This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/LCSYS.2017.2709621, IEEE Control Systems Letters

Preserving Privacy of Finite Impulse Response Systems

Giulio Bottegal, Farhad Farokhi, and Iman Shames

Abstract—**Adding input and output noises for increasing model identification error of finite impulse response (FIR) systems is considered. This is motivated by the desire to protect the model of the system as a trade secret by rendering model identification techniques ineffective. Optimal filters for constructing additive noises that maximizes the identification error subject to maintaining the closed-loop performance degradation below a limit are constructed. Furthermore, differential privacy is used for designing output noises that preserve the privacy of the model.**

Index Terms—**Identification, Estimation, Privacy preservation, Differential privacy**

I. Introduction

INNOVATIVE industries invest resources (e.g., money and time for research and development) to construct new systems and to improve the performance of the previously-deployed ones. To generate revenue and offset the cost of research, they ideally want to capitalize on their achievements. This is sometimes done by restricting the use of their ideas through patents or by hiding the features of their systems as trade secrets. When opting for trade secrets, reverse engineering techniques can be used by competitors to unravel their secrets. For instance, model identification tools can be utilized to identify a black-box system or to extract the parameters of a gray-box system. The gained information can be then used to reverse the financial gains. This motivates the use of methods that can render reverse-engineering techniques ineffective. Such methods, however, most often degrade the performance of the system. Therefore, a framework for balancing the need for preserving the trade secrets against maintaining the performance of the systems is required.

In this paper, linear time-invariant discrete-time finite impulse response (FIR) system are considered. Specifically, the idea of adding noises to the input and output for increasing the error of model identification is explored. A bound on closed-loop performance degradation caused by the additive noise is enforced. An optimal filter for constructing the additive input and output noises that maximizes the identification error subject to maintaining the performance degradation below a threshold is constructed. This is done for both known and unknown input sequences. The former is useful to make the identification difficult for given inputs, such as the optimal experimental design in the model identification literature [1]. The latter, which requires statistics of the input, can accommodate the belief of the designer on the reverse engineering techniques, e.g., a frequently used input for model identification purposes is a sequence of i.i.d.[1] Gaussian noise [2]. Finally, differential privacy framework is used for designing output additive noises that make the system identification difficult without any assumptions on the utilized inputs.

In differential privacy literature, noises are added to the outcome of statistical queries from databases to preserve the privacy of individuals in the database [3]. This framework was more recently used in dynamical systems [4], [5]. In differential privacy literature, most often, additive Laplace noises are used and the parameters of the noise are selected according to the sensitivity of the outcome to variations in the data (that should be kept private). However, weaker variants of differential privacy can be achieved by additive Gaussian noises. This is advantageous as adding Laplace noise can make the designer's task considerably more difficult (in terms of utilizing the outputs of the system), e.g., optimal state estimation when measurements are corrupted by Laplace noise results in non-linearities and memory issues [6].

To the best of our knowledge, the differential privacy has not been explored in the context of preserving the privacy of dynamical systems with the aim of protecting the model as a trade secret. This has been explored thoroughly in one of the sections of the paper. In addition, in this paper, the problem of preserving the privacy of the systems is cast as a concrete optimization problem that balances the need for keeping the privacy with that of the maintaining the performance. This provides a different approach to that of differential privacy in which constraints on the performance degradation cannot be enforced directly to optimally balance between privacy and performance. Finally, note that the problem of releasing the dynamical model of a system under privacy constraints was considered in [7]. In this paper, we take a different approach, i.e., we do not release the model of the system. We want to ensure that inferring an exact model relating inputs and outputs is made difficult.

The rest of the paper is organized as follows. The design of optimal additive input and output noise to hinder system identification is studied in Section II. Section III uses the differential privacy for constructing additive output noises. A numerical example is provided in Section IV. Some concluding remarks are presented in Section V.

F. Farokhi and I. Shames are with the University of Melbourne, Australia. G. Bottegal is with TU Eindhoven, The Netherlands.

e-mails: ffarokhi@unimelb.edu.au (F. Farokhi), ishames@unimelb.edu.au (I. Shames), g.bottegal@tue.nl (G. Bottegal)

The work was supported by a McKenzie Fellowship and the Australian Research Council (LP130100605).

[1] i.i.d. stands for independently and identically distributed.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/LCSYS.2017.2709621, IEEE Control Systems Letters

II. Optimal Additive Noise

Here, we investigate the use of additive noise to preserve the privacy of the model information assuming that the eavesdropper uses the best linear unbiased estimate. These results are subsequently generalized (to the case where the model of the eavesdropper is not known) when using the differential privacy framework.

A. Problem Formulation

In this paper, for sake of simplicity of presentation, linear single-input single-output (SISO) time-invariant discrete-time systems are considered. All the derivations can be extended to multi-input multi-output (MIMO) systems (see Remark 2.2 below). The system is described by the following equation

$$y_t = H(q^{-1})r_t + e_t, \tag{1}$$

where $H(q^{-1})$ represents the transfer function of the system, which is driven by the reference input r_t. The output y_t is corrupted by additive white Gaussian noise with variance σ^2, which is represented by e_t. Assume that $H(q^{-1})$ can be well-represented by a finite-impulse response (FIR) system of order n_h, i.e., $H(q^{-1}) = \sum_{k=0}^{n_h-1} h_k q^{-k}$. Hence, the dynamics of the system is completely characterized by the vector of coefficients $h := [h_0 \ldots h_{n_h-1}]^\top$. In this paper, we assume null initial conditions (that is $r_t = 0$ for $t \leq 0$), though extension to any initial condition is straightforward due to the linearity of the underlying system.

