally unstable. In addition, if this structurally simplistic attack strategy is effective, a more complex strategy may not be necessary to use for the adversary. Let $\varepsilon_k = x_k - \bar{x}_0 \mathbf{1}_N$ where $\bar{x}_0 = \frac{1}{N}\sum_{i=1}^{N} x_0^i$ is the initial average state of all agents. The objective of the system is to reach average consensus, i.e., $\varepsilon_k \to 0$ or become as small as possible due to the presence of disturbances. We assume throughout this paper that $\bar{x}_0 = 0$ without loss of generality due to linearity. We can rewrite (8) as follows:

$$\varepsilon_{k+1} = (I_N - F_k)\varepsilon_k + (F_k + (A - 1)I_N)e_k + w_k \tag{9}$$

The system (9) is internally stable only if $\rho(I_N - F_k) < 1$ leading to consensus in expectation, i.e., $\lim_{k\to\infty} \mathbb{E}\left[\varepsilon_k\right] = 0$. It is however important to note that the violation of this condition does not necessarily mean divergence, or even lack of convergence.

We aim to design T_k such that the system is unstable while there are at most K agents being attacked at each time k. In addition, we would like to minimise the one-step attack energy $\|z_k\|_2^2$. In view of this, we consider the following optimisation problem:

$$\begin{aligned} \min_{T_k} \quad & \|T_k \hat{x}_k\|_2^2 \\ \text{s.t.} \quad & \rho(I_N - F_k) \geq 1 \\ & \|r_k\|_0 \leq K \end{aligned} \tag{P1}$$

where $r_k \in \{0, 1\}^N$ is a binary vector indicating whether or not each agent is attacked at time k, i.e., $r_{k,i} = 0$ if $T_{k,ij} = 0$ for all j and $r_{k,i} = 1$ otherwise while $r_{k,i}, T_{k,ij}$ are the i-th element of r_k and the (i, j)-th element of T_k, respectively. Both constraints of the optimisation problem (P1) are non-convex. In view of this, relaxation and other techniques are adopted to solve the problem.

IV. Main Results

In this section, we will transform the problem (P1) into a convex optimisation problem that is efficiently solvable with the ADMM algorithm described in [17].

A. Lagrangian Relaxation

Instead of a hard constraint on the cardinality of r_k, we consider a new objective function as a combination of attack energy and $\|r_k\|_0$, i.e.,

$$\begin{aligned} \min_{T_k} \quad & \|T_k\hat{x}_k\|_2^2 + \beta\|r_k\|_0 \\ \text{s.t.} \quad & \rho(I_N - F_k) \geq 1 \end{aligned} \tag{P2}$$

where $\beta > 0$ is a penalty weight on the cardinality of r_k. The cardinality term in the objective can be replaced by an equivalent one expressed in T_k. Note that

$$\text{diag}\left(T_k T_k^T\right) = \left[\sum_{j=1}^N T_{k,1j}^2 \cdots \sum_{j=1}^N T_{k,Nj}^2\right]^T$$

From the definition of r_k, we have the convenient equality $\|r_k\|_0 = \|\text{diag}\left(T_k T_k^T\right)\|_0$. The optimisation problem (P2) can then be rewritten as follows:

$$\begin{aligned} \min_{T_k} \quad & \|T_k\hat{x}_k\|_2^2 + \beta\|\text{diag}\left(T_k T_k^T\right)\|_0 \\ \text{s.t.} \quad & \rho(I_N - F_k) \geq 1 \end{aligned} \tag{P3}$$

B. Majorisation-Minimisation Algorithm

For the relaxed problem (P3), the objective function includes a non-convex cardinality term. We approximate the cardinality operator to one that is easier to manipulate while maintaining sufficient resemblance. Traditionally, the cardinality operator is approximated by l_1 norm such as in lasso regression. However, it is a poor approximator as shown in Fig. 1. In view of this, we adopt the following approximation [18]:

$$\|x\|_0 \approx f_\delta(x) = 1 - \log_\delta(x + \delta)$$

for scalar x and some $0 < \delta < 1$. Note that $\lim_{\delta\to 0} f_\delta(x) = \|x\|_0$. For simplicity, let $M_k = \text{diag}(T_k T_k^T)$ and $M_{k,i}$ be the i-th element of M_k, in other words, $M_{k,i} = \sum_{j=1}^N T_{k,ij}^2 \geq 0$. Now we have $\|M_k\|_0 \approx \sum_{i=1}^N f_\delta(M_{k,i}) = N - \sum_{i=1}^N \log_\delta(M_{k,i} + \delta)$.

