Available online at www.sciencedirect.com

ScienceDirect

IFAC PapersOnLine 50-1 (2017) 9496–9501

Optimal Stealthy Attack under KL Divergence and Countermeasure with Randomized Threshold

Enoch Kung * **Subhrakanti Dey** ** **Ling Shi** ***

* *Electronic and Computer Engineering, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong. Email: ekung@ust.hk.*
** *Signals and Systems at Department of Engineering Sciences, Uppsala University, Uppsala, Sweden. Email: Subhrakanti.Dey@signal.uu.se.*
*** *Electronic and Computer Engineering, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong. Email: eesling@ust.hk.*

Abstract: In a cyber-physical system, there are potential sources of malicious attacks that can damage the estimation quality in an underlying network control system. The attacker aims to maximize these damages while the estimator attempts to minimize them. In this paper we define an attack's stealth based on the KL divergence and obtain an optimal attack. Furthermore, we suggest one method in which the estimator may limit the damage to the system while imposing on any attack a probability for it to be non-stealthy.

Keywords: Cyber-Physical Systems, Detection, Security

1. INTRODUCTION

The integral role played by cyber-physical systems (CPS) in the modern era cannot be underestimated. Its scope of applications in communication, transportation, utilities, etc., has motivated extensive research in the area of remote estimation over communication networks. The estimation quality is crucial to the estimator and is therefore an ideal target for potential cyber or physical attacks.

The remote, wireless aspect of the system exposes CPS to external malicious attackers. The study of CPS security is a never ending war between the system designers and attackers. The security problems such as the case of the Maroochy Water Breach [Slay,Miller (2007)] and the SQL Slammer worm attack on the nuclear plant [Kuvshinkova (2003)] are testaments to the damage an attack on CPS can inflict. Therefore, one must study the devastating effects of attacks and at the same time develop countermeasures.

A control system comprises of a plant, sensors, an estimator, and actuators existing in constant communication. An attacker targets this communication in a variety of forms depending on its purpose. For example, a denial-of-service attack [Amin et al. (2009)] interrupts the communication between sensors and the estimator and simply prevents a packet of information from being successfully transmitted. An attacker may also replace the transmitted packet with malicious information [Liu et al. (2009)] further leading the system astray.

Estimators may attempt to detect an attack by setting a detection policy based on incoming data. It performs a hypothesis test to decide whether the data is corrupted or not. For instance, with a false data injection attack with multiple sensors, a sensors data can be cross-checked with those of its neighboring sensors [Ye et al. (2004)] [Shukla and Qiao (2007)]. Under a detection policy, the attacker must trade the magnitude of damage with stealth.

One metric used to define stealthiness in many papers involves the KL divergence. In the numerous ways the attacker can corrupt the system communication, the only requirement is that the corrupted measurement should not differ from the correct measurement by too much in terms of the KL divergence. In [Bai et al. (2015)] and [Kung et al. (2016)], the attack is implemented on the control.

Our work will explore a vector system where the attacker corrupts the measurement transmitted from the sensors. The tradeoff between stealthiness and estimation quality can be clearly described. On the other hand, the attacker's desire to remain undetected is taken advantage of by the estimator. We propose a method in which an attack of any power will have a nonzero probability of being non-stealthy. Our two main contributions are summarized as follows:

1. The estimation error covariance under a stealthy attack is shown to be bounded above and an optimal Gaussian attack achieves this bound.
2. We study the use of randomizing the threshold to further limit the estimation error and increase the probability of discovering the attacker. As far as we

[1] The work by E. Kung and L. Shi is supported by an HKUST Caltech Partnership FP009.

2405-8963 © 2017, IFAC (International Federation of Automatic Control) Hosting by Elsevier Ltd. All rights reserved.
Peer review under responsibility of International Federation of Automatic Control.
10.1016/j.ifacol.2017.08.1587

know, this is the first proposed countermeasure to such attacks.

The paper is organized as follows. In Section 2, the problem background and set up are introduced. Section 3 presents the main contributions of this paper. We provide simulations and comparisons to previous work in Section 4. Section 5 concludes the paper with a few comments on future directions. Notations: $M^{m\times n}$ denotes the space of $m \times n$ matrices. $\mathbb{S}_m^+$ denotes the set of $m \times m$ positive definite matrices. A Gaussian variable z with mean μ and covariance P is written as $z \sim \mathcal{N}(\mu, P)$. For a matrix A, $A^\intercal$ denotes its transpose. For a positive definite matrix X, $X^{1/2}$ is the positive definite square root of X. For a row or column vector v, $(v)_j$ denotes the $j-$th entry of v.

2. PRELIMINARIES

2.1 System Model

We first set up the scenario in which the problem is considered. The state and output variables will follow the equations

$$\begin{aligned} x_{k+1} =& Ax_k + w_k \\ y_k =& Cx_k + v_k; \end{aligned}$$

$A \in M^{n\times n}$, $C \in M^{m\times n}$, $w_k \sim \mathcal{N}(0, Q)$, $v_k \sim \mathcal{N}(0, R)$, where $Q \in \mathbb{S}_n^+$ and $R \in \mathbb{S}_m^+$.

