2017 IEEE Conference on Control Technology and Applications (CCTA)
August 27-30, 2017. Kohala Coast, Hawai'i, USA

Cyber-attack Detection and Accommodation Algorithm for Energy Delivery Systems*

Lalit K. Mestha, *Fellow, IEEE*, Olugbenga M. Anubi, *Member, IEEE*,
Masoud Abbaszadeh, *Senior Member, IEEE*

***Abstract*— Cyber-attack accommodation in a cyber-physical system is to ensure system operation, integrity and availability while maintaining a reasonable operational performance under attack. In this paper, we present a novel cyber-attack accommodation algorithm by estimating the true operational states of the system with new boundary & performance constrained resilient estimators while the system is continuously operating and is under attack. Our approach is based on combining data driven machine learning and physics based domain knowledge with traditional resilient estimation. The results were evaluated using a high fidelity model-based simulation environment.**

I. INTRODUCTION

A cyber-attack/anomaly detection and accommodation (ADA) system for energy delivery system should enable automatic detection of cyber-disruptions caused by unintentional or intentional cyber-attacks, effectively localizing where the attack happened at the critical interfaces that are functional part of the active control system, and then attempt to automatically accommodate the system (i.e., continue operating in normal/degraded condition) through various means for uninterrupted operation. Current methods primarily consider threat detection in information technology (IT: e.g., computers that store, retrieve, transmit and manipulate data) and operational technology (OT: e.g., direct monitoring devices and communication bus interfaces). Cyber-threats can still penetrate through these protection layers and reach the physical domain, similar to stealthy attacks of the kind seen in 2010 with Stuxnet attack [1]. Since then targeted attacks with real consequences are on the rise in critical infrastructures around the world [1]–[3]. Such attacks can diminish the performance of the control system, and may cause total shut down or catastrophic damage to the plant.

Industrial processes are controlled by Programmable Logic Controllers (PLC) with Ethernet ports and IP addresses. Computer worms can live in the PLC and be inactive for many days and can replicate itself into many targets as it finds them [4]. If they go undetected for long, it could be troublesome for many systems, as it will be difficult to pinpoint the source of malware. Therefore, protecting control systems from cyber-attacks requires a completely different approach than IT or OT systems. In the new approach, we need to exploit the knowledge of the system and detect the operational behavior as we are controlling in real-time with sensors and actuators similar to how in the human body our immune system detects and removes / destroys threats from numerous pathogens (e.g., viruses, bacteria and parasites).

Some of the papers in the computational intelligence literature have considered developing Artificial Immune System (AIS) for many applications such as anomaly/fault detection, computer security and adaptive control [5], [6] using structure of immune systems. In [7], an immune system inspired real-time Intrusion Detection System (IDS) using unsupervised clustering is proposed for detecting both known and new attacks. Recent literature on cyber-attack detection algorithms have always concentrated on the power grid [8]–[10], with some based on physics-based modeling and others on estimation theory. Fawzi et. al [10] presented formal methods with estimation and control theory for linear systems to detect maximum number of attacks that can be detected and accurately corrected despite attacks in progression. It exploited the redundancies in sensor path. Pasqualetti et. al [11] have characterized fundamental limitations in attack detection and identification and designed attack monitoring filters for linear systems under attack. Numerous other attack resilient observer-based estimation schemes are also described by Kim et. al [12], including estimators for a class of uniformly observable nonlinear systems with sensing redundancy. Ozay, et. al [13] used feature-based learning algorithms to classify measurements as being either normal or attacked without giving considerations for correcting attacked sensors. If the system is unobservable, conventional state estimation methods fail, and learning methods can solve this kind of detection problem. Feature-based large-scale learning methods have shown tremendous improvements in minimizing false-positive rates in anomaly detection for gas turbine combustors [14] with capabilities to apply for heterogeneous sensing nodes for attack detection and accommodation applications.

In this paper, we propose a new solution to the resilient estimation problem. We use a mixture of feature-based learning and state estimation methods to specifically emphasize detection and neutralization of attacks within one sample period. A boundary and performance constrained resilient estimator (BPRE) is designed by using attack boundaries and performance constraints. The BPRE estimates true operational signals on a sample-by-sample basis from corrupted signals due to attacks so that the system always sees normal signal. Our approach is aligned with the defense in depth

*Research supported by US DOEs (Department of Energy) Cyber Security for Energy Delivery Systems (CEDS) R&D Program and GE Global Research.

L.K. Mestha (corresponding author Lalit.Mestha@ge.com, 518-387-6967), O. Anubi (olugbenga.anubi@ge.com) and M. Abbaszadeh (abbaszadeh@ge.com) are with the GE Global Research Center, Niskayuna, NY 12309 USA.

strategy identified in the DOE Roadmap to achieve security [15]. It can work with heterogeneous multi-modal sensing system. Once implemented, it will essentially operate like an industrial version of an immune system.