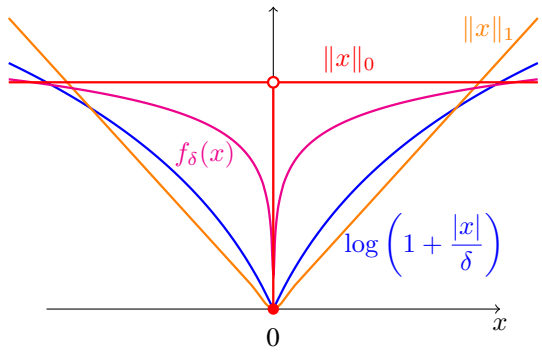


Fig. 1: Approximation of cardinality operator

The objective function remains non-convex as it is a sum of convex and concave functions. To overcome this, we adopt the technique of Majorisation-Minimisation (MM) by exploiting the concavity of the approximation, also known as the reweighted l_1-minimisation method [18]. The basic idea is to replace the non-convex term $\sum_{i=1}^N f_\delta(M_{k,i})$ by a surrogate function and optimise the problem iteratively. As $f_\delta(M_{k,i})$ is concave, its linearisation $s\left(M_{k,i}, M_{k,i}^{(t)}\right)$ is a suitable surrogate:

$$\begin{aligned} & s\left(M_{k,i}, M_{k,i}^{(t)}\right) \\ & = 1 - \log_\delta\left(M_{k,i}^{(t)} + \delta\right) + \frac{\left(\log\delta^{-1}\right)^{-1}}{M_{k,i}^{(t)} + \delta}\left(M_{k,i} - M_{k,i}^{(t)}\right) \end{aligned}$$

where $M_{k,i}^{(t)}$ is the t-th iterative solution of $M_{k,i}$ given the constraints. By replacing the concave term in the objective function by the surrogate and ignoring the constant terms thereof, we then have the following optimisation problem:

$$\begin{aligned} \min_{T_k} \quad & \|T_k\hat{x}_k\|_2^2 + \beta\sum_{i=1}^N w_i^{(t)} M_{k,i} \\ \text{s.t.} \quad & \rho(I_N - F_k) \geq 1 \end{aligned} \tag{P4a}$$

where $w_i^{(t)} = \left(\log\delta^{-1}\left(M_{k,i}^{(t)} + \delta\right)\right)^{-1}$. At each iteration t, the weight $w_i^{(t)}$ increases for smaller values of $M_{k,i}^{(t)}$ and vice versa. The contribution $w_i^{(t)} M_{k,i}$ is minimal when $M_{k,i}^{(t)}$ and $M_{k,i}$ are zero. In other words, the MM algorithm encourages the elements of $M_{k,i}$ to be as small as possible, promoting sparsity in the solution T_k. Recall the definition of $M_{k,i} = \sum_{j=1}^N T_{ij}^2$, we can rewrite (P4a) as

$$\begin{aligned} \min_{T_k} \quad & \|T_k\hat{x}_k\|_2^2 + \beta\|W^{(t)}T_k\|_F^2 \\ \text{s.t.} \quad & \rho(I_N - F_k) \geq 1 \end{aligned} \tag{P4b}$$

with convex objective and $W^{(t)} = \text{diag}\left(w^{(t)}\right)^{1/2}$.