As is well known, the estimate of the state variable comes in the form of update equations of the conditional state mean and covariance. To elaborate, let $\mathcal{I}_k = \{y_1, \dots, y_k\}$ be the history of received signals and define

$$\begin{aligned} \hat{x}_k = \mathbb{E}[x_k \mid \mathcal{I}_k],\ & \hat{x}_{k|k-1} = \mathbb{E}[x_k \mid \mathcal{I}_{k-1}] \\ P_k = & \mathbb{E}[(x_k - \hat{x}_k)(x_k - \hat{x}_k)^\intercal \mid \mathcal{I}_k] \\ P_{k|k-1} = & \mathbb{E}[(x_k - \hat{x}_{k|k-1})(x_k - \hat{x}_{k|k-1})^\intercal \mid \mathcal{I}_{k-1}]. \end{aligned}$$

The Kalman filter gives us the optimal update equations

$$\begin{aligned} \hat{x}_{k|k-1} &= A\hat{x}_{k-1} \\ \hat{x}_k &= \hat{x}_{k|k-1} + K_k(y_k - C\hat{x}_{k|k-1}) \\ P_{k|k-1} &= AP_{k-1}A^\intercal + Q \\ P_k &= (I - K_kC)P_{k|k-1} \\ K_k &= P_{k|k-1}C^\intercal(CP_{k|k-1}C^\intercal + R)^{-1}. \end{aligned}$$

It is also proven that if (A, C) is observable and $(A, \sqrt{Q})$ is controllable, then the sequence $\{K_k\}$ and $\{P_{k|k-1}\}$ converges to a steady state K and P. The steady state error covariance P is the solution of $h \circ g(X) = X$ where

$$\begin{aligned} g(X) &= X - XC^\intercal(CXC^\intercal + R)^{-1}CX \\ h(X) &= AXA^\intercal + Q. \end{aligned}$$

The steady state Kalman gain K is calculated by $PC^\intercal(CPC^\intercal+R)^{-1}$. The covariance of the output y_k would then be written as $\Sigma = CPC^\intercal + R$. We will assume that the system is already in steady state.

2.2 Attack

In our set up, we assume that the sensor transmits y_k to the estimator and that the attacker intercepts y_k and replaces it with a corrupted signal $\tilde{y}_k$. The corrupted signal will affect the state estimation equation, given by

$$\hat{\tilde{x}}_k = A\hat{\tilde{x}}_{k-1} + K(\tilde{y}_k - CA\hat{\tilde{x}}_{k-1}) = (I - KC)A\hat{\tilde{x}}_{k-1} + K\tilde{y}_k.$$

The objective for the attacker is to design $\tilde{y}_k$ that maximizes the Frobenius norm of the difference between the accurate state and the corrupt estimate at each time step k, that is, the attacker aims to maximize

$$\textbf{tr}\ \mathbb{E}[(x_k - \hat{\tilde{x}}_k)(x_k - \hat{\tilde{x}}_k)^\intercal \mid \tilde{\mathcal{I}}_k].$$

Here the history $\mathcal{I}_k$ is the correct measurements up to time k and the corrupted measurements up to time $k - 1$, i.e.,

$$\tilde{\mathcal{I}}_k = \{y_1, \dots, y_k\} \cup \{\tilde{y}_1, \dots, \tilde{y}_{k-1}\}.$$

2.3 Stealthiness

The estimator performs a hypothesis test based on the incoming corrupted signals to decide whether or not the system is under attack. An alarm will be sounded if the estimator decides it is under attack.

We follow the works of [Bai et al. (2015)], [Kung et al. (2016)] and consider the KL divergence between the two distributions P_0 and P_1.

Definition 1. Let P_0 and P_1 be two distributions. The Kullback-Leibler (KL) Divergence between them is given by the expression

$$D(P_0||P_1) = \int_{-\infty}^{\infty} P_0(z)\log\frac{P_0(z)}{P_1(z)}dz.$$

This value denotes the "difference" between the two probability distributions.

Here the attacker wishes to generate a vector $\tilde{y}_k$ following P_0 given the history $\tilde{\mathcal{I}}_k$ such that its difference between P_0 and the predicted distribution P_1 of the uncorrupted y_k based on $\{y_1, \dots, y_{k-1}\}$ is bounded, say by a threshold ϵ. This gives rise to the following definition of stealthiness.

Definition 2. Given $\tilde{\mathcal{I}}_k = \{y_1, \dots, y_k\} \cup \{\tilde{y}_1, \dots, \tilde{y}_{k-1}\}$, an attacker $\{y_i\}_1^k \longrightarrow \tilde{y}_k$ is stealthy if $\tilde{y}_k \sim P_0$, $P_1 = \mathcal{N}(C\hat{x}_{k|k-1}, \Sigma)$ is the conditional probability on $\{y_1, \dots, y_{k-1}\}$, and $D(P_0||P_1) \leq \epsilon$.