Assume that an adversary is interested in inferring on the process relating r_t to y_t by attempting to estimate h from a set of N input/output measurements $\{r_t, y_t\}_{t=1}^N$. To complicate the identification process, an additional component (which is not accessible to the adversary) can be added to the input or to the output of the system to lower the identification accuracy. Let w_t capture such an additional component, which changes the model of the system as

$$y_t = H(q^{-1})r_t + e_t + w_t. \tag{2}$$

This term can capture both the additive input and output noise as discussed, in detail, in what follows.

Assumption 2.1: The malicious entity is unaware of the presence of the additive input or output noise.

This assumption is rather conservative. When using the differential privacy framework in the next section, we can avoid such assumptions. Considering a FIR model for the system and in light of Assumption 2.1, the best linear unbiased estimate (BLUE) of h from perspective of the malicious entity is given by the standard least-squares estimate [8, Ch. 4]. Let us introduce the vectors $y := [y_1 \ldots y_N]^\top$, $e := [e_1 \ldots e_N]^\top$, and $w := [w_1 \ldots w_N]^\top$. Assuming that the system is at rest prior to the data collection (i.e., $r_t = 0$ for all $t \leq 0$) and defining the matrix $R := \begin{bmatrix}\phi_1 & \ldots & \phi_N\end{bmatrix}^T$, where $\phi_t^T := \begin{bmatrix}r_t & \ldots & r_{t-n_h+1}\end{bmatrix}$, it is evident that $y = Rh + w + e$. The least-squares estimate of h is then given by

$$\hat{h} = (R^\top R)^{-1} R^\top y. \tag{3}$$

Note that this estimator is not the true BLUE, which would require the knowledge of the second order statistics of w_t. However, it is the best that the malicious entity can do without the knowledge that w_t exists. This estimator is still unbiased because $\mathbb{E}\{\hat{h}\} = \mathbb{E}\{(R^\top R)^{-1}R^\top(Rh + w + e)\} = h + (R^\top R)^{-1}R^\top \mathbb{E}\{w + e\} = h$. Then, a measure of the accuracy of the estimation of the impulse response is the covariance matrix of $\hat{h}$ [8, Ch. 4], namely

$$P_h := \mathbb{E}\{(\hat{h} - h)(\hat{h} - h)^\top\}. \tag{4}$$

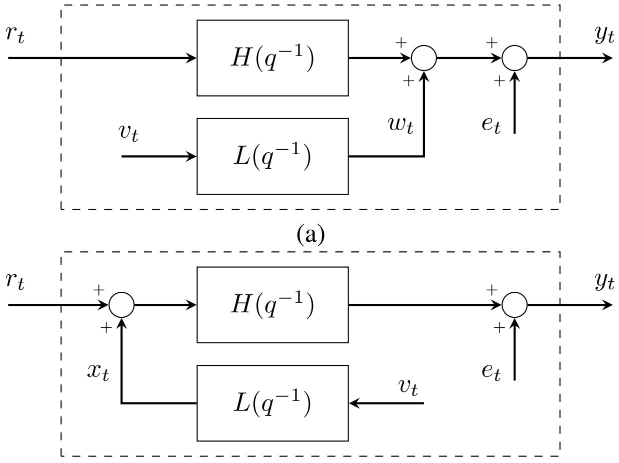
Remark 2.2: Consider a MIMO system with m inputs and p outputs, and assume for simplicity that all the subsystems $H^{ij}(q)$ are FIR systems of the same order n_h, and that all the output channels are corrupted by mutually uncorrelated white noise with the same variance. To obtain a linear regression model analogous to the SISO case, it sufficies to define $\phi_t^T := \begin{bmatrix}r_t^1 & \ldots & r_{t-n_h+1}^1 & \ldots & r_t^m & \ldots & r_{t-n_h+1}^m\end{bmatrix}$, where r_t^i is the the t-th sample of the i-th input (and $r_t^i = 0$ for $t < 1$), $\mathcal{R}_t := I_p \otimes \phi_t^T$, and $\mathbf{R} := \begin{bmatrix}\mathcal{R}_1^T & \ldots & \mathcal{R}_N^T\end{bmatrix}^T$. Then we have that $\mathbf{y} = \mathbf{R}\mathbf{h} + \mathbf{e} + \mathbf{w}$, where $\mathbf{y} = \begin{bmatrix}\mathbf{y}_1 & \ldots & \mathbf{y}_N\end{bmatrix}^T$, with $\mathbf{y}_i := \begin{bmatrix}y_i^1 & \ldots & y_i^p\end{bmatrix}^T$, and $\mathbf{h} = \begin{bmatrix}h^{11\,T} & \ldots & h^{1m\,T} \ldots & h^{p1\,T} & \ldots & h^{pm\,T}\end{bmatrix}^T$ collects the parameters of the impulse responses of the subsystems.