C. Spectral Radius Constraint

To eliminate the last non-convex constraint, we construct a sufficient condition to ensure feasibility and prove that the new constraint is always feasible, thus not over-restrictive. It is straightforward to show that

$$\rho(I_N - F_k) \geq \frac{1}{N}\sum_{i=1}^N \lambda_i(I_N - F_k) \tag{10}$$

$$= 1 + \frac{\epsilon}{N}\left(\text{Tr}(GT_k) - \text{Tr}(L)\right) \tag{11}$$

A sufficient condition for $\rho(I_N - F_k) \geq 1$ is therefore $\text{Tr}(GT_k) \geq \text{Tr}(L)$. As a result, we have

$$\begin{aligned} \min_{T_k} \quad & \|T_k\hat{x}_k\|_2^2 + \beta\|W^{(t)}T_k\|_F^2 \\ \text{s.t.} \quad & \text{Tr}(GT_k) \geq \text{Tr}(L) \end{aligned} \tag{P5}$$

We have yet to show that problem (P5) is feasible when the original problem (P1) is feasible.

Proposition 1. *Problem (P5) is always feasible.*

Proof. Since $\mathcal{G}$ is connected, there exists an index pair (l, m) where $l \neq m$, $l, m = 1, 2, \ldots, N$ such that $G_{lm} \neq 0$. Now

consider the most restricted form of T_k:

$$T_{k,ij} = \begin{cases} T, & (i,j) = (l,m),\ l \neq m \\ 0, & \text{otherwise} \end{cases}$$

for some $T \neq 0$. Then we have $\text{Tr}(GT_k) = G_{lm}T$. If we assign T such that $T \geq {G_{lm}}^{-1}\text{Tr}(L) = {G_{lm}}^{-1}\sum_{i=1}^{N}\sum_{j=1}^{N} G_{ij}$ then the constraint $\text{Tr}(GT_k) \geq \text{Tr}(L)$ in (P5) is satisfied. Thus (P5) is always feasible. □

D. Additional Ad-Hoc Constraints

To study specific attack scenarios, it is sometimes relevant to include additional constraints in (P5). For example, we can include the cosntraint $(I_N - F_k)\mathbf{1} \leq -\mathbf{1}$ or $(I_N - F_k)\mathbf{1} \geq (1+\varsigma)\mathbf{1}$ to ensure that the system cannot reach non-average consensus for some $\varsigma > 0$. Should the agents reach consensus at $x_{k_0} = c\mathbf{1}$, the agents will oscillate in states for the former constraint as

$$\begin{aligned} x_{k_0} &= c\mathbf{1} \\ x_{k_0+1} &= (I_N - F_k)x_{k_0} \leq -c\mathbf{1} + \tilde{w}_{k_0} \\ x_{k_0+2} &= (I_N - F_k)x_{k_0+1} \geq c\mathbf{1} - cw_{k_0} + \tilde{w}_{k_0+1} \\ &\vdots \end{aligned}$$

for $c > 0$ with the inequality signs reversed for $c < 0$ where $\tilde{w}_k = (F_k + \tilde{A})w_k$. As for the latter constraint, the agents will diverge with dynamics

$$x_k \geq c(1+\varsigma)^{k-k_0}\mathbf{1} + \sum_{i=1}^{k-k_0}(1+\varsigma)^{k-k_0-i}\tilde{w}_{k_0+i-1}$$

for $c > 0$ with the signs reversed for $c < 0$. Since $L\mathbf{1} = 0$, we obtain two variants of (P5):

$$\begin{aligned} \min_{T_k} \quad & \|T_k\hat{x}_k\|_2^2 + \beta\|W^{(t)}T_k\|_F^2 \\ \text{s.t.} \quad & \text{Tr}(L - GT_k) \leq 0 \\ & -(GT_k + \varsigma I_N)\,\mathbf{1} \leq 0 \end{aligned} \tag{P6a}$$

$$\begin{aligned} \min_{T_k} \quad & \|T_k\hat{x}_k\|_2^2 + \beta\|W^{(t)}T_k\|_F^2 \\ \text{s.t.} \quad & \text{Tr}(L - GT_k) \leq 0 \\ & (GT_k + 2\epsilon^{-1}I_N)\mathbf{1} \leq 0 \end{aligned} \tag{P6b}$$

