Designing Optimal Watermark Signal for a Stealthy Attacker

Maryam Hosseini, Takashi Tanaka, Vijay Gupta

Abstract—Consider a process being controlled remotely by a controller. Let an attacker have access to the communication channel so that she is able to replace the signal transmitted by the controller with any signal she wishes. The attacker wishes to degrade the control performance maximally without being detected. The controller wishes to detect the presence of the attacker by watermarking signaling information in the control input without degrading the control performance. There are no other assumptions on the capabilities of the attacker and the controller. We present an information theoretic formulation of the problem. Our main result is to show that in the one step version of the problem, if the watermark is a Gaussian distributed random variable, then the maximal performance degradation for any given level of stealthiness for the attacker is achieved when the attacker replaces the control input with the realization of a Gaussian random variable. Conversely, we show the watermark signal that minimizes the stealthiness of a Gaussian attacker is also Gaussian.

I. INTRODUCTION

Cyber-physical system security is by now a well-motivated and popular problem (see, e.g. [1]–[4] and the references therein). One typical formulation of the problem considers plants being controlled remotely. An intruder or an attacker is able to change data transmitted across one or both of sensor-controller and controller-actuator channels. The intruder may have constraints in terms of powers, number of components she can act on and so on. The general problem is to design strategies for intruder/attacker to gain information and degrade the performance of the plant maximally and for the sensor/controller to maintain a guaranteed level of performance in spite of the attacker being present.

In this paper, we consider the attacker acting on the controller-actuator channel. The two works closest to the problem we consider are [4] and [2]. In [4], the authors presented an information theoretic study of the performance degradation that is achievable by an attacker for whom the only constraint is its desire to remain undetected. When the controller is not constrained to conduct any particular detection test, [4] characterized the largest possible estimation error covariance that can be induced by an attacker while remaining stealthy.

However, [4] assumed that the nominal control signal generated by the controller was known to the attacker. As [2] showed, if this assumption does not hold, then the controller can use a watermarking strategy for signaling the presence of an attacker to itself. Specifically, the controller can intentionally add a noisy signal to the nominal control input to detect if an attacker is present. Obviously, adding a noisy signal to the optimal control input degrades the performance of the system. In [2], the authors posed the problem of designing the watermark signal for stationary Gaussian processes to maximize the Kullback-Leibler distance between the compromised and noisy control inputs in the case of a replay attack and proposed a particular (although suboptimal) solution.

We consider the same framework as that in [4] but remove the assumption of the attacker having access to the nominal control input. As mentioned above, this introduces the possibility for the controller to watermark its input. However, differently from [2], we do not limit the attacker to a replay attack. The only constraint we place on the attacker is stealthiness. Intuitively, stealthiness is defined as the difference or distance between the nominal and the corrupted signal that characterizes the difficulty with which a detector can detect whether an attack is in progress. We use mutual information for measuring the distance. Mutual Information (MI) is an information theoretic metric proposed by Shannon which characterizes the amount of information that one random variable can provide about another random variable [5]. After presenting a notion of stealthiness in terms of MI, the main contribution of the paper is twofold. First, we consider the problem of identifying the watermark signal that minimizes the similarity (as measured by MI) between the watermarked control input and the control input as corrupted by the attacker, while the degradation in the LQG performance as compared to the nominal case is bounded. We show that the optimal watermark when an attacker is replacing the control input with a values that are realizations of a Gaussian random variable is a Gaussian signal. Further, this watermark can be obtained by solving a Semi-Definite Programming (SDP) problem. Then, we consider the problem of designing the control input that the attacker should substitute that is optimal in the sense that it is as similar as possible to the watermarked control input (for stealthiness), while being as dissimilar as possible to the nominal control input (so that the performance degradation is maximized). We show that

Maryam Hosseini is with the Department of Electrical Engineering University of Hawaii at Manoa, HI 96822 (e-mail:hosseini@hawaii.edu)
Takashi Tanaka is with ACCESS Linnaeus Centre, KTH Royal Institute of Technology, 100 44 Stockholm, Sweden (ttanaka@kth.se)
Vijay Gupta is with the Department of Electrical Engineering, University of Notre Dame, IN 46556 (e-mail:vgupta2@nd.edu).
if the controller is using a Gaussian watermark signal, the attacker should introduce a control input that is distributed according to a Gaussian random variable. In this paper, we consider only the one-step version of the problem.