The additional input w_t determines the quality of the estimated system $\hat{h}$ by entering into the expression of the parameter covariance matrix P_h. Intuitively, the higher the power of w_t, the higher P_h (and thus the lower the identification accuracy). On the other hand, w_t has an undesired effect on the output power. Therefore, the additive noise is designed to increase the total variance of $\hat{h}$ (expressed through the trace of P_h) while keeping low the contribution of w_t to the variance of y_t. Let $\lambda_y := \mathbb{E}\left[y_t^2 | r_t = 0,\, t \in \mathbb{Z}\right]$ be such contribution. Note that, if $r_t = 0$, the output is driven only by the stationary noise processes e_t and w_t and so λ_y is constant in t.

Problem 2.3: For a given input r, find an appropriate additive noise w_t to maximize the identification error $\mathrm{tr}(P_h)$ while keeping the performance degradation small by guaranteeing $\lambda_y \leq \gamma_1$.

In Problem 2.3, γ_1 is a pre-selected constant that reflects the maximum tolerable output variance, which is a measure of the performance degradation caused by the additive input and output noises. If γ_1 is very small, the optimal solution is add no noise. In this case, the closed-loop performance is far superior to protecting the model. However, if γ_1 is too large, the output of the system is drowned in noise and thus the system becomes practically useless.

Here, the additive noise is designed for a given sequence of inputs captured by r. This might not be generally feasible as, when dealing with causal systems, the additive noise should be designed and employed prior to receiving the entire sequence of inputs. This design methodology is however very useful to make the identification difficult for a given input, such as those in optimal experimental design in the model identification literature [1]. Alternatively, a distribution for the input signal can be considered. Furthermore, the length of the experiment N that the malicious entity is collecting to identify the system is also unknown a priori, and shall be treated as a random quantity.



(a)

(b)

Fig. 1. The schematic diagram of the closed-loop system with additive output (a) and input (b) noises. The eavesdropper only has access to the signals outside of the dashed box.

Assumption 2.4: Let $N \in \mathbb{N}$ be a random number distributed according to $\mathbb{P}\{N = \ell\} = p(\ell)$ for some $p : \mathbb{N} \to [0,1]$ such that $\sum_{\ell \in \mathbb{N}} p(\ell) = 1$. For a given N, assume that $r \in \mathbb{R}^N$ is distributed according to the conditional probability density function $p(\cdot|N)$ such that $\mathbb{P}\{r \in \mathcal{R}|N\} = \int_{r' \in \mathcal{R}} p(r'|N)\mathrm{d}r'$ for all Lebesgue-measurable sets $\mathcal{R} \subseteq \mathbb{R}^N$.

Remark 2.5: In general, the probability density function of the input signals might not be known in advance. In that case, an online or adaptive approach can be used to estimate the statistical properties of the input as more inputs are revealed over time and design (or update the design of) privacy-preserving filters based on the additional gathered information. The result of this paper can serve as a first step in that direction. This is because if rigorous treatment of the problem for known deterministic inputs or random inputs with known probability distributions is not well understood, the analysis of the online approach would not be possible (or straightforward to say the least).

In this case, the identification error P_h which is used as a measure of privacy should be replaced with $\mathbb{E}\{P_h\}$ with the expectation being taken over random variables r and N. This allows us to generalize the problem of the interest as follows.

Problem 2.6: For given distributions of random variables N and r following Assumption 2.4, find an appropriate additive noise w_t to maximize the identification error $\mathrm{tr}(\mathbb{E}\{P_h\})$ while keeping the performance degradation small by guaranteeing $\lambda_y \leq \gamma_1$.

In this paper, two families of additive noise are considered, namely, additive output noise and additive input noise. In the remainder of this section, these two families are described.

1) Additive Output Noise: Figure 1 (a) illustrates the schematic diagram of the closed-loop system with additive output noise. The additive noise w_t is modelled by a zero-mean moving-average (MA) stochastic process of the form

$$w_t = L(q^{-1})v_t, \tag{5}$$

where v_t is a sequence of i.i.d. zero-mean noise (which is not necessarily Gaussian) of unit variance and $L(q^{-1}) := \sum_{k=0}^{n_l} l_k q^{-k}$ is a FIR filter of prescribed order n_l. Then, w_t is a stationary process with zero-mean and well-defined autocovariance function [9]. The additive noise $w := [w_1 \ldots w_N]^\top$ can be expressed as $w = Lv$, where $v := [v_{-n_l+2} \ldots v_0\, v_1 \ldots v_N]^\top$ and

$$L := \begin{bmatrix} l_{n_l-1} & \ldots & l_0 & 0 & 0 & \ldots & 0 \\ 0 & l_{n_l-1} & \ldots & l_0 & 0 & \ldots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \ldots & 0 & l_{n_l-1} & \ldots & l_0 & 0 \\ 0 & 0 & \ldots & 0 & l_{n_l-1} & \ldots & l_0 \end{bmatrix}. \tag{6}$$

The identification error covariance, in this case, is

$$\begin{aligned} P_h &= (R^\top R)^{-1} R^\top \mathrm{Var}[w+e] R(R^\top R)^{-1} \\ &= (R^\top R)^{-1} R^\top (LL^\top + \sigma^2 I_N) R(R^\top R)^{-1}. \end{aligned} \tag{7}$$

Further, the output variance can be determined by

$$\lambda_y := \mathbb{E}\{y_t^2 | r_t = 0\} = \mathbb{E}\{(w_t + e_t)^2\} = \|l\|^2 + \sigma^2, \tag{8}$$

where $l = [l_0 \ldots l_{n_l-1}]^\top$.