E. ADMM

We present an ADMM algorithm to solve (P6a) and (P6b) adopting an extension to ADMM. We first reformulate problem (P6a) and (P6b) as follows:

$$\begin{aligned} \min_{T_{k,i},Z} \quad & \sum_{i=1}^{2}\left(\|T_{k,i}\hat{x}_k\|_2^2 + \beta\|W^{(t)}T_{k,i}\|_F^2\right) \\ \text{s.t.} \quad & \text{Tr}(L - GT_{k,1})^+ = 0 \\ & g(T_{k,2})^+ = 0 \\ & T_{k,i} = Z \quad \forall i \in \{1,2\} \end{aligned} \tag{P7}$$

where $g(T_k) = -(GT_k + \varsigma I_N)\mathbf{1}$ for problem (P6a) and $g(T_k) = (\epsilon GT_k + 2I_N)\mathbf{1}$ for (P6b). The augmented Lagrangian for (P7) is

$$\begin{aligned} & L_\eta(T_{k,i}, Z, \mu_i, \Lambda_i) \\ &= \sum_{i=1}^{N}\left(\|T_{k,i}\hat{x}_k\|_2^2 + \beta\|W^{(t,m)}T_{k,i}\|_F^2 + \text{Tr}\left(\Lambda_i^T(T_{k,i} - Z)\right)\right) \\ &\quad + \frac{\eta}{2}\left(\left(\text{Tr}(L - GT_{k,1})^+\right)^2 + \|g(T_{k,2})^+\|_2^2\right) \\ &\quad + \frac{\eta}{2}\sum_{i=1}^{N}\text{Tr}\left((T_{k,i} - Z)^T(T_{k,i} - Z)\right) \\ &\quad + \mu_1\text{Tr}(L - GT_{k,1})^+ + \mu_2^T g(T_{k,2})^+ \end{aligned}$$

where μ_i, Λ_i are Lagrangian multipliers (dual variables) and $\eta > 0$ is the step size for dual ascents. We can then solve (P7) with the following update rules

$$\begin{aligned} T_{k,i}^{(t,m+1)} &= \arg\min_{T_{k,i}} L_\eta\left(T_{k,i}, Z^{(t,m)}, \mu_i^{(t,m)}, \Lambda_i^{(t,m)}\right) \\ Z^{(t,m+1)} &= \arg\min_{Z} L_\eta\left(T_{k,i}^{(t,m+1)}, Z, \mu_i^{(t,m)}, \Lambda_i^{(t,m)}\right) \\ \mu_1^{(t,m+1)} &= \mu_1^{(t,m)} + \eta\text{Tr}\left(L - GT_{k,1}^{(t,m+1)}\right)^+ \\ \mu_2^{(t,m+1)} &= \mu_2^{(t,m)} + \eta g\left(T_{k,2}^{(t,m+1)}\right)^+ \\ \Lambda_i^{(t,m+1)} &= \Lambda_i^{(t,m)} + \eta\left(T_{k,i}^{(t,m+1)} - Z^{(t,m+1)}\right) \end{aligned}$$

Theorem 1. *The updates rule for $T_{k,i}$ is given by*

$$T_{k,i}^{(t,m+1)} = \textit{vec}^{-1}\left(S^{-1}\,\textit{vec}\left(\eta Z^{(t,m)} - \Lambda_i^{(t,m)}\right)\right)$$

if the above solution satisfies the constraints $\text{Tr}(L - GT_{k,1}) \leq 0$ and $g(T_{k,2}) \leq 0$ respectively, where $S = \left(2\beta I_N \otimes W^{(t)} + (2X_k + \eta I_N) \otimes I_N\right)$. Otherwise, it is the solutions to the nonlinear matrix equations