The paper is organized as follows. Section III begins with the feature-based heterogeneous multi-modal sensing and methods used to compute attack decision boundary. Section IV shows problem formulation and theory of resilient estimator. In Section IV we show main simulation results for few sensor attack cases followed by conclusions in Section VI.

II. PROBLEM DEFINITION AND NOTATION

Considering a cyber-physical system under an adversarial attack in some monitoring node, the goal of this work is to come up with an "accommodation"[1] algorithm which facilitates safe operation of the system while reasonable operational performance is preserved. The accommodation mechanism will be used to guarantee normal operation, integrity and availability of the system. The problem is formulated as an online optimization as explained in Section IV.

The following notions and conventions are employed throughout the paper: $\mathbb{R}, \mathbb{R}^n, \mathbb{R}^{n\times m}$ denote the space of real numbers, real vectors of length n and real matrices of n rows and m columns respectively. $\mathbb{R}_+$ denotes positive real numbers. $X^\top$ denotes the transpose of the quantity X. Normal-face lower-case letters ($x \in \mathbb{R}$) are used to represent real scalars, bold-face lower-case letter ($\mathbf{x} \in \mathbb{R}^n$) represents vectors, while normal-face upper case ($X \in \mathbb{R}^{n\times m}$) represents matrices. For a matrix X, $X^{1:p}$ and $X_{1:p}$ denote the first p rows and columns respectively. For a vector $\mathbf{x}$, $\mathbf{x}_i$ denotes its ith element. $\underline{\mathbf{0}}, \mathbf{O}$ denote the vectors of zeros and matrix of zeros of appropriate dimensions respectively. $\underline{\mathbf{1}}$ denotes the vectors of ones of appropriate dimensions. $Q \succeq 0$ denotes positive semi-definite symmetric matrix, i.e $\mathbf{x}^\top Q\mathbf{x} \geq 0\ \forall \mathbf{x} \neq 0$. $\mathcal{H}$ denotes Hilbert space of continuous functions endowed with the inner product $\langle f, g\rangle \triangleq \int f(t)g(t)dt$.

III. HETEROGENEOUS MULTIMODAL SENSING & ATTACK BOUNDARY COMPUTATION

A. Representing signals from heterogeneous sensors

The proposed sensing approach should handle many types of inputs from multiple heterogeneous data stream in complex hyper connected energy delivery systems. Signals from time domain are converted to features using multi-modal-multi-disciplinary (MMMD) feature discovery framework employed as in machine learning discipline [14]. A feature may refer to, for example, mathematical characterizations of data and is computed in each overlapping batch of data stream. Examples of features as applied to sensor data can be classified broadly into knowledge-based, shallow and deep features. Knowledge-based features use domain or engineering knowledge of physics of the system to create features. These features can be simply statistical descriptors (e.g., max, min, mean, variance), and different orders of moments, calculated over a window of a time-series signal and its corresponding FFT spectrum as well. Shallow features are from unsupervised learning (e.g., k-means clustering), manifold learning and nonlinear embedding (e.g., isoMap, LLE), low dimension projection (e.g., principal component analysis, independent component analysis), and neutral networks, along with genetic programming and sparse coding. Deep learning features can be generated using deep learning algorithms which involves learning good representations of data through multiple levels of abstraction. For simplicity, we will describe the use of Principal Components Analysis (PCA) for extracting features in Section IV.

B. Attack Decision Boundary in Feature Space

In order to make the decision about attack/abnormal behavior, a classifier is trained by considering all the nodes. In the case that both normal and abnormal data is available, a supervised learning technique is used for developing the classifier. In the case that abnormal data is not available (e.g. using legacy normal data from the plant), we use semi-supervised learning techniques. In this study, we have used a classifier based on support vector machines (SVM) both for supervised and semi-supervised learning along with a Gaussian kernel to cover the nonlinearity on the feature space. In this work, we have created both normal and attack data sets using GE ARTEMIS™ high fidelity power plant simulation platform. ARTEMIS is a simulation agent that runs the high fidelity plant dynamics (developed in Easy5), the real-time Mark VIe controllers and the HMI (the operator console) in one unified platform. The decision boundary is a hyperplane in the kernelled feature space of the same dimension as the feature vector. The shape of the decision boundary in the original feature space is nonlinear and could be very complex. Using this kernel trick, we are still able to separate nonlinearly distributed data using linear hyperplanes in the kernelled space. During the real-time operation, the feature vector from monitoring nodes is compared with the pre-stored decision boundary. For each instant of the feature vector, a score is associated with the data, whose sign determines whether the data lies inside the decision boundary (hence normal) or outside (hence attack), while its value represents some notion of closeness to the decision boundary.