To summarize, the attacker's objective is to solve at each time k

$$\max_{\tilde{y}_k} \textbf{tr}\ \mathbb{E}[(x_k - \hat{\tilde{x}}_k)(x_k - \hat{\tilde{x}}_k)^\intercal \mid \tilde{\mathcal{I}}_k] \text{ with } D(P_0||P_1) \leq \epsilon.$$

Taking $P_1 = \mathcal{N}(C\hat{x}_{k|k-1}, \Sigma)$, $D(P_0||P_1)$ is given by

$$\begin{aligned} & D(P_0||P_1) \\ =& \frac{1}{2}\log((2\pi)^m|\Sigma|) \\ & + \frac{1}{2}\textbf{tr}\ \Sigma^{-1}\mathbb{E}[(\tilde{y}_k - C\hat{x}_{k|k-1})(\tilde{y}_k - C\hat{x}_{k|k-1})^\intercal \mid \tilde{\mathcal{I}}_k] \\ & - \Big(-\int_{-\infty}^{\infty} P_0(z)\log P_0(z)dz\Big). \end{aligned}$$

The final term is known as the differential entropy and it satisfies the inequality

$$-\int_{-\infty}^{\infty} P_0(z)\log P_0(z)dz \leq \frac{1}{2}\log\Big((2\pi e)^m|\tilde{\Sigma}_k|\Big)$$

with equality when P_0 is a Gaussian distribution.

The constraint $\epsilon \geq D(P_0||P_1)$ implies

$$\begin{aligned} \epsilon \geq & \frac{1}{2}\textbf{tr}\ \Sigma^{-1}\mathbb{E}[(\tilde{y}_k - C\hat{x}_{k|k-1})(\tilde{y}_k - C\hat{x}_{k|k-1})^\intercal \mid \tilde{\mathcal{I}}_k] \\ & - \frac{1}{2}\log\left(|\Sigma^{-1}\tilde{\Sigma}_k|\right) - \frac{m}{2}. \end{aligned} \quad (1)$$

All stealthy attacks will also satisfy this inequality, so for now, we consider all attacks that satisfy (1). We will perform optimization over all attacks in this enlarged set and show that it is stealthy.

Be mindful that the expectation in (1) is the expectation over P_0.

Remark 1: Stealthiness can also be defined by setting P_1 to be the predicted distribution of the measurement using corrupted state estimate, i.e., $P_1 \sim \mathcal{N}(C\hat{\tilde{x}}_{k|k-1}, \Sigma)$. The stealthy attacks under this definition strictly includes those defined by Definition 1.

3. MAIN RESULT

3.1 Optimal Attack

This section deals with the optimal attack on the measurement. The attacker intercepts the measurement and, based on its knowledge of past measurements and attacks, alters this measurement. We provide an explicit construction.

First we expand the error covariance term

$$
\begin{aligned}
&\mathbb{E}[(x_k - \hat{\tilde{x}}_k)(x_k - \hat{\tilde{x}}_k)^\intercal \mid \tilde{\mathcal{I}}_k] \\
=&\mathbb{E}[(x_k - \hat{x}_k + \hat{x}_k - \hat{\tilde{x}}_k)(x_k - \hat{x}_k + \hat{x}_k - \hat{\tilde{x}}_k)^\intercal \mid \tilde{\mathcal{I}}_k] \\
=&P + \mathbb{E}[(x_k - \hat{x}_k)](\hat{x}_k - \hat{\tilde{x}}_k)^\intercal] \\
&\quad + (\hat{x}_k - \hat{\tilde{x}}_k)\mathbb{E}[(x_k - \hat{x}_k)]^\intercal + (\hat{x}_k - \hat{\tilde{x}}_k)(\hat{x}_k - \hat{\tilde{x}}_k)^\intercal.
\end{aligned}
$$

Since $\mathbb{E}[x_k] = \hat{x}_k$, the term above becomes

$$
P + \mathbb{E}[(\hat{x}_k - \hat{\tilde{x}}_k)(\hat{x}_k - \hat{\tilde{x}}_k)^\intercal \mid \tilde{\mathcal{I}}_k].
$$

Defining $e_k = \hat{x}_k - \hat{\tilde{x}}_k$, we have

$$
e_k = (I - KC)Ae_{k-1} + K(y_k - \tilde{y}_k).
$$

After considerable algebra,

$$
\begin{aligned}
\mathbb{E}[e_k e_k^\intercal \mid \tilde{\mathcal{I}}_k] =&\blacksquare - K\mathbb{E}[\tilde{y}_k \mid \tilde{\mathcal{I}}_k]((I - KC)Ae_{k-1} + Ky_k)^\intercal \\
&- ((I - KC)Ae_{k-1} + Ky_k)\mathbb{E}[\tilde{y}_k \mid \tilde{\mathcal{I}}_k]^\intercal K^\intercal \\
&+ K\mathbb{E}[(\tilde{y}_k - \mathbb{E}[\tilde{y}_k])(\tilde{y}_k - \mathbb{E}[\tilde{y}_k])^\intercal \mid \tilde{\mathcal{I}}_k]K^\intercal \\
&+ K\mathbb{E}[\tilde{y}_k \mid \tilde{\mathcal{I}}_k]\mathbb{E}[\tilde{y}_k \mid \tilde{\mathcal{I}}_k]^\intercal K^\intercal,
\end{aligned}
$$

where $\blacksquare$ represents a group of terms determined by the history $\tilde{\mathcal{I}}_k$. For simplicity, let $(I - KC)Ae_{k-1} + Ky_k = \overline{e}_k$, which is also deterministic under the given history.

