Int. J. Appl. Math. Comput. Sci., 2025, Vol. 35, No. 4, 591–600
DOI: 10.61822/amcs-2025-0042

amcs

ROBUSTNESS ENHANCEMENT OF A DYNAMIC OBJECT MODEL AGAINST ADVERSARIAL ATTACKS

WOJCIECH SOPOT [a,*], PAWEŁ WACHEL [a], GRZEGORZ MZYK [a]

[a]Faculty of Information and Communication Technology
Wrocław University of Science and Technology
Wyb. Wyspiańskiego 27, 50-370, Wrocław, Poland
e-mail: {wojciech.sopot,pawel.wachel,grzegorz.mzyk}@pwr.edu.pl

The aim of this work is to find a compromise between the accuracy of reproducing the behaviour of a nominal (Wiener-type) object examined in a laboratory under noise-free conditions and its robustness to intentional external attacks disrupting the input signal. By linearizing the model at the operating points and replacing the computationally expensive minimax optimization criterion with a simpler one, we construct a technique that leads to models robust to adversarial attacks of bounded intensity. Simulation experiments demonstrate the robustness of the obtained models against adversarial disruptions, highlighting the method's potential applications in fields requiring high resilience, such as control systems and safety-critical environments.

Keywords: nonlinear dynamic systems, adversarial attacks, error-in-variables, Wiener model, robustness.

1. Introduction

In the past decade, much research has focused on finding ways to handle spoofing attacks on nonlinear dynamic systems. As a prominent example, one can consider attacks on the global navigation satellite system (GNSS) (Ceccato *et al.*, 2020), usually performed by jamming a satellite signal with a stronger one (Lemieszewski *et al.*, 2021; Khoei *et al.*, 2022). While such attacks can be harmful and destructive, they are often difficult to perform, partly due to the cost of required hardware. Another problem with such procedures is that there are already methods to detect them, which reduces their effectiveness. There is, however, still a possibility to perform the mentioned attacks by introducing smaller, controlled perturbations (of reduced intensity) into system input or output that may remain undetected. Although hidden within preexisting signals, such disruptions might nevertheless significantly impact a system's output.

In the context of nonlinear dynamic systems, experimental research was conducted on the impact of the aforementioned attacks on social systems (see, e.g., Avram *et al.*, 2019). An excellent example is disinformation attacks that aim to spread disinformation in society (Frąszczak, 2023). In such a case, an adversary, using subtle seeding, aims to create a situation where members of society echo initially seeded misinformation so that it disseminates, as described by Diaz Ruiz and Nilsson (2023).

It is worth noting that, as malicious attacks on social graphs become more common, the significance of research in this area is turning out to be crucial for understanding and mitigating these threats. Those attacks can take various forms, like the mentioned disinformation attacks, disruption of connectivity graphs in multi-agent systems, such as drone swarms (Reily *et al.*, 2022), or even selecting nodes in social graphs that pose the biggest threat in case of epidemic outbreaks (Bucur and Holme, 2020). It is, therefore, evident that developing safety measures against such interferences is crucial in many domains, including social media and various drone applications.

Consequently, it is not surprising that the mentioned strategies, in some contexts called adversarial attacks, and their potential consequences have recently attracted the attention of researchers in many fields of science and engineering. Currently, research on adversarial attacks mainly focuses on cases related to machine learning, particularly neural networks (Li *et al.*, 2022). One of the

*Corresponding author

most prominent examples was presented by Goodfellow *et al*. (2020), who, by slightly perturbing the values of the image's pixels, made it possible to change the neural network response.

More conceptual research was also done on adversarially trained linear regression (Ribeiro and Schön, 2023; Ribeiro *et al.*, 2022; 2023), where its relation to the least absolute shrinkage and selection operator, Lasso (Tibshirani, 1996), and Ridge regressions (Hoerl and Kennard, 1970), was shown (see also the work of Xu *et al*. (2008), where the robustness of these estimators is analyzed). Nevertheless, there is still an unexplored area for research on adversarial attacks on nonlinear dynamic systems. In the system identification field, input disturbance's difficulty has been considered for many decades in a slightly different context. The so-called error-in-variables problems are widely discussed by, e.g., Chen and Zhao (2014) or Söderström (2018). Unlike these standard approaches, the task set in our work assumes, however, that the input disturbance is malicious and can be interpreted as an intentional attack. We, therefore, optimize the model for the worst case, assuming that the value of the upper limit on the energy of the jamming signal is given a priori. More precisely, this paper addresses the problem of modelling discrete-time nonlinear systems subjected to adversarial attacks by contaminating system inputs. Unlike in the usual modelling/identification methods, we are not interested in finding the model that will fit any input to any output, by instead the goal is to find model that will be most stable (in some sense), given the control signal. Such a model can be further used to construct the mentioned physical device. Since the area is not well explored yet, we focus on a class of Wiener systems, mostly due to their structural simplicity and ability to approximate relatively complex phenomena (their extensions are used, e.g., in epidemiology models (Sunusi *et al*., 2022)), and due to the multitude of already developed tools (Mzyk, 2013; 2014). The original contribution along with the differences between the proposed solution and the existing methods can be summarized as follows:

- unlike traditional system identification tasks, where one is interested in finding a model that approximates a system, here an a priori known model is corrected/tuned for a specific input sequence, the execution of which is scheduled in a serial manner;
- the developed method is parameterized by the initially assumed attack power level on the input signal (see Definition 1), that is used to express the allowed level of external disruption,
- the proposed methodology is universal in the sense that it can be used for various types of input perturbation, whose nature is expressed as a vector norm (which leads to various analogies, i.e., Lasso regression);
- the idea of linearization of the model around the nominal operating points allows avoiding the need to solve a minimax optimization problem;
- extensive simulation studies are carried out illustrating the dependence of the form of the optimal (adversarial) model on the input disturbance power.

The paper is organized as follows. Section 2 contains problem formulation, and Section 3 focuses on a minimax criterion upper bound. The main result of the paper is presented in Section 4, where the minimax criterion is replaced by a simpler one, incorporating local linearization of the system. Sections 5 and 6 contain presentation of selected numerical results and concluding remarks, respectively.[1]

2. Problem formulation

To formally describe our modelling problem, let us consider some prototype device that can be modeled by a SISO Wiener system with finite memory, as mentioned in Introduction. In our setup, a given control sequence $\{u_t\}_{t=1-H}^{N}, N \in \mathbb{N}$, is used to excite the above-mentioned device, and the corresponding output of the system, $\{y_t\}_{t=0}^{N}$, is measured in a noise-free environment. Clearly, in such a scenario, distance

$$\frac{1}{N}\sum_{t=1}^{N}\left|y_t - f\left(\sum_{i=0}^{H}\lambda_i u_{t-i}\right)\right|^d \tag{1}$$

is equal to zero, where $\lambda \in \mathbb{R}^{H+1}$ is the impulse response of the linear subsystem dynamics, $f \in \mathcal{F}$ is a static nonlinearity (some known parametric class $\mathcal{F}$), and $d \in [1, +\infty)$. The configuration considered is shown in Fig. 1.

In the paper, both f and λ are treated as known, as per prototype's design, or as a result of initial, precise identification. This means that we assume that we are able to determine a model with a Wiener structure, of the real noise-free system, for which approximation error is negligible.

Assumption 1. For a given nonlinear system excited by a known input sequence $\{u_t\}_{t=1-H}^{N}$, the corresponding output measurements $\{y_t\}_{t=1}^{N}$ have a signal-to-noise ratio (SNR) satisfactory low, and thus can be regarded as noise-free. This condition is only assumed to hold for data acquired *prior* to the system's deployment in operational conditions.

[1]Scripts used to perform the experiments for this work are open-source, available at `https://github.com/cyber-physical-systems-group/adversarial-training-in-nonlinear-system-modelling`.

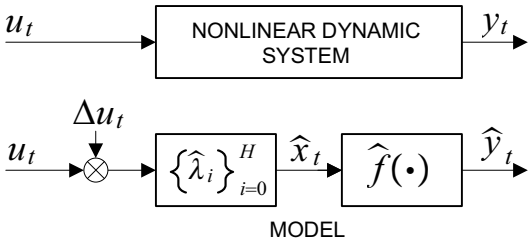

Fig. 1. Nominal system and its adversarial model.

Assumption 1 requires a comment due to a *noise-free* character of the output measurements. Note that in many production processes, at some intermediate (construction) phases, the device under development is thoroughly examined in laboratory conditions, where a control signal is precisely designed, and the system response is measured with high accuracy, i.e., with a satisfactorily high SNR. Another situation when such a scenario might occur is when there is available a blueprint of some device, which allows precise predictions, given the input data.

Nevertheless, these idealised conditions are usually not met in post-production applications of the device. This motivates the manufacturing of devices following the desired output trajectory with the highest guaranteed accuracy, in case the nominal control process (system input) is randomly or intentionally disturbed. To model such a disturbance, we use a class of signals, $PLS_p(\delta, H)$, with some a priori known $\delta > 0$ (interpreted as attack intensity), defined as follows.

Definition 1. A discrete-time signal $\{s_t\}$ is a power limited signal, denoted as $PLS_p(\delta, H)$, if, for a given $\delta > 0$, $H \in \mathbb{N}$, $p \in [1, +\infty)$, and any segment $b_t = [s_t, s_{t-1}, \ldots, s_{t-H}]^\top$, it holds that $\|b_t\|_p \leq \delta$.