IV. ATTACK NEUTRALIZATION WITH RESILIENT ESTIMATOR

A. Attack neutralizer for an industrial immune system

To understand the mechanisms involved in neutralizing cyber-attacks, let us consider a general system (e.g., cyber physical system, software system, bio-mechanical system, network system, communication system, etc.) that contains access to continuous stream of data from monitoring nodes in the form of time series signals. The time series signals might be generated from set of output sensor nodes ('$\mathbf{y}$ '; both physical and/ virtual sensors), set of actuator nodes ('$\mathbf{u}$ '; both hard and/ soft actuators generated from open

[1]Accommodation is used in this paper to mean the mechanism to achieve system resilience under adversarial attack.

or closed loop system), set of output of controller nodes ('**c** '; controller node signals), set of reference nodes ('**r** '; reference signals). In this context, logical data are also considered as time series signals. Total number of signals used for providing immunity to a system are equal to total number of nodes that exist in sensors, actuators, controllers and reference nodes. Some or all combinations of these nodes can be used for monitoring and neutralization. The neutralizer comprises of many computational blocks (Feature Transform, Boundary and Performance Constrained Resilient Estimator and Inverse Feature Transform) as shown schematically in Figure 1. A multi-dimensional decision boundary and various mapping parameters are input to the neutralizer. Each block is described in detail next. The attack neutralizer is designed to perform like a filter to remove attack signatures present in each monitoring nodes. At the end, the output contains true estimates of signals at these nodes for use with the rest of the system for operational normalcy.

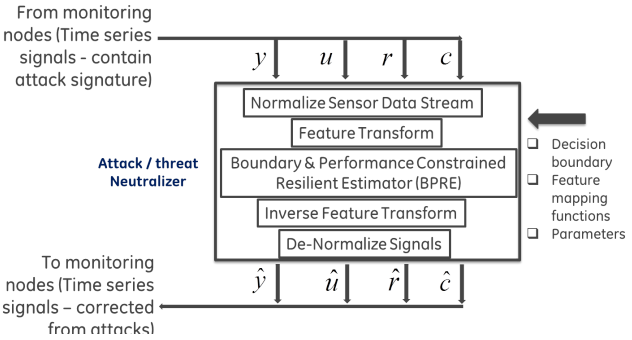

Fig. 1. Schematic diagram of the proposed industrial immune system

B. Normalize & De-Normalizing Sensor Data Stream

Often it will be useful to normalize sensor signals to some nominal operating condition temporally and spatially before performing feature transform. Consider, for example, temporal normalization of monitoring node data is performed as follows at every sample along a time axis:

$$\mathbf{y}_{\mathrm{N}} = \frac{(\mathbf{y}_{\mathrm{nom}} - \mathbf{y}_{\mathrm{UN}})}{\bar{\mathbf{y}}_{\mathrm{nom}}} \tag{1}$$

Where $\mathbf{y}_{\mathrm{N}}$ is the normalized signal, $\mathbf{y}_{\mathrm{UN}}$ is the un-normalized signal and $\mathbf{y}_{\mathrm{nom}}$ is the sensor signal for nominal operating condition and $\bar{\mathbf{y}}_{\mathrm{nom}}$ is a temporal average of a time series signal for normal operating conditions. In this case, to obtain feature transform a temporal average is computed for a batch length (e.g., 45 seconds).

De-normalization involves obtaining the sensor signals in original un-normalized space by inverting the equation above.

C. Forward and Inverse Feature Transform

Feature transform involves transforming signals from a Hilbert space ($\mathbb{H}$) to a finite dimensional Euclidean space. The components in the Euclidean space are referred to features in this paper.[2] In order to facilitate easy inversion, this transformation is done via orthogonal functional basis set. For the signal $y \in \mathbb{H}$, the feature, $\mathbf{w} \in \mathbb{R}^n$, is given by:

$$\mathbf{w}(t) = \underset{\mathbf{x}\in\mathbb{R}^n}{\arg\min} \int_{t-T}^{t} \left(y(\tau) - \sum_{j=1}^{n} \mathbf{x}_j \Psi_j(\eta_t(\tau)) \right)^2 d\tau \tag{2}$$

where $\eta_t : [t-T,\ t] \mapsto [-1,\ 1]$ and $\Psi_j \in \mathbb{H}^{[-1,\ 1]}, j = 1, \ldots, n$ are orthogonal basis functions. Hence, the solution to the above optimization problem is given by:

$$\mathbf{w}_j = \frac{< \Psi_j, y >}{< \Psi_j, \Psi_j >} \tag{3}$$

The inverse feature transform is then given by

$$\hat{y}(t) = \sum_{j=1}^{n} \mathbf{w}_j(t) \Psi_j(\eta_t(t)). \tag{4}$$

In general, there are many available choices for the orthogonal basis; from trigonometric, to orthogonal polynomials like Legendre, Laguerre, Hermite and Chebyshev. However, the signals considered in this paper are sampled at specific intervals and for each signal, there is a design of experiment done to collect a finite data set consisting of points that characterizes the operating space of that signal. Consequently, PCA is used to obtain an orthonormal set which serves as the functional basis for this application. The feature transform for the ith monitoring nodes is then obtained as:

$$\mathbf{y}_{\mathrm{N}_i} = \mathbf{y}_{0_i} + \mathbf{\Psi}_i \mathbf{w}_i \tag{5}$$

where $\mathbf{y}_{0_i} \in \mathbb{R}^{n_T}$ is the ensemble average for the signal, $\mathbf{\Psi}_i \in \mathbb{R}^{n_T \times n}$ is a matrix whose columns are the basis vectors, $\mathbf{w}_i \in \mathbb{R}^{n_T}$ is the corresponding feature, and n_T is the number of sample points in the interval $[t-T,\ t]$. It is noteworthy that the features are obtained from a sliding window batch of data as elucidated in (2). Hence, the features themselves are time series signals.

D. Boundary & Performance Constrained Resilient Estimator

We start by generating a labeled data set consisting of attack and normal scenarios. The data set have been carefully generated from realistic runs on high fidelity industry-standard models. The different normal and attack scenarios considered were carefully selected by the system domain experts.

First, a decision boundary that separates attack from safe is constructed using the kernel methods described in Section III-B. The Resilient Estimator combines concepts from compressed sensing with the binary classifier decision function to estimate the "true" features of the time series signals. The objective is to find the features that best explains the measurement and are also classified as "normal" by

[2]The notion of features used here is different from those used in Kernel-based learning methods. There, the transformation is usually from a finite dimensional Euclidean space to a higher dimensional reproducing kernel Hilbert space (RKHS). Features in that sense then refers to components in the RKHS.

the decision function. Normal features would represent safe operation with reasonable operational performance. As a result, the approach successfully fuses domain knowledge with compressed sensing. Another merit of this approach is that the domain knowledge, in the form of the decision function, can be updated online whenever new knowledge becomes available.

1) Boundary Decision Function: First, feature transform is done independently for each monitoring node. These nodal features are referred to as *local features*. The local features from all monitoring nodes are then stacked into one big vector, on which further dimensionality reduction is carried out to obtain what is referred to as *global features*. The local features capture time-related signatures from the signals from each monitoring nodes while the global features capture cross-relational signatures among monitoring nodes. The two-level feature transform is described by the following:

$$\mathbf{y}_{\mathrm{N}} = \mathbf{y}_0 + \mathbf{\Psi}\mathbf{w};\ \ \mathbf{w} = \mathbf{w}_0 + \mathbf{\Phi}\mathbf{g}$$

where $\mathbf{y}_{\mathrm{N}} = \left[\mathbf{y}_{\mathrm{N}_1}^\top, \ldots, \mathbf{y}_{\mathrm{N}_{n_s}}^\top\right]^\top$, $\mathbf{w} = \left[\mathbf{w}_1^\top, \ldots, \mathbf{w}_{n_s}^\top\right]$, $\mathbf{\Psi} = \mathsf{blkdiag}\left\{\mathbf{\Psi}_1, \ldots, \mathbf{\Psi}_{n_s}\right\}$, with n_s the number of monitoring nodes, $\mathbf{w}_i \in \mathbb{R}^{n_L}$ the ith local feature, $\mathbf{\Psi}_i \in \mathbb{R}^{n_T \times n_L}$ the ith local bases matrix, $\mathbf{g} \in \mathbb{R}^{n_G}$ the global features vector, $\mathbf{\Phi} \in \mathbb{R}^{n_T \times n_G}$ the global bases matrix., and $n_L \& n_G$ are the numbers of local and global features respectively. The 0th terms are ensemble averages of corresponding quantities. After combining the equations above, the conversion from global feature space to signal space is given by:

$$\mathbf{y}_{\mathrm{N}} = \underbrace{\mathbf{y}_0 + \mathbf{\Psi}\mathbf{w}_0}_{\mathbf{y}_b} + \underbrace{\mathbf{\Psi}\mathbf{\Phi}}_{\mathbf{\Omega}}\mathbf{g}. \tag{6}$$

The parameters $\mathbf{y}_0$, $\mathbf{w}_0$, $\mathbf{\Psi}$ and $\mathbf{\Phi}$ are obtained offline and can be updated independently of the resilient estimator.