This paper is organized as follows. Section II presents the system model and the problem formulation. In Section III-A, we prove that the best watermarking signal for a Gaussian controller is given by a Gaussian random variable. In Section III-B, we prove that the worst attack for a Gaussian watermark is also Gaussian.

**Notation:** A sequence of variables \( \{x_0, x_2, \ldots, x_N\} \) is denoted by \( x_0 \) or simply as \( x \) if the lower and upper limits are clear from the context. If \( N \to \infty \) we denote the infinite sequence by \( x_0^\infty \). \( N(\mu, \sigma^2) \) refers to a random variable with a Gaussian pdf with mean \( \mu \) and variance \( \sigma^2 \). \( M \times N \) matrices \( M \) and \( N \) implies that \( M \times N \) is positive definite (respectively, positive semidefinite).

Let \( f \) and \( g \) be two probability measures on the same measurable space. Let \( df/dg \) be the Radon-Nikodym derivative of \( f \) with respect to \( g \) [6]. The Kullback-Leiber distance \( D(f||g) \) between \( f \) and \( g \) is defined as:

\[
D(f||g) = \int \left( \log \frac{df}{dg} \right) df, \quad \text{if} \ \frac{df}{dg} \ \text{exists.}
\]

The mutual information between two random variable \( X \) and \( Y \) is defined as:

\[
I(X;Y) = D(f_XY||f_X \times f_Y),
\]

where \( f_X \times f_Y \) denotes the product measure. \( E[X] \) denotes the expectation of random variable \( X \).

**II. PROBLEM FORMULATION**

**a) System Model:** We consider a time invariant process described by:

\[
x_{t+1} = ax_t + u_t + w_t,
\]

\[
y_t = cx_t + v_t,
\]

where \( x_t \in \mathbb{R} \) is the state at time \( t \) and \( y_t \in \mathbb{R} \) is the sensor observation. The process noise sequence \( w_1^\infty \sim \mathcal{N}(0, \sigma_w^2) \) and the measurement noise sequence \( v_1^\infty \sim \mathcal{N}(0, \sigma_v^2) \) are white noise sequences. The initial condition \( x_0 \sim \mathcal{N}(0, P_0) \) is independent of these noise sequences. The control input \( u_t \in \mathbb{R} \) is the output of a pre-designed LQG controller.

To detect if an attacker is present, the controller may want to change the control input \( u_t \) to a watermarked version \( u_t^* \). In the sequel, we consider \( u_t^* \) to be the control input transmitted by the controller, with \( u_t = u_t^* \) if no watermarking is performed.

**b) Attack Model:** The attacker has access to the communication channel from the controller to the actuator. The attacker can replace any control input \( u_t^* \) by an input of its choice \( \tilde{u}_t \). To design this attack signal, we assume that the attacker has access to the system parameters \((a, c, \sigma_w^2, \sigma_v^2, P_0)\), measurements \( y_0^t \) and the nominal control inputs \( u_0^t \). However, even though the attacker knows that watermarking may be performed, it does not know the watermarked control inputs \( u_t^* \).

**Remark** In this paper, we focus on the one step case. In other words, we consider the case when \( t = 0 \). For simplicity, we thus remove the subscript \( t \) for all the signals. Extension of the work to multi-stage case is an interesting problem but beyond the scope of this work.

**c) Problem Statement:** To motivate the problem statement, let us consider the problem from the point of view of the controller. The controller wants to design the best possible watermark, i.e., it wants to design a watermark signal \( u^* \) so that the similarity between the watermark signal \( u^* \) and the attacker signal \( \tilde{u} \) is minimized. We propose the use of mutual information as the similarity metric between \( u^* \) and \( \tilde{u} \). At the same time, adding noise to the nominal control input degrades the performance and the controller wants to minimize this degradation. As a proxy for the LQG performance, we consider the constraint \( E[||u - u^*||^2] < \epsilon \). Thus the problem from the point of view of the controller is given by:

\[
\begin{align*}
\text{minimize} & \quad I(\tilde{u}; u^*) \\
\text{subject to} & \quad E[||u - u^*||^2] < \epsilon.
\end{align*}
\]

(1)

Now, let us consider the problem from the point of view of the attacker. The attacker aims to generate an attack signal \( \tilde{u} \) that is as similar as possible to the watermarked control input \( u^* \) to remain undetected or stealthy. Once again, to capture the notion of stealthiness, we propose the use of mutual information as the similarity metric between \( u^* \) and \( \tilde{u} \). At the same time, the attacker wants to substitute the nominal control input \( u \) by a signal \( \tilde{u} \) that is as dissimilar as possible to the nominal control input so that the performance of the process is degraded maximally. Thus the problem that attacker is interested in is given by:

\[
\begin{align*}
\text{maximize} & \quad I(\tilde{u}; u^*). \\
\end{align*}
\]

(2)

The main result of this paper is that (i) if \( \tilde{u} \) is distributed according to a Gaussian random variable (i.e., for a Gaussian attacker), the optimal solution of (1) is a Gaussian watermarking signal; and (ii) if \( u^* \) is distributed according to a Gaussian random variable (i.e., for a Gaussian watermarking signal), the optimal solution of (2) is a Gaussian attacker.