$$\begin{aligned} & T_{k,1}(2X_k + \eta I_N) + 2\beta W^{(t)}T_{k,1} + \Lambda_1^{(t,m)} - \eta Z^{(t,m)} \\ & \qquad - (\mu_1 + \eta h_\alpha(\text{Tr}(L - GT_{k,1})))\,H_\alpha(T_{k,1})G = 0 \\ & T_{k,2}(2X_k + \eta I_N) + 2\beta W^{(t)}T_{k,2} + \Lambda_2^{(t,m)} - \eta Z^{(t,m)} \\ & \qquad + \sum_{i=1}^{N}(\mu_2 + \eta h_\alpha\left(g(T_{k,2})_i\right))\,H_{\alpha,2i}(T_{k,2})\hat{G}_i = 0 \end{aligned}$$

where $h_\alpha(z) = \alpha^{-1}\ln(1 + \exp(\alpha z))$, with derivative $h'_\alpha(z) = \exp(\alpha z)/(1 + \exp(\alpha z))$, $H_\alpha(T) = h'_\alpha(\text{Tr}(L - GT_{k,1}))$, $H_{\alpha,2i}(T) = h'_\alpha(g(T_{k,2})_i)$, $[\hat{G}_i]_{mn} = -G_{im}$ for (P6a), $[\hat{G}_i]_{mn} = \epsilon G_{im}$ for (P6b) and the update for Z is

$$Z^{(t,m+1)} = \frac{1}{2\eta}\sum_{i=1}^{2}\left(\eta T_{k,i}^{(t,m+1)} + \Lambda_i^{(t,m)}\right)$$

Proof. Since $L_\eta(T_{k,i}, Z, \mu_i, \Lambda_i)$ is strongly convex, it has a unique minimiser. However, it is not a smooth function at $\text{Tr}(L - GT_{k,1}) = 0$ and $g(T_{k,2}) = 0$. In view of this, we replace all $f(\cdot)^+$ by $h_\alpha \circ f(\cdot)$ in the augmented Lagrangian. This transformation is due to $\lim_{\alpha\to\infty} h_\alpha(z) = z^+$ and it becomes strongly convex and smooth in $T_{k,i}$ and Z. Note that any positive value of α leads to a feasible solution, but a larger α results in a more accurate solution and slower

convergence as it becomes less Lipschitz continuous. We can then seek to solve the equations $\nabla_{T_{k,1}} L_\eta(T_{k,i}, Z, \mu_i, \Lambda_i) = 0$ and $\nabla_Z L_{\eta,}(T_{k,i}, Z, \mu_i, \Lambda_i) = 0$ for the unique minimiser:

$$\nabla_Z L_\eta(T_{k,i}, Z, \mu_i, \Lambda_i) = \sum_{i=1}^{2} \eta(Z - T_{k,i}) - \Lambda_i = 0$$

$$Z = \frac{1}{2\eta} \sum_{i=1}^{2} (\eta T_{k,i} + \Lambda_i)$$

Let $X_k = \hat{x}_k \hat{x}_k^T$ and for large α,

$$\begin{aligned}
&\nabla_{T_{k,1}} L_\eta(T_{k,i}, Z, \mu_i, \Lambda_i) \\
&= 2T_{k,1}X_k + 2\beta W^{(t)} T_{k,1} + \mu_1 H_\alpha(T_{k,1}) G^T \\
&\quad + \eta h_\alpha \left(\mathrm{Tr}(L - GT_{k,1})\right) H_\alpha(T_{k,1}) G^T + \Lambda_1 + \eta(T_{k,1} - Z) \\
&= T_{k,1}(2X_k + \eta I_N) + 2\beta W^{(t)} T_{k,1} + \Lambda_1 - \eta Z \\
&\quad + \left(\mu_1 + \eta h_\alpha(\mathrm{Tr}(L - GT_{k,1}))\right) H_\alpha(T_{k,1}) G \qquad (12)
\end{aligned}$$

If the Sylvester equation $T_{k,1}(2X_k + \eta I_N) + 2\beta W^{(t)} T_{k,1} + \Lambda_1 - \eta Z = 0$ has a solution satisfying $\mathrm{Tr}(L - GT_{k,1}) \leq 0$, it is the optimal solution minimising the augmented Lagrangian because the last term would be zero should the condition be met. Also, the above Sylvester equation can be written as $S\mathrm{vec}(T_{k,1}) = \mathrm{vec}(\eta Z - \Lambda_1)$ as S is invertible:

$$T_{k,1} = \mathrm{vec}^{-1}\left(S^{-1}\ \mathrm{vec}(\eta Z - \Lambda_1)\right)$$