It is obvious that maximizing the estimation error is equivalent to maximizing $\mathbf{tr}\ \mathbb{E}[\, e_k e_k^\intercal \mid \tilde{\mathcal{I}}_k]$ and thus maximizing

$$
- 2\overline{e}_k^\intercal K\mu_k + \mathbf{tr}\ \left[K\tilde{\Sigma}_k K^\intercal\right] + \mathbf{tr}\ \left[K\mu_k\mu_k^\intercal K^\intercal\right] \text{ where}
$$

$$
\mu_k = \mathbb{E}[\tilde{y}_k \mid \tilde{\mathcal{I}}_k] \text{ and } \tilde{\Sigma}_k = \mathbb{E}[(\tilde{y}_k - \mu_k)(\tilde{y}_k - \mu_k)^\intercal \mid \tilde{\mathcal{I}}_k].
$$

The main result of the section is given in the following two theorems.

Theorem 3. Let Q be the orthogonal matrix that diagonalizes $\Sigma^{1/2}K^\intercal K\Sigma^{1/2}$, that is, $Q^\intercal\Sigma^{1/2}K^\intercal K\Sigma^{1/2}Q = \mathcal{K} = diag(k_1, \ldots, k_m)$. Then

$$
\begin{aligned}
&- 2\overline{e}_k^\intercal K\mu_k + \mathbf{tr}\ \left[K\tilde{\Sigma}_k K^\intercal\right] + \mathbf{tr}\ \left[K\mu_k\mu_k^\intercal K^\intercal\right] \\
\leq& - \sum_{j=1}^{m} (2\overline{e}_k^\intercal K\Sigma^{1/2}Q)_j s_{k,j}(\eta^*) + k_j\lambda_{k,j}(\eta^*) + k_j s_{k,j}(\eta^*)^2
\end{aligned}
$$

such that the following are satisfied:

$$
s_{k,j}(\eta^*) = \frac{-\left(\bar{e}_k^\intercal K\Sigma^{1/2}Q\right)_j}{\eta^* - k_j},\ \tilde{\lambda}_{k,j}(\eta^*) = \frac{1}{1 - \frac{k_j}{\eta^*}}
$$

$$
2\epsilon = \sum_{j=1}^{m} \tilde{\lambda}_{k,j}(\eta^*) + \sum_{j=1}^{m} s_{k,j}(\eta^*)^2 - m - \sum_{j=1}^{m} \log \tilde{\lambda}_{k,j}(\eta^*).
$$

Theorem 4. Let $s_k = \begin{bmatrix} s_{k,1} & \ldots & s_{k,m} \end{bmatrix}$ and let $(\tilde{\lambda}_{k,1}, \ldots, \tilde{\lambda}_{k,m})$ be the eigenvalues of the matrix $\tilde{\Lambda}_k$. The attack given by

$$
\tilde{y}_k \sim \mathcal{N}(\Sigma^{1/2}Qs_k + C\hat{x}_{k|k-1}, \Sigma^{1/2}Q\tilde{\Lambda}_k Q^\intercal\Sigma^{1/2})
$$

achieves the upper bound and is stealthy.

Proof. [Theorem 3] First we perform a change of variables to facilitate the calculations,

$$
\begin{aligned}
s_k =&Q^\intercal\Sigma^{-1/2}(\mu_k - C\hat{x}_{k|k-1}) \\
\tilde{\Lambda}_k =&Q^\intercal\Sigma^{-1/2}\tilde{\Sigma}_k\Sigma^{-1/2}Q.
\end{aligned} \tag{2}
$$

Once the optimal s_k and $\tilde{\Lambda}_k$ are determined, these equations give us the optimal mean μ_k and $\tilde{\Sigma}_k$. The following expression

$$
-2\overline{e}_k^\intercal K\mu_k + \mathbf{tr}\ \left[K\tilde{\Sigma}_k K^\intercal\right] + \mathbf{tr}\ \left[K\mu_k\mu_k^\intercal K^\intercal\right] \tag{3}
$$

is written, in new variables after eliminating all constant terms (which plays no part in the optimization), as

$$
\begin{aligned}
&\sum_{j=1}^{m} -2q_{k,j}s_{k,j} + \mathbf{tr}\ \mathcal{K}\tilde{\Lambda}_k + \mathbf{tr}\ \mathcal{K}s_k s_k^\intercal \\
\leq&\sum_{j=1}^{m} -2q_{k,j}s_{k,j} + k_j\lambda_{k,j} + k_j s_{k,j}^2.
\end{aligned} \tag{4}
$$

Here, $q_{k,j}$ is the $j-th$ entry of the vector $\overline{e}_k^\intercal K\Sigma^{1/2}Q$. The inequality is a simple result of matrix theory. Similarly the constraint on the KL divergence can be rewritten from (1) to

$$
\begin{aligned}
2\epsilon \geq&\mathbf{tr}\ \Sigma^{-1}\tilde{\Sigma}_k + \mathbf{tr}\ \Sigma^{-1}(\mu_k - C\hat{x}_{k|k-1})(\mu_k - C\hat{x}_{k|k-1})^\intercal \\
&- \log(|\Sigma^{-1}\tilde{\Sigma}_k|) - m \\
\geq&\sum_{j=1}^{m} \tilde{\lambda}_{k,j} + s_{k,j}^2 - 1 - \log \tilde{\lambda}_{k,j}
\end{aligned} \tag{5}
$$

From (4) we see that over all matrices $\tilde{\Lambda}_k$ with a given set of eigenvalues $(\tilde{\lambda}_{k,1}, \ldots, \tilde{\lambda}_{k,m})$, $\tilde{\Lambda}_k = diag(\tilde{\lambda}_{k,1}, \ldots, \tilde{\lambda}_{k,m})$ maximizes our objective function. Therefore, we narrow our consideration to diagonal $\tilde{\Lambda}_k$. Furthermore, both the right hand expression in (5) and the objective function (3) is increasing on at least one eigenvalue $\tilde{\lambda}_{k,j}$ (because if not, then $\tilde{\Sigma}_k \leq \Sigma$, which improves estimation quality). If any attack satisfies the strict inequality in (5), one can always increase that eigenvalue $\tilde{\lambda}_{k,j}$, consequently increasing (3) while continuing to satisfy (5).