**Remark** Note that the problem (1) is formally similar to the sequential rate-distortion problem in [7]. However, the problem in [7] considers the same two signals in the
optimization objective and the constraint. In our problem, the optimization objective is the MI between $\tilde{u}$ and $u^*$, while the constraint is in terms of the variables $u$ and $u^*$. Similarly, while the problem formulation in (1) is similar to privacy-accuracy tradeoff problem in [8], but the solution here is completely different than the one presented in [8]. The problem (2) is also formally similar to the one considered in [9] and our proof follows similar principles.

III. MAIN RESULTS

A. Optimal Watermark for a Gaussian Attacker

In this section, we assume that the attacker’s policy is Gaussian, i.e., $p(\tilde{u}|u)$ can be written in the form of

$$\tilde{u} = \Sigma_{21}^{-1} u + \xi, \quad \xi \sim \mathcal{N}(0, \Sigma_{22} - \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12})$$

where $\xi$ is independent of $u$, and that the joint distribution $p(u, \tilde{u})$ is a zero-mean Gaussian distribution with a covariance matrix

$$\begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \triangleq \mathbb{E}_p \begin{bmatrix} u \\ \tilde{u} \end{bmatrix} \begin{bmatrix} u \\ \tilde{u} \end{bmatrix}^\top.$$  (4)

Under this assumption, we next observe that an optimal watermarking policy for (1) is also Gaussian.

**Theorem 1.** There exists an optimal policy $p(u^*|u)$ for (1) that can be written in the form of

$$u^* = \Sigma_{31}^{-1} u + \eta, \quad \eta \sim \mathcal{N}(0, \Sigma_{33} - \Sigma_{31} \Sigma_{11}^{-1} \Sigma_{13})$$

with some matrix $\Sigma_{31}, \Sigma_{33}$ such that $\Sigma_{33} - \Sigma_{31} \Sigma_{11}^{-1} \Sigma_{13} \succeq 0$, where $\eta$ is independent of $u$.

**Proof** Suppose $p(u^*|u)$ is an arbitrary feasible watermarking policy for (1) such that the value of the objective function is $c$. It is sufficient to prove that there always exists another feasible watermarking policy $p'(u^*|u)$ in the form of (5) such that the value of the objective function is $c' \leq c$.

To this end, let $p(u, u^*) \triangleq p(u^*|u)p(u)$ be a joint distribution induced by a given watermarking policy $p(u^*|u)$, and without loss of generality, assume $p(u, u^*)$ is zero-mean. Let

$$\begin{bmatrix} \Sigma_{11} & \Sigma_{13} \\ \Sigma_{31} & \Sigma_{33} \end{bmatrix} \triangleq \mathbb{E}_p \begin{bmatrix} u & u^* \\ \tilde{u} & \tilde{u}^* \end{bmatrix} \begin{bmatrix} u & u^* \\ \tilde{u} & \tilde{u}^* \end{bmatrix}^\top.$$  (6)

be the covariance matrix. Notice that if $p(u^*|u)$ is given, a joint distribution

$$p(u, \tilde{u}, u^*) \triangleq p(u^*|u)p(\tilde{u}|u)p(u)$$

is also determined, and its covariance matrix is

$$\Sigma \triangleq \mathbb{E}_p \begin{bmatrix} u & u & u^* \\ \tilde{u} & \tilde{u} & \tilde{u}^* \end{bmatrix} \begin{bmatrix} \Sigma_{11} & \Sigma_{12} & \Sigma_{13} \\ \Sigma_{21} & \Sigma_{22} & \Sigma_{23} \\ \Sigma_{31} & \Sigma_{32} & \Sigma_{33} \end{bmatrix}.$$  (7)

We have used the fact that

$$\mathbb{E}_p \tilde{u} u^T = \mathbb{E}_p (\Sigma_{21} \Sigma_{11}^{-1} u + \xi) u^T = \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{13}.$$  