Remark 2.7: It should be noted that by increasing the order of the noise generation filter n_l, the performance can only be improved while maintaining the same privacy guarantee. This is because the optimal solution from the lower order is always feasible in the optimization problem relating to the higher order noise filters. The order of the system is thus only dictated by the available resources for preserving the privacy of the model.

2) Additive Input Noise: Figure 1 (b) shows the schematic diagram of the closed-loop system with additive input noise. In this case, the additive input noise is denoted by x_t and is modeled by a zero-mean MA stochastic process of the form

$$x_t = L(q^{-1})v_t, \tag{9}$$

where, similarly, v_t is a sequence of i.i.d. zero-mean noise of unit variance and $L(q^{-1})$ is a FIR filter of prescribed order n_l determining the autocorrelation of x_t. Then, the new system is described by

$$\begin{aligned} y_t &= H(q^{-1})(r_t + x_t) + e_t \\ &= H(q^{-1})(r_t + L(q^{-1})v_t) + e_t. \end{aligned} \tag{10}$$

The additive noise w_t, in this case, is the contribution of x_t to the output, i.e., $w_t = H(q^{-1})L(q^{-1})v_t$. Define

$$F(q^{-1}) := H(q^{-1})L(q^{-1}), \tag{11}$$

which can be expressed as $F(q^{-1}) = \sum_{k=0}^{n_f-1} f_k q^{-k}$ with $n_f = n_h + n_l - 1$. Note that $x := [x_1 \ldots x_N]^\top$ can be expressed as $x = Fv$ with $v := [v_{-n_f+2} \ldots v_0 v_1 \ldots v_N]^\top$ and F is defined similarly to L in (6). The identification error covariance becomes

$$\begin{aligned} P_h &= (R^\top R)^{-1} R^\top \mathrm{Var}[w+e] R(R^\top R)^{-1} \\ &= (R^\top R)^{-1} R^\top (FF^\top + \sigma^2 I_N) R(R^\top R)^{-1}. \end{aligned} \tag{12}$$

Finally, it can be shown that $\lambda_y = \|f\|^2 + \sigma^2$, where $f = [f_0 \ldots f_{n_f-1}]^\top$.

B. Deterministic Input

This part is dedicated to solving Problem 2.3. The results are first presented for the output noise case.

1) Additive Output Noise: For additive output noise, Problem 2.3 can be rewritten as

$$\arg\max_{l\in\mathbb{R}^{n_l}}\ \mathrm{tr}(P_h), \tag{13a}$$

$$\text{s.t.} \qquad \lambda_y \le \gamma_1, \tag{13b}$$

where γ_1 denotes the maximum tolerated output variance. Define the performance degradation ratio

$$\rho := \frac{\mathbb{E}\{y_t^2 | r_t = 0\}}{\mathbb{E}\{y_t^2 | r_t = 0, w_t = 0\}} = \frac{\lambda_y}{\sigma^2}.$$

If the goal of the designer is to keep the performance degradation ratio below ϵ, the constant γ_1 can be selected to be smaller than $\sigma^2\epsilon$. The following lemma is instrumental to obtain an analytic solution of (13).

Lemma 2.8: Let

$$E := R(R^\top R)^{-1}(R^\top R)^{-1}R^\top\,, \tag{14a}$$

$$c := \mathrm{tr}(\sigma^2(R^\top R)^{-1})\,, \tag{14b}$$

and denote by Q_l a selection matrix such that $\mathrm{vec}(L) = Q_l l$, where $\mathrm{vec}(L)$ is a vector composed of all the columns of the matrix L. Then, for the additive noise model, $\mathrm{tr}(P_h) = l^\top Q_l^\top (I_{N+n_l-1} \otimes E) Q_l l + c$.

Proof: See [10] for the proof. ■

Defining $M := Q_l^\top (I_{N+n_l-1} \otimes E) Q_l$ and noting that the term c is independent of l (and thus can be discarded from the optimization problem), we transform (13) into

$$\arg\max_{l\in\mathbb{R}^{n_l}}\ l^\top M l \tag{15a}$$

$$\text{s.t.} \qquad l^\top l \le \gamma_1 - \sigma^2. \tag{15b}$$

The following result can be immediately proved.

Theorem 2.9: The solution of (15) is $l^* = \sqrt{\gamma_1 - \sigma^2}\eta^*$, where η^* is the normalized eigenvector corresponding to the largest eigenvalue of M.