The maximization problem is now reduced to

$$
\max_{s_k, \tilde{\Lambda}_k} \sum_{j=1}^{m} -2q_{k,j}s_{k,j} + k_j\tilde{\lambda}_{k,j} + k_j s_{k,j}^2 \text{ subject to}
$$

$$
2\epsilon = \sum_{j=1}^{m} \tilde{\lambda}_{k,j} + s_{k,j}^2 - 1 - \log \tilde{\lambda}_{k,j}.
$$

The equations obtained from Lagrange multipliers are

$$2s_{k,j}\eta = -2q_{k,j} + 2k_j s_{k,j},\ \ k_j = \eta\left(1 - \frac{1}{\tilde{\lambda}_{k,j}}\right).$$

Solving this yields

$$s_{k,j}(\eta) = \frac{-q_{k,j}}{\eta - k_j},\ \ \tilde{\lambda}_{k,j}(\eta) = \frac{1}{1 - \frac{k_j}{\eta}}.$$

To satisfy the constraint, we require the η^* to satisfy

$$\sum_{j=1}^{m} \tilde{\lambda}_{k,j}(\eta^*) + s_{k,j}(\eta^*)^2 - 1 - \log \tilde{\lambda}_{k,j}(\eta^*) = 2\epsilon \qquad (6)$$

and η^* can be calculated by root-finding algorithms such as Bisection Algorithm or Newton's method.

There are two solutions η possible, lying in the branches $(-\infty, 0) \cup (\max_j k_j, \infty)$ and we choose the η that yields the larger value for the objective function.

The theorem is thus proved.

Proof. [Theorem 4] The previous proof constructs s_k and $\tilde{\Lambda}_k$ and by (2) we obtain μ_k and $\tilde{\Sigma}_k$. Take $\tilde{y}_k \sim P_0 = \mathcal{N}(\mu_k, \tilde{\Sigma}_k)$. This achieves the upper bound. It remains to prove its stealthiness.

Because P_0 is Gaussian, the differential entropy is equivalent to $\frac{1}{2}\log\left((2\pi e)^m|\tilde{\Sigma}_k|\right)$ and $D(P_0||P_1)$ equals

$$\begin{aligned}
&\frac{1}{2}\Big[\mathbf{tr}\ \Sigma^{-1}\tilde{\Sigma}_k + \mathbf{tr}\ \Sigma^{-1}\mu_k\mu_k^\intercal - m - \log\left(|\Sigma^{-1}\tilde{\Sigma}_k|\right)\Big] \\
=&\frac{1}{2}\Big[\mathbf{tr}\ Q\tilde{\Lambda}_k Q^\intercal + \mathbf{tr}\ Q s_k s_k^\intercal Q^\intercal - m - \log\left(|Q\tilde{\Lambda}_k Q^\intercal|\right) \\
=&\frac{1}{2}\sum_{j=1}^{m}\left[\tilde{\lambda}_{k,j} + s_{k,j}^2 - m - \log\tilde{\lambda}_{k,j}\right] = \epsilon.
\end{aligned}$$

The final equality is the result of the previous theorem. This proves that this attack is stealthy.

Remark 2: The line of analysis performed here can be easily done to our alternative definition of stealthy, where the mean of P_1 is calculated by corrupted state estimate. In that case, we simply take

$$s_k = Q^\intercal\Sigma^{1/2}(\mu_k - C\hat{\tilde{x}}_{k|k-1}).$$

The Lagrange multipliers approach produces the optimal attack with ease.

We can consider more generally the system

$$x_k = Ax_{k-1} + Bu_{k-1} + Kz_k.$$

All results remain the same because the term e_k does not contain any control terms. This allows a common set up in which we can compare the effects of attack only on measurement compared to the attack only on control. This is done numerically in Section 4.

3.2 Randomized Threshold

Now that the relationship between the threshold and the damage to estimation performance has been established, we propose that the threshold may be randomly generated so as to achieve the effect of limiting the attacker's damage while raising the possibility of detection.

The design parameters in this problem is $g(\epsilon')$, the distribution of the threshold over $[0, \epsilon]$, as well as $\beta(D, \epsilon')$, the probability that an attack $\tilde{y}_k$ will be considered non-stealthy given its KL divergence value D. The attacker then chooses a factor of risk $\gamma(D) = \int_0^\infty \beta(D, \epsilon')g(\epsilon')d\epsilon'$, which is the probability that its attack $\tilde{y}_k$ is non-stealthy.

In this probabilistic approach, we seek the optimal design of β for a certain g. Optimality of β is given as follows.