Next, using the attack/safe labels of the data set together with global features, a kernel-based binary classifier is constructed. The classification decision function (or score) is of the form [16]

$$s(\mathbf{g}) = \sum_{i=1}^{m} q_i \kappa(\mathbf{x}_i, \mathbf{g}) + b_s \tag{7}$$

where $\kappa(.,.)$ is the kernel function which computes dot products in the high dimensional RKHS, $\mathbf{x}_i$ are the support vectors, q_i and b_s are the associated weights and bias. The decision function based on the score above is

$$\text{Status}(\mathbf{g}) = \begin{cases} \text{attack} & \text{if} \quad s(\mathbf{g}) > 0 \\ \text{normal} & \text{if} \quad s(\mathbf{g}) < 0 \end{cases} \tag{8}$$

2) Resilient Estimator: Suppose that the measured signal is given by:

$$\mathbf{y}_{\mathrm{N}} = \mathbf{y}_{\mathrm{N}}^* + \mathbf{y}_a, \tag{9}$$

where $\mathbf{y}_a$ is the malicious addition due to an adversarial agent, the objective is to estimate the true signal $\mathbf{y}_{\mathrm{N}}^*$. The attack signal $\mathbf{y}_a$ is not limited to a constant bias and could be any dynamically changing and intelligently designed malicious signal. It is assumed that the attack is sparse. In other words, the ratio of compromised nodes is such that there is enough redundancies in the system to facilitate reconstruction. The exact level of redundancy needed to ensure accurate reconstruction of compromised signal is a topic of future work. Sparse attack assumption is common in the literature on resilient estimation [11], [17], [18]. Equivalently, (9) can be written in terms of local features as:

$$\mathbf{w} = \mathbf{w}^* + \mathbf{w}_a. \tag{10}$$

Consequently, the resilient estimator is cast as the optimization problem

$$\text{Minimize} \left\|\mathbf{w}_k - \mathbf{w}_0 - \mathbf{\Phi}\mathbf{g}\right\|_{\ell_0};\ \text{s.t. } s(\mathbf{g}) < 0 \tag{11}$$

The subscript k is used to indicate the time dependence of the local features. The optimization problem above is NP-hard mainly due to the index minimization objective. This type of objective has been shown to be able to identify attack signals [11]. While, the constraint $s(\mathbf{g}) < 0$ improves the overall estimation problem by appending domain-level knowledge, it introduces additional complexity because it is in general non-convex. In what follows, a new problem composed of a convex relaxation of the objective and a local approximation of the kernel decision function is described.

Most approaches in literature [11], [17], [18] suggests replacing the index minimization objective with a ℓ_1-norm. It was shown in [18] that if the matrix $\mathbf{\Phi}$ satisfies certain conditions, this convexification is loss-less. Thus, the objective in (11) is replaced with $\left\|\mathbf{w}_k - \mathbf{w}_0 - \mathbf{\Phi}\mathbf{g}\right\|_{\ell_1}$.

Next, the local approximation of the kernel-decision function is described. The normal constraint $s(\mathbf{g}) < 0$ is equivalent to:

$$-\sum_{q_i<0} |q_i| \kappa(\mathbf{x}_i, \mathbf{g}) < -\sum_{q_j>0} |q_j| \kappa(\mathbf{x}_j, \mathbf{g}). \tag{12}$$

Since $\mathrm{sgn}(q_i)$ is really the label on the ith support vector, the constraint separates the support vectors into two sets and requires that the solution be closer to the "normal" set, where closeness is measured by the dot product in RKHS given by the kernel function. Subsequently, it is assumed that the kernel function is *isotropic*[3] i.e $\kappa(\mathbf{x}_i, \mathbf{g}) = f_\kappa(\|\mathbf{x}_i - \mathbf{g}\|_Q^2)$, $Q \succeq 0$. The function $f_\kappa : \mathbb{R}_+ \mapsto \mathbb{R}_+$ is continuous and monotonically decreasing. It is clear at this point that the non-convexity is due to the RHS of (12). Hence, the following approximation is considered:

$$\hat{s}(\mathbf{g}) \triangleq \sum_{q_i<0} \|\mathbf{x}_i - \mathbf{g}\|_{Q_i}^2 + \sum_{q_j>0} \mathbf{q}_j^\top (\mathbf{x}_j - \mathbf{g}), \tag{13}$$

where Q_i and $\mathbf{q}_i$ are parameters that are determined locally at each time step. The convex LHS is approximated with a quadratic and the non-convex RHS is approximated with linear function. A local binary classification problem is then solved with two requirements:

- preserve the label of the current measurement

[3]With slight modification, the result in this paper can be reproduced for other types of kernel

- reproduce the labels of the support vectors as much as possible with more emphasis placed on the ones closest to the current operating point.