Proof: The change of variable $\eta = l/\sqrt{\gamma_1 - \sigma^2}$ transforms the optimization problem in (15) to

$$\begin{aligned}\eta^* \in \arg\max_{\eta\in\mathbb{R}^{n_l}}\ & \eta^\top M \eta \\ \text{s.t.} \qquad & \eta^\top \eta \le 1.\end{aligned}$$

Note that $M \ge 0$ has at least one positive eigenvalue (as otherwise $M = 0$). Therefore, Courant–Fischer–Weyl min-max principle [11, p. 58] shows η^* is the normalized eigenvector corresponding to the largest eigenvalue of M. ■

It can be seen that the quality of the model identification drops linearly with increasing γ_1. At the same time, the performance degradation ratio increases linearly with γ_1. This capture the trade-off between these two objectives. Note that, for instance, simply increasing the noise variance σ^2 to the upper bound γ_1 would determine a linear increase of the identification error, as P_h is proportional to σ^2. However, this strategy is non-optimal, and Theorem 2.9 shows how to obtain the best trade-off between performance degradation and model quality degradation, namely how to get highest linear gain. A comparison between these two strategies is given in Section IV.

If, for a given application, the linear dependency between model quality degradation and system performance degradation is not suitable, one can use the following alternative formulation of the problem:

$$\arg\min_{l\in\mathbb{R}^{n_l}}\ (\mathrm{tr}(P_h))^{-1} + \gamma_2\lambda_y\,, \tag{16}$$

where γ_2 determines weight on the performance versus the privacy. This formulation is useful when the constraint on the performance is not hard (i.e., the degradation does not need to be maintained under a given level but large output variations are not pleasant). This problem is rewritten as

$$\arg\min_{l\in\mathbb{R}^{n_l}} (l^\top M l + c)^{-1} + \gamma_2\|l\|^2\,, \tag{17}$$

where c is defined in (14).

Theorem 2.10: Let $\lambda_1 \ge \lambda_2 \ge \ldots \ge \lambda_{n_l} \ge 0$ be the eigenvalues of M and $v_1, v_2, \ldots, v_{n_l}$ denote the corresponding eigenvectors. The solution of (17) is

$$l^* = \begin{cases} 0, & \lambda_1 \le \gamma_2 c^2, \\ \sqrt{1/\sqrt{\gamma\lambda_1} - c/\lambda_1}v_1, & \text{otherwise.}\end{cases}$$

Proof: See [10] for the proof. ■

2) Additive Input Noise: Similarly, Problem 2.3 can be expressed as

$$\arg\max_{l\in\mathbb{R}^{n_l}}\ \mathrm{tr}(P_h), \tag{18a}$$

$$\text{s.t.} \qquad \lambda_y \le \gamma_1. \tag{18b}$$

Using the same line of reasoning as in Lemma 2.8, we introduce the following instrumental result.

Lemma 2.11: Let Q_f be a selection matrix such that $\mathrm{vec}(F) = Q_f f$. Then, for the additive input noise model,

$$\mathrm{tr}(P_h) = f^\top Q_f^\top (I_{N+n_f-1} \otimes E) Q_f f + c\,, \tag{19}$$

where E and c are defined in (14).

Proof: The proof follows the same line of reasoning as in Lemma 2.8. ■

Now, note that the coefficients of the filter $L(q^{-1})$ and filter $F(q^{-1}) = H(q^{-1})L(q^{-1})$ are related according to

$$f = Hl, \tag{20}$$

where $H \in \mathbb{R}^{n_f \times n_l}$ is a Toeplitz matrix formed by the coefficients of h. Substituting (20) in (19) gives $\mathrm{tr}(P_h) = l^\top H^\top Q_f^\top (I_{N+n_f-1} \otimes E) Q_f H l + c$. Therefore, the optimization problem in (18) can be transformed into

$$\arg\max_{l\in\mathbb{R}^{n_l}}\ l^\top M' l, \tag{21a}$$

$$\text{s.t.} \qquad l^\top H^\top H l \le \gamma_1 - \sigma^2, \tag{21b}$$

where $M' = H^\top Q_f^\top (I_{N+n_f-1} \otimes E) Q_f H$. The following result can be immediately proved.

Theorem 2.12: Assume $H^\top H > 0$. The solution of (21) is $l^* = \sqrt{\gamma_1 - \sigma^2}(H^\top H)^{1/2}\eta^*$, where η^* is the normalized eigenvector corresponding to the largest eigenvalue of $(H^\top H)^{-1/2} M' (H^\top H)^{-1/2}$.

Proof: Introducing $\eta = (H^\top H)^{-1/2} l/\sqrt{\gamma_1 - \sigma^2}$ transforms the optimization problem in (15) to

$$\begin{aligned}\eta^* \in \arg\max_{\eta\in\mathbb{R}^{n_l}}\ & \eta^\top (H^\top H)^{-1/2} M' (H^\top H)^{-1/2} \eta \\ \text{s.t.} \qquad & \eta^\top \eta \le 1.\end{aligned}$$

The rest of the proof follows the same line of reasoning as in the proof of Theorem 2.9. ■

The condition $H^\top H > 0$ is satisfied so long as H has full column rank. This is guaranteed if $h_{n_h} \neq 0$, i.e., no fewer than n_h parameters are required for describing filter $H(q^{-1})$.