Definition 5. $\beta^* : (0, \infty) \times [0, \epsilon] \longrightarrow [0, 1]$ is optimal if

1. $\beta^*(D, \epsilon') = 0$ if $D < \epsilon'$
2. $\beta^*(D, \epsilon')$ should be increasing with D
3. $\int_0^\infty \beta^*(D, \epsilon')g(\epsilon')d\epsilon' = \int_0^\infty \beta(D', \epsilon')g(\epsilon')d\epsilon' \Rightarrow D \leq D'$.

We briefly explain the intuition behind these conditions. The first simply states that any attack with KL divergence of D is stealthy if it does not exceed the threshold. The second condition means that a larger D value yields a larger probability of being non-stealthy. The final condition states that because β gives a relation between D and the risk factor $\gamma(D)$, the optimal β^* matches a given risk factor γ with the smallest D possible over all choices of β, and hence yields the slightest damage to the estimation quality.

The main result is that β^* is the same regardless of g chosen.

Theorem 6. For any given distribution $g : [0, \epsilon] \to [0, 1]$,

$$\beta^*(D, \epsilon') = \begin{cases} 1 & D \geq \epsilon' \\ 0 & \text{otherwise} \end{cases}$$

Proof. Let β be another viable function such that

$$\gamma = \int_0^\infty \beta^*(D, \epsilon')g(\epsilon')d\epsilon' = \int_0^\infty \beta(D', \epsilon)g(\epsilon)d\epsilon.$$

By condition 1, we can reduce this to

$$\int_0^D d\epsilon' = \int_0^{D'} \beta(D', \epsilon')d\epsilon' \Longrightarrow$$

$$\int_0^D (1 - \beta(D', \epsilon'))d\epsilon' = \int_D^{D'} \beta(D', \epsilon')d\epsilon'.$$

The left hand term is positive since the integrand is positive. The right hand term has a positive integrand, hence it is positive only if $D \leq D'$.

As a result, we have the function

$$\gamma(D) = \int_0^D g(\epsilon')d\epsilon'.$$

Example: Suppose we take $g(\epsilon)$ to be the uniform distribution between $(0, \Delta)$. Then

$$\gamma(D) = \begin{cases} \dfrac{D}{\Delta} & D \in (0, \Delta) \\ 1 & D \in [\Delta, \infty) \end{cases}.$$

We can see that if we choose $D = \epsilon$, the corresponding risk factor remains non-zero. Even when the attacker chooses to attack with KL divergence value of ϵ, there is a chance that the attack is non-stealthy. This is a better result compared with setting a strict threshold at ϵ in which the attack is fully stealthy.

3.3 Alternate formulation

Another formulation of the problem in the previous section is to have the attacker consider a linear tradeoff between risk of detection and attacking power. In this new scenario, we describe the optimal distribution of the threshold ϵ, which turns out to be either a uniform distribution or a combination of uniform distributions.

We set this up by considering the objective function

$$\gamma(\beta, g, D) = \int \beta(D, \epsilon')g(\epsilon')d\epsilon' - \lambda\mathcal{L}(D),$$

$\mathcal{L}(D)$ being the estimation error. λ acts as a conversion factor so that $\mathcal{L}(D)$ and $P[Alarm \mid D]$ can be compared. Note that $\mathcal{L}(D)$ is unbounded hence its minimum will be achieved for $D \to \infty$, so we add an upper bound $\overline{D}$ to possible values of D. This can be a result of the attacker's power constraints. It is only required that it has the power to achieve ϵ, that is, $[0, \epsilon] \subset [0, \overline{D}]$.

For now, we simplify the integral $\int_0^{\overline{D}} \beta(D, \epsilon')g(\epsilon')d\epsilon'$ to a function $F(D)$. This function is increasing with $F(0) = 0$ and $F(D) = 1$ for $D \geq \epsilon$. Then by abuse of notation, $\gamma(\beta, g, D) = \gamma(F, D)$.

For a given F, λ, and $\bar{D}$, the attacker would minimize the function $\gamma(F, D)$. This corresponds F to a $D(F)$ such that

$$\gamma(F, D(F)) = \min_{D'} \gamma(F, D').$$

The objective for the estimator is to design F such that the following optimality criterion is achieved: F^* is optimal iff $D(F^*) = \min_F D(F)$. The solution to this problem is constructive and will be split into cases. The following lemma shows that constructing an optimal β and g is equivalent to constructing F.

Lemma 7. If $F(D)$ is continuous and piecewise linear, then there exists β and g such that

$$F(D) = \int_0^{\overline{D}} \beta(D, \epsilon')g(\epsilon')d\epsilon'. \tag{7}$$

Proof. The interval $[0, \overline{D}]$ is divided into partitions $[d_i, d_{i+1}]$ on which F is linear, i.e., of the form $F(D) = a_iD + b_i$, where by continuity, $a_id_i + b_i = a_{i-1}d_i + b_{i-1}$ starting with $b_0 = 0$. We may take $\beta = 0$ for $D < \epsilon'$ and 1 otherwise, yielding

$$F(D) = \int_0^D g(\epsilon')d\epsilon'.$$

Define $g(\epsilon') = a_i$ for $\epsilon' \in [d_i, d_{i+1}]$. The viability of these designs can be checked simply.

Now onto our construction of F. First observe that $\mathcal{L}$ decreases with η convexly. η decreases with D convexly also, so the composition of this, $\mathcal{L}(D)$, is concave and increasing. This tells us that $\frac{d\mathcal{L}}{dD}$ is positive and decreasing.