If the condition is not satisfied, one would need to solve the nonlinear matrix equation (12) set to 0. Similarly for $T_{k,2}$, if the above solution does not satisfy the constraint $g(T_{k,2}) \leq 0$, one should solve the following nonlinear equation for the update of $T_{k,2}$ from the gradient L_η:

$$\begin{aligned}
&T_{k,2}(2X_k + \eta I_N) + 2\beta W^{(t)} T_{k,2} + \Lambda_2^{(t,m)} \\
&- \eta Z^{(t,m)} + \sum_{i=1}^{N} \left(\mu_2 + \eta h_\alpha\left(g(T_{k,2})_i\right)\right) H_{\alpha,2i}(T_{k,2}) \hat{G}_i = 0
\end{aligned}$$

thus concluding the proof. □

V. Simulation Results

In this section, some simulation results with the proposed linear false data injection attack algorithm are presented. We consider the following graph:

$$G = \begin{bmatrix} 0 & 0 & 4.4 & 2.3 \\ 0 & 0 & 1.9 & 0.2 \\ 4.4 & 1.9 & 0 & 1.1 \\ 2.3 & 0.2 & 1.1 & 0 \end{bmatrix}$$

with $n = 1, A = 0.8, C = 1.2, Q_i = R_i = 0$ for all $i \in \mathcal{V}$.

We first simulate the problem (P5) with $\beta = 100$ and the initial states $x_0 = [1, -1.5, 3.5, -3]^T$ and $W_{ij}^{(0)} = 1$ for all i, j in the finite time horizon $k = 1, 2, \ldots, 100$. The agent states, relative consensus errors and attack inputs are plotted in Fig. 2a-2c. The agents diverged from $\bar{x}_0$ with diminishing rate as k grows but did not converge within the time horizon. The attacker continuously attacked agent 1 and 3 with $\max_k \|z_k\|_2 \approx 4.1$ while the attack inputs gradually decayed for $k \geq 14$.

We then solve problem (P6a) with $\varsigma = 0.1$, shown in Fig. 2d-2f. It can be observed that the agents also diverged but at a much higher rate, approximately 7 times, than (P5). The relative consensus errors $x_k^i - \bar{x}_k$ were bounded within ± 4 while $\max_k \|z_k\| \approx 12.4$. Similar to (P5), the attack input also gradually decayed after $k \geq 44$.

On the other hand, the solution for (P6b) showed an entirely different behaviour where the agents converged to a fixed point at $x_k = [0.6974, 1.0267, -0.3055, -0.1743]^T$ without consensus. The attack input peaked at $\|z_5\|_2 \approx 5.8$ and converged to $z_k = [-1.606, 0, 1.458, 0]^T$ at $k = 35$.

Of the three problems considered in this simulation, (P6b) is the most capable as it generates low attack energy $\|z_k\|_2^2$ while keeping the agents from consensus. In addition, it also keeps the agents state finite to promote stealthiness.

VI. Conclusion and Future Work

We considered a linear injection attack against communication links in the multi-agent consensus protocol. We presented a formulation of the problem as an optimisation problem aiming to minimise one-step attack energy while ensuring instability of the system, as well as an algorithm to solve the non-convex problem efficiently. Numerical examples showed that it is possible to drive the multi-agent system to diverge by attacking only one agent at each time step with bounded average attack power.

A main limitation of this work is that we did not consider any attack detection or countermeasure. This is important as it relates to how the attacker should choose the strategy in order to avoid detection. A plausible direction to overcome this drawback is to formulate the two problems simultaneously in a game setting and seek for the Nash equilibrium to analyse the system performance.

References

[1] R. Olfati-Saber and R. M. Murray, "Consensus problems in networks of agents with switching topology and time-delays," *IEEE Transactions on Automatic Control*, vol. 49, no. 9, pp. 1520–1533, 2004.