Remark 2.13: The derivations of this section hold for arbitrary noise distributions as only the first and the second moments of the noise were considered. However, the choice of the Gaussian noise is highly preferred as it makes the integration of the closed-loop system with other control loops much easier. This is an important feature as, most often, off-the-shelf systems are interconnected to achieve complex tasks. Other noise distributions do not lend themselves that easily to integration as they might violate assumptions in the design of the control loops (e.g., Laplace noise results in an increased false alarm rate for fault detection schemes).

C. Extension to regularized least-squares

We now modify the proposed privacy-preserving technique to cope with regularized least-squares estimators. The cost function associated with this type of estimators is

$$J_{\text{RLS}}(h) = \|y - Rh\|_2^2 + \eta\|h\|_{K^{-1}}^2\,, \tag{22}$$

where K is a positive semidefinite matrix (usually called a kernel) inducing desired properties in the estimates $\hat{h}$, see [12] for details on regularized methods for system identification. The solution to (22) is $\hat{h} = (R^\top R + \eta K^{-1})^{-1}R^\top y = Cy$, with obvious defintion of C. This solution is biased. Further, it can be verified (see, e.g., [12]) that the mean square error (MSE) of the estimate is given by

$$\begin{aligned}\text{MSE} =&\mathbb{E}\{(h-\hat{h})(h-\hat{h})^\top\} \quad (23)\\ =&(I_{n_h}{-}CR)hh^\top(I_{n_h}{-}CR)^\top{+}CLL^\top C^\top{+}\sigma^2CC^\top,\end{aligned}$$

the first term on the right hand side corresponding to the bias induced by the regularization penalty. Then, the results of Theorems 2.9 and 2.10 hold by redefining

$$E := C^\top C\,, \tag{24a}$$

$$c := \text{tr}((I_{n_h} - CR)hh^\top(I_{n_h} - CR)^\top + \sigma^2CC^\top)\,, \tag{24b}$$

and, accordingly, updating the definition of matrix M. Note that the identification performance depends on the parameter η, regulating the bias-variance trade off, and on the kernel matrix K. These are user choices, which are not accessible to privacy-preserving device. One possible way to circumvent this issue is to consider the best possible choice of kernel, which is given by $K = hh^\top$ [12].

D. Random Inputs

The problem of designing an additive output noise is only considered in this section. The results can be easily extended to the design of input noises following the same line of reasoning. Problem 2.6 can be cast as

$$\arg\max_{l\in\mathbb{R}^{n_l}}\ \text{tr}(\mathbb{E}\{P_h\}) \tag{25a}$$

$$\text{s.t.} \qquad \lambda_y \leq \gamma_1. \tag{25b}$$

Note that $\text{tr}(P_h) = \mathbb{E}\{c(r,N)\} + l^\top\mathbb{E}\{Q_l(N)^\top(I_{N+n_f-1}\otimes E(r,N))Q_l(N)\}l$. Although having the same definition, $Q_l(N)$, $E(r,N)$, $c(r,N)$ are used instead of Q_l, E, and c to emphasize they are functions of random variables N and r. Define $M'' := \mathbb{E}\{Q_l(N)^\top(I_{N+n_f-1}\otimes E(r,N))Q_l(N)\}$. The optimization problem in (25) can be rewritten as

$$\arg\max_{l\in\mathbb{R}^{n_l}}\ l^\top M''l, \tag{26a}$$

$$\text{s.t.} \qquad l^\top l \leq \gamma_1 - \sigma^2. \tag{26b}$$

Theorem 2.14: The solution of (26) is $l^* = \sqrt{\gamma_1 - \sigma^2}\eta^*$, where η^* is the normalized eigenvector corresponding to the largest eigenvalue of M''.

Proof: The proof follows the same line of reasoning as in Theorem 2.9. ■

Unfortunately, calculating M'' in an explicit from as a function of the distributions of N and r is generally difficult. The following remark provides a numerical algorithm for constructing an approximation of this matrix.

Remark 2.15 (Monte Carlo Simulation): Samples of possible input length N^i, $i \in \{1,\dots,\theta\}$, are selected randomly. For each N^i, ϑ samples of the inputs of length N^i can be selected. Let these samples be denoted by r^{ij}. Define $\hat{M}'' = (1/(\theta\vartheta))\sum_{i=1}^{\theta}\sum_{j=1}^{\vartheta}Q_l(N^i)^\top(I_{N^i+n_f-1}\otimes E(r^{ij},N^i))Q_l(N^i)$. Evidently, $\mathbb{P}\{\|\hat{M}'' - M''\| \geq \epsilon\} \to 0$ as both θ and ϑ tend to infinity for all $\epsilon > 0$. Therefore, by selecting enough samples, an arbitrarily close approximation of M'' with a high probability can be constructed.

III. Relationship to Differential Privacy

Throughout this section, the design of an additive output noise is only considered. The results for the additive input noise can be constructed similarly. Furthermore, h is assumed to belong to a compact set $\mathcal{H} \subseteq \mathbb{R}^{n_h}$.