Case 1:-$\lambda\mathcal{L}(0) < 1 - \lambda\mathcal{L}(\bar{D})$

In this case, ideally F should be designed such that $\gamma(F, D)$ have no local minima, which will result in the attacker having no desire to attack, i.e., $D(F) = 0$. This can be done by setting

$$F(D) = \min\{\ \ell D, 1\ \} \quad \text{where} \quad \ell = \max\{\ \lambda\frac{d\mathcal{L}}{dD}(0), \frac{1}{\epsilon}\ \}.$$

It can easily be seen that it satisfies the boundary conditions. It suffices to show there are no local minima. For $D \in (0, \frac{1}{\ell})$, $\frac{dF}{dD} > \lambda\frac{d\mathcal{L}}{dD}$, hence γ is increasing. As a consequence $\gamma(F, D) > \gamma(F, 0)$ for $D \in (0, \frac{1}{\ell}]$. Then for $D \in (\frac{1}{\ell}, \bar{D})$, $\frac{dF}{dD} = 0 < \lambda\frac{d\mathcal{L}}{dD}$. γ would decrease the second half to $\gamma(F, \bar{D})$. We can conclude that $\gamma(F, D) \geq \gamma(F, \bar{D}) > \gamma(F, 0)$. Hence $\gamma(F, 0)$ is the global minimum.

Case 2:$-\lambda\mathcal{L}(0) \geq 1 - \lambda\mathcal{L}(\bar{D})$

Since $\mathcal{L}$ increases with D, there must be a unique $E \in (0, \bar{D})$ such that $-\lambda\mathcal{L}(E) = 1 - \lambda\mathcal{L}(\bar{D})$. The global minimum in this scenario would either be at $D = \bar{D}$ or is a local minimum. We hope to design F such that the local minimum is located at the smallest D possible. Because $-\lambda\mathcal{L}(D) > 1 - \lambda\mathcal{L}(\bar{D})$ for $D < E$, the global minimum cannot be located in the interval $(0, E)$. The next best thing would be if the global minimum is achieved at a point D as close to E as possible.

The analysis can be split into two subcases.

Subcase 1: $E \in (0, \epsilon)$

Construct $F(D)$ in the following way. Set $F(D) = 0$ for $D \in (0, E]$. Then for a sufficiently small dE, let

$$F(D) = \min\{\ \lambda\frac{d\mathcal{L}}{dD}(E + dE)(D - E), 1\ \}$$
$$\text{for } D \in (E, \bar{D}) \text{ if } \lambda\frac{d\mathcal{L}}{dD}(E + dE) > \frac{1}{\epsilon - E}$$

and otherwise let

$$\begin{cases} \lambda\frac{d\mathcal{L}}{dD}(E + dE)(D - E) & \text{if } D \in (E, E + 2dE] \\ \frac{1 - F(E + 2dE)}{\epsilon - E - 2dE}(D - E - 2dE) & \\ & \text{if } D \in (E + 2dE, \epsilon) \\ 0 \text{ otherwise} & \end{cases}.$$

The design is chosen so that γ has negatives slope in $(E, E + dE)$ but positive slope in $(E + dE, \epsilon)$ and have $F(D) = 1$ for $D > \epsilon$.

This gives a global minimum at $E + dE$ since by design it is the location of the only local minimum and $\gamma(E + dE)$ has a lesser value than $-\lambda\mathcal{L}(E) = 1 - \lambda\mathcal{L}(\bar{D})$. With this design of F, the attacker will choose to attack with $D = E + dE$ rather than $\bar{D}$.

Since dE can be selected arbitrarily small, we can design g using the lemma to get the global minimum to be arbitrarily close to E. However taking limit of $dE \longrightarrow 0$ would lead to the global minimum being $-\lambda\mathcal{L}(E) = 1 - \lambda\mathcal{L}(\bar{D})$. Therefore, the best result that can be achieved is having the global minimum located arbitrarily close to such E.

Subcase 2: $E \in [\epsilon, \bar{D})$

In this case, for any design of F that satisfies our requirements will not be able to keep the global minimum located in the interval $(0, \epsilon)$. The only scenario is that the attacker plays $D \in (\epsilon, \bar{D})$ and the estimator raises the alarm.

Here we find the relation between ϵ and $\bar{D}$. For a given $\bar{D}$ and its corresponding E such that $\lambda\mathcal{L}(E) = 1 - \lambda\mathcal{L}(\bar{D})$, the estimator needs to choose a threshold ϵ that includes E and employ the given design of β and g. This leads the attacker to choose an attack that achieves KL divergence value arbitrarily close to E.