Therefore, the local classification problem considered is:

$$\begin{aligned}&\text{Minimize:}\\&\quad \tfrac{1}{2}\left(\sum_{q_i<0}\|Q_i\|_{\mathrm{F}}^2+\sum_{q_j>0}\|\mathbf{q}_j\|^2+\gamma\sum_{i=1}^{m}d(\mathbf{x}_i,\mathbf{g}_{k-1})e_i^2\right)\\&\text{Subject to:}\\&\quad s_k\hat{s}(\mathbf{g}_k)\geq 1\\&\quad s_i\hat{s}(\mathbf{x}_j)-1+e_i=0,\ i=1,\ldots,m,\end{aligned}\tag{14}$$

where

- $\mathbf{g}_{k-1}$ is estimated global feature from the previous step. It is used as an indication of the current operating condition,
- $d(\mathbf{x}_i,\mathbf{g}_{k-1})$ is a semi-positive valued function used to penalize the importance of the ith support vector. e.g $10\exp(-\|\mathbf{x}_i-\mathbf{g}_{k-1}\|^2)+1$.
- $(\mathbf{g}_k, s_k)$ are the respective global features and SVM score corresponding to the current measurement
- (s_i, e_i) are the respective SVM score and degree of misclassification of the ith support vector

The first two terms in the objective corresponds to Vapnik-Chervonenkis (VC) dimension [20] which helps to improve the generalization of the decision function. For simplicity and faster computation, the Q_is are restricted to a class of diagonal matrices ($Q_i = \text{diag}(\mathbf{q}_i)$). As a result, the optimization problem in (14) reduces to:

$$\begin{aligned}&\text{Minimize:}\\&\quad \tfrac{1}{2}\left(\|\mathbf{q}\|^2+\gamma\|\mathbf{e}\|_D^2\right)\\&\text{Subject to:}\\&\quad 1-s_kK_k\mathbf{q}\leq 0\\&\quad SK\mathbf{q}-\underline{\mathbf{1}}+\mathbf{e}=0\end{aligned}\tag{15}$$

where $\mathbf{q}=[\mathbf{q}_1^\top,\ldots,\mathbf{q}_m^\top]^\top$, $\mathbf{e}=[e_1,\ldots,e_m]^\top$,
$D=\text{diag}([d(\mathbf{x}_1,\mathbf{g}_{k-1}),\ldots,d(\mathbf{x}_m,\mathbf{g}_{k-1})])$,
$K_k=[(\mathbf{x}_1-\mathbf{g}_k)^\top\text{diag}(\mathbf{x}_1-\mathbf{g}_k),\ldots,(\mathbf{x}_m-\mathbf{g}_k)^\top\text{diag}(\mathbf{x}_m-\mathbf{g}_k)]$,
$S=\text{diag}([s_1,\ldots,s_m])$, and

$$K=\begin{bmatrix}\underline{\mathbf{0}}^\top & \cdots & (\mathbf{x}_1-\mathbf{x}_m)^\top\text{diag}(\mathbf{x}_1-\mathbf{x}_m)\\ \vdots & \ddots & \vdots\\ (\mathbf{x}_m-\mathbf{x}_1)^\top\text{diag}(\mathbf{x}_m-\mathbf{x}_1) & \cdots & \underline{\mathbf{0}}^\top\end{bmatrix}$$

This is a simple linearly-constrained quadratic program and it's dual;

$$\underset{\alpha\geq 0}{\text{Minimize}}\ \frac{1}{2}a\alpha^2-b\alpha,\tag{16}$$

where

$$\begin{aligned}H&=I+\gamma K^\top DK\\a&=K_k^\top H^{-1}K_k\\b&=1-\gamma s_kK_k^\top H^{-1}K^\top SD\underline{\mathbf{1}},\end{aligned}$$

has the closed-form solution $\alpha^*=\max\left\{\frac{b}{a},0\right\}$. Hence the parameters of $\hat{s}(.)$ as given by the solution of the primal problem (15) are:

$$\mathbf{q}^*=H^{-1}\left(\gamma K^\top SD\underline{\mathbf{1}}+\alpha^*s_kK_k\right).\tag{17}$$

Remark. *In order to avoid computing the inverse of a possibly large matrix at every time step, a rule-of-thumb is suggested for the weight matrix D. Let $K=U\Sigma V^\top$, where $\Sigma=[\Sigma_1\ \boldsymbol{O}]$, be the singular value decomposition of K– which can be calculated apriori. Select $D=U\Lambda U^\top$, where $\Lambda=diag([d(\mathbf{x}_1,\mathbf{g}_{k-1}),\ldots,d(\mathbf{x}_m,\mathbf{g}_{k-1})])$. Thus, the inverse*

$$H^{-1}=V\left[\begin{array}{c|c}(I+\gamma\Sigma_1\Lambda\Sigma_1)^{-1} & \\ \hline & I\end{array}\right]V^\top\tag{18}$$

only involves online computation of the inverse of a smaller diagonal matrix.