[2] R. Olfati-Saber, J. A. Fax, and R. M. Murray, "Consensus and cooperation in networked multi-agent systems," in *Proceedings of the IEEE*, vol. 95, no. 1, 2007, pp. 215–233.

[3] A. Nedić and A. Ozdaglar, "Convergence rate for consensus with delays," *Journal of Global Optimization*, vol. 47, no. 3, pp. 437–456, 2010.

[4] X. Zhang and M. Chen, "Event-triggered consensus for second-order leaderless multi-agent systems," in *2013 25th Chinese Control and Decision Conference (CCDC)*, 2013, pp. 4395–4399.

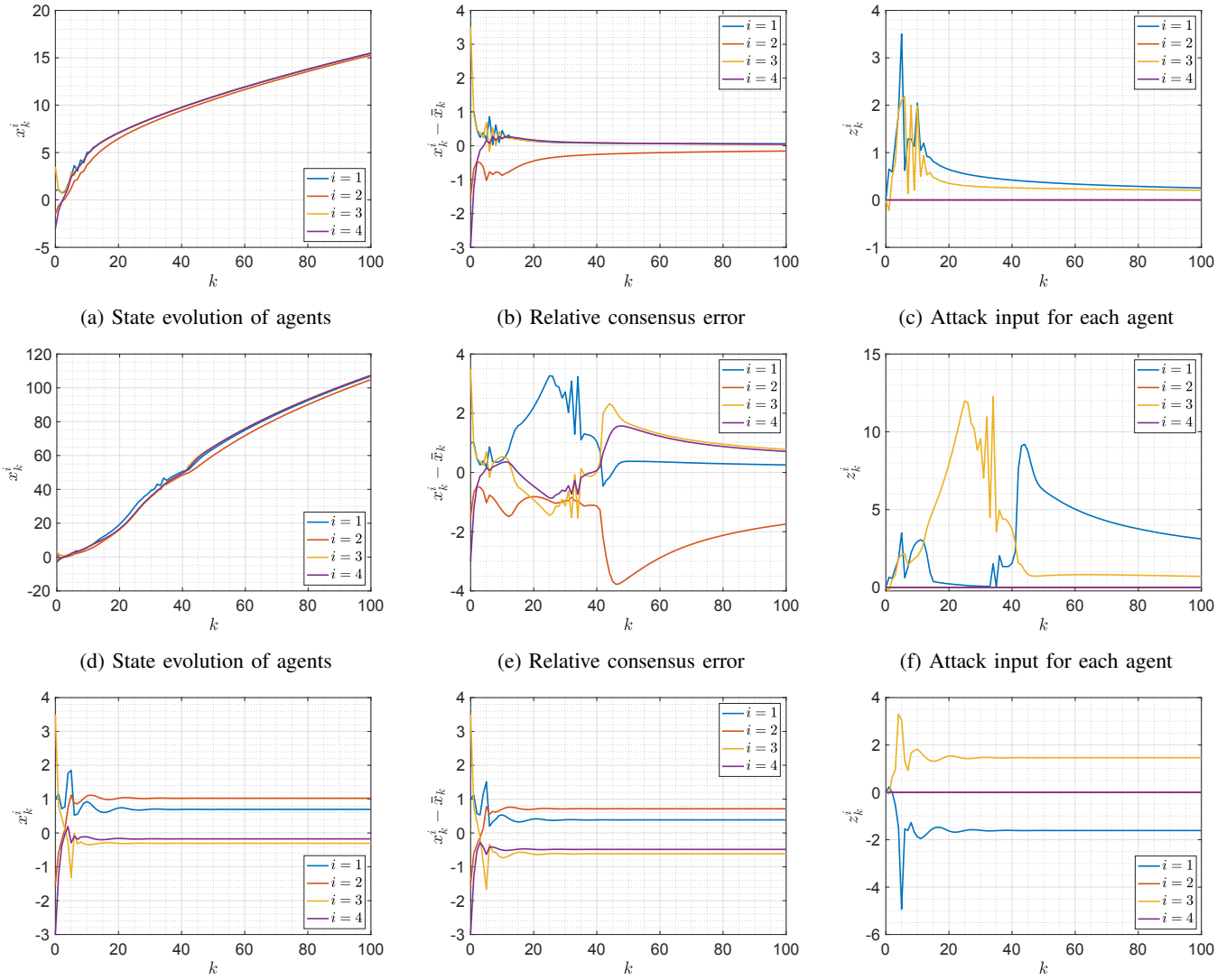
[5] K. F. E. Tsang, J. Wu, and L. Shi, "Zeno-free stochastic distributed event-triggered consensus control for multi-agent systems," in *2019 American Control Conference (ACC)*, Philadelphia, PA, 2019, pp. 778–783.