Now, consider an alternative policy $p'(u^*|u)$ defined by

$$u^* = \Sigma_{31}^{-1} u + \eta, \quad \eta \sim \mathcal{N}(0, \Sigma_{33} - \Sigma_{31} \Sigma_{11}^{-1} \Sigma_{13})$$

where $\eta$ is independent of $u$, and denote the induced joint distribution by

$$p'(u, \tilde{u}, u^*) \triangleq p'(u^*|u)p(\tilde{u}|u)p(u).$$

Clearly, $p(u, \tilde{u}, u^*)$ and $p'(u, \tilde{u}, u^*)$ share the same covariance matrix (7). Thus, if $p(u^*|u)$ is a feasible policy (i.e., $\mathbb{E}_p\|u - u^*\|^2 < c$) then $p'(u^*|u)$ is also feasible, since $\mathbb{E}_p\|u - u^*\|^2 = \mathbb{E}_p\|u - u^*\|^2$ follows from the fact that $p(u, \tilde{u}, u^*)$ and $p'(u, \tilde{u}, u^*)$ have the same covariance matrix. Moreover,

$$I_p(\tilde{u}; u^*) - I_{p'}(\tilde{u}; u^*)$$

$$= \int \log \frac{dp_{\tilde{u},u^*}}{dp_{\tilde{u},u^*}} dp_{\tilde{u},u^*} - \int \log \frac{dp'_{\tilde{u},u^*}}{dp'_{\tilde{u},u^*}} dp'_{\tilde{u},u^*}$$

$$= \int \log \frac{dp_{\tilde{u},u^*}}{dp_{\tilde{u},u^*}} dp_{\tilde{u},u^*} - \int \log \frac{dp'_{\tilde{u},u^*}}{dp'_{\tilde{u},u^*}} dp'_{\tilde{u},u^*}$$

$$= \int \log \frac{dp_{\tilde{u},u^*}}{dp'_{\tilde{u},u^*}} dp_{\tilde{u},u^*} - \int \log \frac{dp'_{\tilde{u},u^*}}{dp'_{\tilde{u},u^*}} dp'_{\tilde{u},u^*}$$

$$= \int \left( \int \log \frac{dp_{\tilde{u},u^*}}{dp'_{\tilde{u},u^*}} dp_{\tilde{u},u^*} \right) dp_{\tilde{u}}$$

$$= \int D(dp_{\tilde{u},u^*}||dp'_{\tilde{u},u^*}) dp_{\tilde{u}} \geq 0$$

where (9) follows from the fact that $\int \log \frac{dp_{\tilde{u},u^*}}{dp'_{\tilde{u},u^*}} dp_{\tilde{u},u^*} = \int \log \frac{dp'_{\tilde{u},u^*}}{dp_{\tilde{u},u^*}} dp_{\tilde{u},u^*}$, since $\log \frac{dp_{\tilde{u},u^*}}{dp'_{\tilde{u},u^*}}$ is a quadratic function and $p$ and $p'$ have the same second order moment. Thus we have constructed a Gaussian policy $p'(u^*|u)$ attaining the objective value $c' \leq c$.

Our second result presents a computationally efficient method to synthesize an optimal watermark policy. In particular, we show that optimal watermark policy can be obtained by solving a semidefinite program. Note that from Theorem 1, to characterize the optimal watermark policy, we need to specify $\Sigma_{13}$ and $\Sigma_{33}$.

**Theorem 2.** Suppose (3) and (4) are fixed. Then, an optimal watermark policy is given by (5), where $\Sigma_{13}$ and $\Sigma_{33}$ are obtained as the optimal solution to a determinant-
maximization problem

\[
\minimize_{\Pi > 0, \Sigma_{31}, \Sigma_{33}} \frac{1}{2} \log \det \Sigma_{22} - \frac{1}{2} \log \det \Pi \\
\text{subject to} \quad \Sigma \succeq 0, \quad \text{Trace}(\Sigma_{11} - \Sigma_{31} - \Sigma_{13} + \Sigma_{33}) \leq \epsilon
\]

Thus, \( I(\tilde{u}; u|u^*) \) is minimized, and consequently \( I(\tilde{u}; u^*) \) is maximized, if \( u \) and \( \tilde{u} \) are jointly Gaussian. 2

\[
I(\tilde{u}; u|u^*) = I(\tilde{u}; u, u^*) - I(\tilde{u}; u^*) = I(\tilde{u}; u) + I(\tilde{u}; u^*|u) - I(\tilde{u}; u^*) = I(\tilde{u}; u) - I(\tilde{u}; u^*),
\]

where (11) is obtained by using the fact that \( \tilde{u} \rightarrow u \rightarrow u^* \) is a Markov chain. Given that \( I(\tilde{u}; u) = I_u \), maximizing \( I(\tilde{u}; u^*) \) is equivalent to minimizing \( I(\tilde{u}; u|u^*) \).