Definition 3.1: The system is ϵ-differential private if $\mathbb{P}\{y \in \mathcal{Y}|h\} \leq \exp(\epsilon)\mathbb{P}\{y \in \mathcal{Y}|h'\}$ for all Lebesgue-measurable sets $\mathcal{Y} \subseteq \mathbb{R}$ and $h, h' \in \mathcal{H}$ that differ in at most only one entry, i.e., $\|h - h'\|_0 \leq 1$. The system is (ϵ,δ)-differential private if $\mathbb{P}\{y \in \mathcal{Y}|h\} \leq \exp(\epsilon)\mathbb{P}\{y \in \mathcal{Y}|h'\} + \delta$.

Note that a random variable w is said to follow the Laplace distribution with mean μ and (scaling) parameter $b > 0$ if $\mathbb{P}\{w \in \mathcal{W}\} = \int_{w\in\mathcal{W}}(2b)^{-1}\exp(-|w-\mu|/b)\text{d}w$ for all Lebesgue-measurable sets $\mathcal{W} \subseteq \mathbb{R}$.

Theorem 3.2: Assume w_t is i.i.d. Laplace random variables with $b \geq \sup_{h,h'\in\mathcal{H}:\|h-h'\|_0\leq 1}\|Rh - Rh'\|_1/\epsilon$. Then, the system is ϵ-differential private.

Proof: See [10] for the proof. ■

Note that $\sup_{h,h'\in\mathcal{H}:\|h-h'\|_0\leq 1}\|Rh - Rh'\|_1$ exists and is finite because $\mathcal{H}$ is assumed to be a compact set.

Theorem 3.3: Assume w_t is i.i.d. Laplace random variables with scaling parameter b. Then, $\lambda_y = 2b^2 + \sigma^2$.

Proof: The proof follows from that $\lambda_y := \mathbb{E}\{y_t^2|r_t = 0\} = \mathbb{E}\{w_t^2\} + \mathbb{E}\{e_t^2\} = 2b^2 + \sigma^2$. ■

Combination of Theorems 3.2 and 3.3 illustrates the trade-off between preserving privacy and closed-loop performance because as ϵ tends to zero (to achieve a higher level of privacy), the performance degrades (i.e., λ_y goes to infinity).

Proposition 3.4: Let $\mathcal{H} := \{h \in \mathbb{R}^{n_h} \,|\, \underline{h} \leq h_i \leq \overline{h}, \forall i\}$. Then, $\sup_{h,h' \in \mathcal{H}:\|h-h'\|_0 \leq 1} \|Rh - Rh'\|_1 = (\overline{h} - \underline{h}) \sum_{k=1}^{N} |r_k|$.

Proof: See [10] for the proof. ■

Proposition 3.4 illustrates that the parameter of the Laplace noise b should be increased upon admitting larger input sequences. This is because, with larger N, there are more data to extract the system parameters and, thus, the employed mechanism needs to be more conservative to avoid leaking the private information. Some relaxations of the differential privacy, e.g., (ϵ, δ)-differential privacy, that lend themselves to using a Gaussian noise, e.g., [4]. Let for any ϵ and δ define $\kappa(\epsilon, \delta) = (\mathcal{Q}^{-1}(\delta) + \sqrt{\mathcal{Q}^{-1}(\delta)^2 + 2\epsilon})/2$ with $\mathcal{Q}^{-1}$ denoting the inverse of $\mathcal{Q} : x \mapsto \int_x^\infty 1/\sqrt{2\pi} \exp(-u^2/2) \mathrm{d}u$.

Theorem 3.5: Assume w_t is i.i.d. zero-mean Gaussian noise with $\sigma \geq \kappa(\epsilon, \delta) \sup_{h,h' \in \mathcal{H}:\|h-h'\|_0 \leq 1} \|Rh - Rh'\|_2/\epsilon$. Then, the system is (ϵ, δ)-differential private.

Proof: The proof is similar to that of Theorem 3.2 and can be found in [4]. ■

IV. Numerical Examples

Consider the discrete-time system $y_t = G(q^{-1}) r_t + e_t$, where $G(q^{-1}) = (q^{-1} - 0.2q^{-2})/(1 - 0.9q^{-1} + 0.17q^{-2})$. Clearly, $G(q^{-1})$ is not a FIR system. This system can be approximated by the FIR filter $H(q^{-1}) = q^{-1} + 0.7q^{-2} + 0.46q^{-3} + 0.295q^{-4} + 0.1873q^{-5} + 0.1184q^{-6} + 0.0747q^{-7} + 0.0471q^{-8} + 0.0297q^{-9}$. The quality of the approximation is $\|H(q^{-1}) - G(q^{-1})\| = 0.0507$. In the following, we consider the deterministic input and the random input cases.

1) Deterministic inputs: We assume that a sequence of $N = 200$ input samples is injected by the malicious entity. The sequence is generated by filtering a white noise process through the low-pass filter $W(q^{-1}) = 1/(1 - 0.95q^{-1})$. We set $\sigma^2 = 1$ and $\gamma_1 = 2$, so that we are allow to double the variance of the output. First, we consider the least-squares estimator (3). We compute the identification error, given by $\mathrm{tr}(P_h)$, of least-squares equipped with the proposed privacy preserving technique using output additive noise case with $n_l = 10$, and the identification error of least-squares without any privacy preserving device. To get a fair comparison, in the latter case the noise variance is equal to the total noise variance of the former case, that is $\mathrm{tr}(FF')/N + \sigma^2$. The noise filter designed by the privacy preserving device yields $\mathrm{tr}(P_h) = 0.25$, while the variance obtained using standard least-squares is $\mathrm{tr}(P_h) = 0.17$; we have thus obtained an error increase of approximately 50%.