The relaxed and local approximate optimization problem for the resilient problem is:

$$\begin{aligned}&\text{Minimize}\quad \|\mathbf{w}_k-\mathbf{w}_0-\mathbf{\Phi g}\|_{\ell_1}\\&\text{Subject to}\\&\qquad \mathbf{g}^\top P\mathbf{g}+\mathbf{p}^\top\mathbf{g}+r<0,\end{aligned}\tag{19}$$

where

$$\begin{aligned}P&=\sum_{q_i<0}Q_i\\\mathbf{p}&=-2\sum_{q_i<0}Q_i\mathbf{x}_i-\sum_{q_j>0}\mathbf{q}_j\\r&=\sum_{q_i<0}\|\mathbf{x}_i\|_{Q_i}^2+\sum_{q_j>0}\mathbf{x}_i^\top\mathbf{q}_j.\end{aligned}\tag{20}$$

At this point, the matrix P can in general be indefinite – since there is nothing saying otherwise. One could explicitly impose that constraint in the optimization problem in (15), in which case the ability to derive a closed-form solution is lost and one resorts to iterative solution for the local approximation problem. This would be a good option if there is enough processing speed to handle iterative solutions in both the local approximation and resilient estimation problems. On the other hand, one could allow indefinite P in favor of a closed-form solution for the local approximation problem – as done already. If that is the case, the optimization problem in (19) can be further relaxed by considering a semi-definite program (SDP) variant [21].The resulting SDP can then be solve sufficiently fast using interior-point methods [22].

V. NUMERICAL SIMULATION

Figure 2 illustrates an example of the simulated performance of a boundary and performance constrained resilient estimator as described above. In this example, 45 seconds of a time series signal is shown for six sensor nodes (sensors 1 to 6) obtained from GE ARTEMIS high fidelity power plant simulation platform.

In this example, signals with large deviations (red signal) corresponds to the behavior of the system under a spoofing attack (i.e., when sensor 6 was subjected to a bias attack). The middle (green) signal of each graph corresponds to the normal operational condition prior to the attack while the

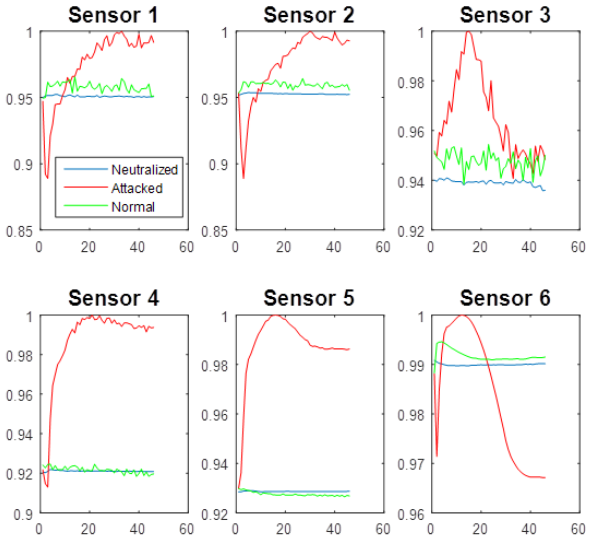

Fig. 2. Numerical simulation of neutralization with BPRE for a simulated spoofing attack to sensor 6. The x-axis in the figure is time in seconds and the y-axis are the corresponding normalized sensor values.

bottom (blue) signal of each graph corresponds to neutralized signals obtained after processing through the resilient estimator. In this example, features were computed using local and global basis vectors obtained from linear principal component analysis on the data set obtained with several normal and attack scenarios. This simulation used 15 support vectors.

VI. CONCLUSIONS

This paper shows a framework for using advanced techniques developed in machine learning with control and optimization theory to create an industrial equivalent of a human immune system. When pathogens enter our bodies, the body's immune system automatically detects the attack and triggers a response to fight off the pathogens over time. In an industrial equivalent system, even though we cannot remove the pathogens automatically, as of now, at the minimum their abnormal behavior on the system can be neutralized with algorithms discussed in this paper provided there is enough system level redundancies in the sensing system. In our method, we can use heterogeneous sensing modality and feature-based discovery & learning to transform the signals to feature space, detect abnormal features in the feature-space using kernel-based decision boundary constraints and then estimate their true values close to the values the system was producing during the operational normalcy just before the attack took place. True values are estimated in feature-space and then transformed to time domain. Use of feature space reduces the dimensionality of the system.

ACKNOWLEDGMENT

This study is supported under contract number DEOE0000833 awarded in 2016 by the United States Department of Energy (DOE)'s Cyber-security for Energy Delivery Systems (CEDS) R&D Program. Authors acknowledge the DOE/CEDS R&D Staff, and Justin John and Daniel Holzhauer from GE Global Research for useful discussions and providing simulation infrastructure with high fidelity models.