[6] D. V. Dimarogonas, E. Frazzoli, and K. H. Johansson, "Distributed event-triggered control for multi-agent systems," *IEEE Transactions on Automatic Control*, vol. 57, no. 5, pp. 1291–1297, 2012.

[7] X. Yi, K. Liu, D. V. Dimarogonas, and K. H. Johansson, "Dynamic event-triggered and self-triggered control for multi-agent systems," *IEEE Transactions on Automatic Control*, vol. 64, no. 8, pp. 3300–3307, 2019.

[8] K. Kikuchi, A. Cetinkaya, T. Hayakawa, and H. Ishii, "Stochastic communication protocols for multi-agent consensus under jamming attacks," in *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*, Melbourne, 2017, pp. 1657–1662.

[9] W. Xu, D. W. Ho, J. Zhong, and B. Chen, "Event/self-triggered control for leader-following consensus over unreliable network with dos attacks," *IEEE Transactions on Neural Networks and Learning Systems*, 2019.



(a) State evolution of agents (b) Relative consensus error (c) Attack input for each agent

(d) State evolution of agents (e) Relative consensus error (f) Attack input for each agent

(g) State evolution of agents (h) Relative consensus error (i) Attack input for each agent

Fig. 2: Simulation Results with $\beta = 100$ for (P5): (a-c), (P6a): (d-f) and (P6b): (g-i)

[10] S. Sundaram and C. N. Hadjicostis, "Distributed function calculation via linear iterative strategies in the presence of malicious agents," *IEEE Transactions on Automatic Control*, vol. 56, no. 7, pp. 1495–1508, 2011.

[11] H. J. LeBlanc, H. Zhang, S. Sundaram, and X. Koutsoukos, "Consensus of multi-agent networks in the presence of adversaries using only local information," in *Proceedings of the 1st International Conference on High Confidence Networked Systems*, 2012.

[12] L. Ma, Z. Wang, and Y. Yuan, "Consensus control for nonlinear multi-agent systems subject to deception attacks," in *2016 22nd International Conference on Automation and Computing, ICAC*, 2016, pp. 21–26.

[13] U. Meyer and S. Wetzel, "A man-in-the-middle attack on umts," in *Proceedings of the 2004 ACM Workshop on Wireless Security*, 2004, pp. 90–97.

[14] Z. Guo, D. Shi, K. H. Johansson, and L. Shi, "Worst-case stealthy innovation-based linear attack on remote state estimation," *Automatica*, vol. 89, pp. 117–124, 2018.

[15] Y. Mo, S. Weerakkody, and B. Sinopoli, "Physical authentication of control systems: Designing watermarked control inputs to detect counterfeit sensor outputs," *IEEE Control Systems Magazine*, vol. 35, no. 1, pp. 93–109, 2015.

[16] F. Callegati, W. Cerroni, and M. Ramilli, "Man-in-the-middle attack to the https protocol," *IEEE Security and Privacy*, no. 1, pp. 78–81, 2009.

[17] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2010.

[18] Y.-B. Zhao and D. Li, "Reweighted l1-minimization for sparse solutions to underdetermined linear systems," *SIAM Journal on Optimization*, vol. 22, no. 3, pp. 1065–1088, 2012.