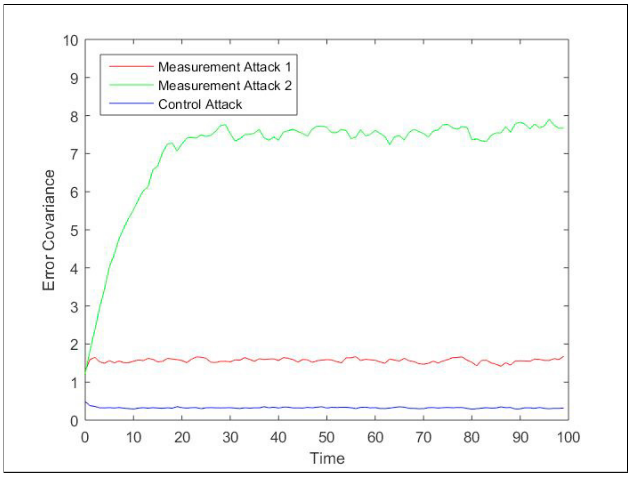

Figure 1. Measurement Attack 1, Measurement Attack 2, Control Attack with $\epsilon = 0.3$

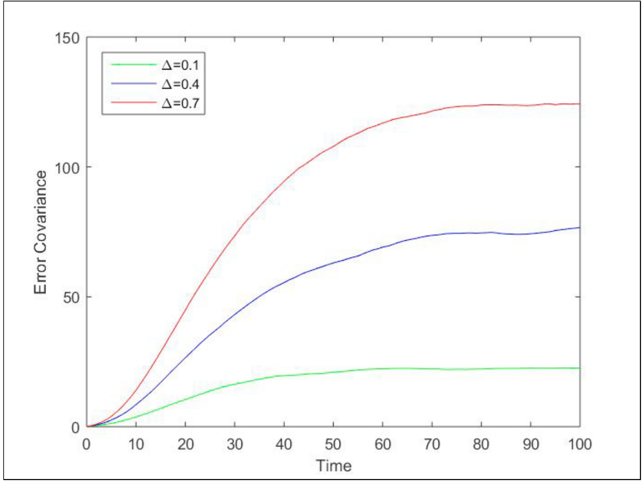

Figure 2. Error Covariance with $\Delta = 0.1$, $\Delta = 0.4$, and $\Delta = 0.7$

4. NUMERICAL SIMULATION

All of the following Monte Carlo simulations are calculated over 100 time steps, averaging over 1000 realizations.

In Figure 1. we assume the parameters

$$A = 0.9, Q = 0.3, C = 1, R = 0.6.$$

We compare three attacks. The first is the attack, labelled Measurement Attack 1, that is stealthy with respect to Definition 1. The second, Measurement Attack 2, is stealthy according to our alternate definition, where the mean of P_1 is calculated by corrupted state estimate (see Remark 1). The third is the attack on control in the work of [Kung et al. (2016)].

It can be seen that the attack on measurements proposed in this paper are comparatively more damaging than the control attack. Also, note that the control attack yields a relatively constant error covariance because it, too, does not depend on the previous history.

Comparing the two measurement attacks, the attack based on our alternate definition is more damaging. This is intuitive since the set of attacks stealthy under this alternate definition strictly includes all attacks stealthy with respect to Definition 1.

Finally, Figure 2. plots the error covariance as a result of choosing g to be uniform distributions over different intervals $(0, \Delta_i)$ under the risk factor $\gamma = 0.5$. The parameters are set at

$$A = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, \; Q = \begin{bmatrix} 0.01 & 0 \\ 0 & 0.04 \end{bmatrix}$$

$$C = [1 \quad 1], \; R = 0.2.$$

It can be seen that as Δ_i is closer to zero, the error covariance also decreases. That is because if g_i is the uniform distribution over $[0, \Delta_i]$,

$$0.5 = \frac{D}{\Delta_i} \Longrightarrow D = 0.5\Delta_i.$$

A smaller Δ_i corresponds to a smaller D.

5. CONCLUSION & FUTURE WORK

In this paper, we follow several previous works considering system attacks using KL divergence as a constraint. The recurring conclusion is that one may always find a Gaussian attack that is optimal. We further observe that there are ways for the estimator to manipulate the attacker into reducing its attack on the estimator quality by taking advantage of its necessity to remain stealthy.

There are two lines of research from this problem. One may look at optimal attacks, not only at each time k, but over a finite or infinite time horizon that minimizes average error covariance. Also, now that the attack has been characterized for both control and measurement, the attacker may consider an interplay between the two attacks in a game environment.

REFERENCES

C. Z. Bai, F. Pasqualetti, V. Gupta. Security in Stochastic Control Systems: Fundamental Limitations and Performance Bounds. In American Control Conference, 2015, Chicago, IL, July 1-3, pp. 195-200.

E. Kung, S. Dey, and L. Shi. The Performance and Limitations of ϵ-Stealthy Attacks on Higher Order Systems. IEEE Transactions on Automatic Control, accepted and to appear,2016.

F. Ye, H. Luo, S. Lu, and L. Zhang. Statistical en-route filtering of injected false data in sensor networks. In Proceedings of IEEE INFOCOM,2004.

J. Slay and M. Miller. Lessons learned from the Maroochy water breach, Critical Infrastructure Protection. vol. 253, pp. 7382, 2007.

S. Kuvshinkova SQL Slammer worm lessons learned for consideration by the electricity sector. North American Electric Reliability Council, 2003.

S. Amin, A. Cardenas, and S. Sastry, Safe and secure networked control systems under denial-of-service attacks, Hybrid Syst.: Comput. Control, vol. 5469, pp. 3145, Apr. 2009.

V. Shukla and D. Qiao, Distinguishing data transience from false injection in sensor networks, in SECON 2007, 2007.

Y. Liu, M. K. Reiter, and P. Ning, False data injection attacks against state estimation in electric power grids, in Proc. ACM Conf. Comput. Commun. Security, Chicago, IL, USA, Nov. 2009, pp. 2132.