REFERENCES

[1] R. Langner, Stuxnet: Dissecting a cyberwarfare weapon, IEEE Security Privacy, vol. 9, no. 3, pp. 49-51, 2011.

[2] R. Loughin, Cyber Attack on Illinois Water Utility System, http://www.chemicalprocessing.com/blogs/chemical-security-action/cyber-attack-on-illinois-water-utility-system/, 2011.

[3] R.M. Lee, M.J. Assante, T. Conway, German Steel Mill Cyber Attack, ICS CP/PE (Cyber-to-Physical or Process Effects) case study paper, SANS, ICS Defense Use Case Dec 30, 2014; same authors working with the E-ISAC, Analysis of the Cyber Attack on the Ukrainian Power Grid, SANS, ICS Defense Use Case March 18, 2016.

[4] R. Spenneberg, M. Brggemann, H. Schwartke, PLC-Blaster: A Worm Living Solely in the PLC, Black Hat Asia 2016

[5] D. Dasgupta, Advances in Artificial Immune Systems, IEEE Computational Intelligence Magazine, Nov 2006.

[6] P.J.C Branco, J.A. Dente, R.V. Mendes, Using Immunology Principles for Fault Detection, IEEE Trans. On Industrial Electronics, Vol. 50, No. 2, April 2003.

[7] M. Jha, R. Acharya, An Immune inspired Unsupervised Intrusion Detection System for Detection of Novel Attacks, Security Informatics (ISI), 2016

[8] A. Giani, E. Bitar, M. Garcia, M. McQueen, P. Khargonekar, K. Poolla, Smart Grid Data Integrity Attacks, in Smart Grid, IEEE Transactions on, vol.4, no.3, pp.1244-1253, Sept. 2013.

[9] C-F. Tsai et al., Intrusion detection by machine learning: A review, Expert Systems with Applications 36.10, 11994-12000, 2009.

[10] H. Fawzi, P. Tabuada, S. Diggavi, Secure Estimation and Control for Cyber-Physical Systems Under Adversarial Attacks, IEEE Trans. On Automatic Control, Vol. 59, No. 6, June 2014.

[11] F. Pasqualetti, F. Dorfler, and F. Bullo, Attack detection and identification in cyber-physical systems, IEEE Trans. Autom. Control, vol. 58, no. 11, pp. 2715-2729, 2013.

[12] J. Kim, C. Lee, H. Shim, Y. Eun, J.H. Seo, Detection of Sensor Attack and Resilient State Estimation for Uniformly Observable Nonlinear Systems, 55th IEEE Conf on Decision and Control, Dec 12-14, Las Vegas, 2016

[13] M. Ozay et al., Machine Learning Methods for Attack Detection in the Smart Grid, IEEE Trans. on Neural Networks and Learning Systems, Mar 2015.

[14] W. Yan, L. Yu, On Accurate and Reliable Anomaly Detection for Gas Turbine Combustors: A Deep Learning Approach, Annual Conference of the Prognostics and Health Management Society, 2015.

[15] DOE Roadmap to Achieve Energy Delivery Systems Cybersecurity, www.controlsystemsroadmap.net, Sept 2011.

[16] Scholkopf, Bernhard, and Alexander J. Smola. Learning with kernels: support vector machines, regularization, optimization, and beyond. MIT press, 2001.

[17] Fawzi, Hamza, Paulo Tabuada, and Suhas Diggavi. Secure estimation and control for cyber-physical systems under adversarial attacks. IEEE Transactions on Automatic Control 59.6 (2014): 1454-1467.

[18] Candes, Emmanuel J., and Terence Tao. Decoding by linear programming. IEEE transactions on information theory 51.12 (2005): 4203-4215.

[19] Ahsen, Mehmet Eren, Niharika Challapalli, and Mathukumalli Vidyasagar. Two New Approaches to Compressed Sensing Exhibiting Both Robust Sparse Recovery and the Grouping Effect. arXiv preprint arXiv:1410.8229 (2014).

[20] Blumer, Anselm, et al. Learnability and the Vapnik-Chervonenkis dimension. Journal of the ACM (JACM) 36.4 (1989): 929-965.

[21] Boyd, Stephen, and Lieven Vandenberghe. Semidefinite programming relaxations of non-convex problems in control and combinatorial optimization. Communications, Computation, Control, and Signal Processing. Springer US, 1997. 279-287.

[22] Kim, Seung-Jean, et al. An Interior-Point Method for Large-Scale ℓ_1-Regularized Least Squares. IEEE journal of selected topics in signal processing 1.4 (2007): 606-617.