Also, since \( u \) and \( u^* \) are jointly Gaussian, we can write \( u^* = au + \xi \), where \( \xi \) is a Gaussian random variable that is independent of \( u \). Since \( \xi \) is also independent of \( \tilde{u} \), we can use the entropy power inequality to write

\[
e^{2h(u^*|u)} \geq e^{2h(au|u)} + e^{2h(\xi|\tilde{u})} = e^{2h(au|u)} + e^{2h(\xi|\tilde{u})}
\]

with equality if and only if \( u \) and \( \tilde{u} \) are jointly Gaussian. Now, note that

\[
e^{2h(u^*|u)} = e^{2h(u^*) - 2I(\tilde{u}; u^*)} = 2\pi e^e \xi^e.
\]

Thus, substituting (12) in (13) we have

\[
2\pi e^e \xi^e e^{-I(\tilde{u}; u^*)} \geq 2\pi e^e \xi^e e^{-I(\tilde{u}; u)}
\]

\[
|a|^2 + 2\pi e^e \xi^e e^{-I(\tilde{u}; u)}
\]

\[
\Rightarrow -I(\tilde{u}; u^*) \geq \log \left( (2\pi e^e \xi^e) e^{-I(\tilde{u}; u)} \right)
\]

\[
|a|^2 + 2\pi e^e \xi^e e^{-I(\tilde{u}; u)} - \log(2\pi e^e \xi^e),
\]

where, once again, the inequality holds with equality if and only if \( u \) and \( \tilde{u} \) are jointly Gaussian. Finally, substituting (14) in (11), yields

\[
I(\tilde{u}; u|u^*) \geq I(\tilde{u}; u) + \log \left( (2\pi e^e \xi^e) e^{-I(\tilde{u}; u)} |a|^2 + (2\pi e^e \xi^e) - \log(2\pi e^e \xi^e), \right)
\]

with equality if and only if \( u \) and \( \tilde{u} \) are jointly Gaussian. Now, notice that \( I(\tilde{u}; u) = I_u \), which is a constant. Thus, \( I(\tilde{u}; u|u^*) \) is minimized, and consequently \( I(\tilde{u}; u^*) \) is maximized, if \( u \) and \( \tilde{u} \) are jointly Gaussian.  \( \square \)

B. Worst Attacker for a Gaussian Watermark Signal

We now present the result complementary to Theorem 1 by showing that the solution of the optimization problem in (2) is also a Gaussian signal. In this optimization problem, the attacker aims to maximize the performance degradation while retaining its stealthiness. Thus, it wants to design an attack signal that is as similar to the watermark signal to make sure that the attack \( \tilde{u} \) is undetectable, and at the same time, it wishes to ensure that the similarity between the nominal control input and attacker doesn’t exceed an specific level \( I_u \).

Theorem 3. The optimal solution \( \tilde{u} \) of the optimization problem in (2) is a Gaussian random variable.

Proof The proof technique follows from [9] and [10]. We obtain a lower bound for \( I(\tilde{u}; u|u^*) \) and show that the lower bound is achieved if and only if \( u \) and \( \tilde{u} \) are jointly Gaussian.
IV. CONCLUSIONS

In this paper, we considered a process being controlled remotely by a controller when an attacker can replace the signal transmitted by the controller with any signal she wishes. The attacker wishes to degrade the control performance maximally without being detected. The controller wishes to detect the presence of the attacker by watermarking signaling information in the control input without degrading the control performance. There are no other assumptions on the capabilities of the attacker and the controller. We presented an information theoretic formulation of the problem. Our main result was to show that in the one step version of the problem, if the watermark is a Gaussian distributed random variable, then the maximal performance degradation for any given level of stealthiness for the attacker is achieved when the attacker replaces the control input with the realization of a Gaussian random variable. Conversely, we showed the watermark signal that minimizes the stealthiness of a Gaussian attacker is also Gaussian. We also presented a SDP to identify the optimal watermarking signal.

V. ACKNOWLEDGMENT

The work of Vijay Gupta was supported partially be NSF grants 0846631 and 150016.

REFERENCES