We now consider regularized least-squares estimators, as described in Subsection II-C. We employ as regularization kernel the stable spline kernel $K_{i,j} = \beta^{\max(i,j)}$ (see [12]), with $\beta = 0.7$. The trade off parameter η is set as $\eta = 0.1$. Using the proposed privacy preserving technique the obtained MSE of the estimated system is 0.17, while without privacy preservation (and with the same noise variance) we get a MSE equal to 0.13. Increasing η, the privacy preserving device tends to have a milder effect on the MSE, because the regularized least-squares estimator gives higher weight to the prior knowledge, penalizing the information acquired from data.

2) Random inputs: Assume that the malicious entity injects a sequence of i.i.d. zero-mean unit-variance Gaussian variables of length N chosen with equal probability from $\{10, \ldots, 20\}$. The approach of Subsection II-D is considered for constructing an optimal additive output noise with $n_l = 5$. In this example, M'' is approximated using the method of Remark 2.15 with $\theta = 100$ and $\vartheta = 1000$. Set $\sigma^2 = 0.1$ and $\gamma_1 = 0.2$. Therefore, the performance degradation ratio is upper-bounded as $\rho \leq 2$ (indeed the upper bound is tight due to the nature of the optimal solution). The optimal additive input noise, in this case, is driven by the FIR filter $L(q^{-1}) = 0.1450 + 0.0799q^{-1} + 0.2125q^{-2} + 0.0799q^{-3} + 0.1450q^{-4}$. Using the Monte Carlo simulation, it can be shown that $\mathrm{tr}(\mathbb{E}\{P_h\})/\mathrm{tr}(\mathbb{E}\{P_h | w_t = 0\}) \approx 1.9639$. Therefore, the system identification error has been approximately doubled at the expense of doubling the output variance. From Theorem 2.14, it can be inferred that $\mathrm{tr}(\mathbb{E}\{P_h\})/\mathrm{tr}(\mathbb{E}\{P_h | w_t = 0\}) = 1 + (\eta^{*\top} M'' \eta^*)/\mathbb{E}\{c(r, N)\}(\gamma_1 - \sigma^2)$.

V. Conclusions

Adding input and output noises for increasing the model identification error was considered. Optimal filters for constructing additive coloured noises were designed to maximize the identification error while maintaining the closed-performance degradation below a threshold. Differential privacy was also explored for designing output noises that preserve the privacy of the model. Future work can focus on developing performance measures for selecting the best injection point (i.e., input or output) of privacy preserving noise.

References

[1] C. R. Rojas, J. S. Welsh, G. C. Goodwin, and A. Feuer, "Robust optimal experiment design for system identification," *Automatica*, vol. 43, no. 6, pp. 993–1008, 2007.

[2] M. Gevers, "A personal view of the development of system identification: A 30-year journey through an exciting field," *Control Systems, IEEE*, vol. 26, no. 6, pp. 93–105, 2006.

[3] C. Dwork, "Differential privacy," in *Automata, Languages and Programming: 33rd International Colloquium, ICALP 2006, Venice, Italy, July 10-14, 2006, Proceedings, Part II* (M. Bugliesi, B. Preneel, V. Sassone, and I. Wegener, eds.), pp. 1–12, Berlin, Heidelberg: Springer, 2006.

[4] J. Le Ny and G. J. Pappas, "Differentially private filtering," *IEEE Transactions on Automatic Control*, vol. 59, no. 2, pp. 341–354, 2014.

[5] Z. Huang, Y. Wang, S. Mitra, and G. E. Dullerud, "On the cost of differential privacy in distributed control systems," in *Proceedings of the 3rd International Conference on High Confidence Networked Systems*, pp. 105–114, 2014.

[6] F. Farokhi, J. Milosevic, and H. Sandberg, "Optimal state estimation with measurements corrupted by laplace noise," in *Proceedings of the 55th Conference on Decision and Control*, pp. 302–307, IEEE, 2016.

[7] J. Le Ny and G. J. Pappas, "Privacy-preserving release of aggregate dynamic models," in *Proceedings of the 2nd ACM International Conference on High Confidence Networked Systems*, pp. 49–56, 2013.

[8] T. Söderström and P. Stoica, *System identification*. Prentice-Hall, 1988.

[9] A. Lindquist and G. Picci, *Linear Stochastic Systems*. Springer, 2015.

[10] G. Bottegal, F. Farokhi, and I. Shames, "Preserving privacy of finite impulse response systems," Technical Report, Available Online: http://people.eng.unimelb.edu.au/ffarokhi/files/PrivacyFIR2017.pdf, 2017.

[11] R. Bhatia, *Matrix Analysis*. Graduate Texts in Mathematics, New York, US: Springer, 1997.

[12] G. Pillonetto, F. Dinuzzo, T. Chen, G. De Nicolao, and L. Ljung, "Kernel methods in system identification, machine learning and function estimation: A survey," *Automatica*, vol. 50, no. 3, pp. 657–682, 2014.