Observe that, for a given δ, $PLS_p(\delta, H)$ can be interpreted as a family of disturbances with constrained energy on time intervals of length H. The p-norm chosen for normalization will depend on the nature of the disruption. For example, in mechanical systems, $p = 1$ and $p = 2$ may refer to sequences of potential and kinetic energies, respectively. Our goal is to construct a model $\{\hat{f}^* \in \mathcal{F}, \hat{\lambda}^* \in \mathbb{R}^{H+1}\}$, robust to adversarial attacks on input signal, that will serve as a reference for mass-produced devices. Examples of procedures enabling the synthesis of appropriate characteristics of a real system include the use of mechanical elements with appropriate mass or elasticity, the selection of capacitance/inductance in electrical systems (Wing, 2009), the diameter of wires in flow systems, oil density (per Bernoulli's principle), the use of materials with appropriate heat capacity (Siyu and Jian-Qiao, 2012), and others (Norquay *et al.*, 1998). This new model, $\{\hat{f}^*, \hat{\lambda}^*\}$, should be robust with respect to input perturbation $\{\Delta u_t\}$ of malicious nature; see Assumption 2.

Assumption 2. The input perturbations $\{\Delta u_t\} \in PLS_p(\delta, H)$ are unknown, although their intensity δ is known and they are adversarial for a given model $\{\hat{f}, \hat{\lambda}\}$, i.e., they maximize the mean error

$$R_d(\hat{f}, \hat{\lambda}; \{\Delta u_t\}) = \frac{1}{N} \sum_{t=1}^{N} \left| y_t - \hat{f} \left(\sum_{i=0}^{H} \hat{\lambda}_i \left(u_{t-i} + \Delta u_{t-i} \right) \right) \right|^d \quad (2)$$

for a given control sequence, $\{u_t\}_{t=1-H}^{N}$, output, $\{y_t\}_{t=1}^{N}$, and known $d \in [1, +\infty)$, as mentioned before.

Such limitation of a disturbance can be justified by physical constraints of an adversary, like a limit on the power that can introduced to the system, or by having some secondary mean of protection against attacks of higher intensity.

Thus, obtaining the new model will require solving the problem of adversarial training, formulated as a minimization

$$\{\hat{f}^*, \hat{\lambda}^*\} = \arg\min_{\hat{f}, \hat{\lambda}} Q_d(\hat{f}, \hat{\lambda}), \quad (3)$$

where

$$Q_d(\hat{f}, \hat{\lambda}) = \max_{\{\Delta u_t\} \in PLS_p(\delta, H)} R_d(\hat{f}, \hat{\lambda}; \{\Delta u_t\}). \quad (4)$$

To avoid well-known ambiguities related to the identifiability of Wiener systems and ensure the uniqueness of $\{\hat{f}, \hat{\lambda}\}$, we assume that the admissible class of models $\{\hat{f}, \hat{\lambda}\}$ contains elements with memory $\hat{\lambda}$ of known length $H < \infty$ and normalized such that $\|\hat{\lambda}\|_q = 1$, where q meets the condition $\frac{1}{p} + \frac{1}{q} = 1$.

3. Minimax criterion upper bound

As noted above, the problem considered has a *minimax* nature and thus is particularly intricate to handle in practice. In this section, we propose a *non-minimax* counterpart of (4) and explore its tightness. Obviously, a sum of the max of absolute values is greater than or equal to the max of a sum of absolute values, hence

$$Q_d(\hat{f}, \hat{\lambda}) \leq \frac{1}{N} \sum_{t=1}^{N} \max_{||\delta_t||_p \leq \delta} \left| y_t - \hat{f} \left(\sum_{i=0}^{H} \hat{\lambda}_i \left(u_{t-i} + \Delta u_{t-i} \right) \right) \right|^d, \quad (5)$$

where $\delta_t = [\Delta u_t, \Delta u_{t-1}, \ldots, \Delta u_{t-H}]^\top$. Assuming differentiability of $\hat{f}$ in working points (cf. Assumption 4),

we write the right-hand side of Eqn. (5) in terms of its deviation from linearity. Then it holds that

$$Q_d(\hat{f},\hat{\lambda}) \leq \frac{1}{N}\sum_{t=1}^{N}\max_{||\delta_t||_p\leq\delta}\Bigg|y_t - \hat{f}\left(\hat{\lambda}^\top\phi_t\right) - \hat{f}'\left(\hat{\lambda}^\top\phi_t\right)\left(\hat{\lambda}^\top\delta_t\right) + r_t\Bigg|^d, \quad (6)$$

where $\phi_t = [u_t, u_{t-1}, \ldots, u_{t-H}]^\top$ and r_t is a remainder, in general not negligible. It can be noticed that, in general, the system's parameters cannot be changed freely but just slightly augmented. This can be expressed by the following assumption.

Assumption 3. The norm of the difference between the original λ and any estimated $\hat{\lambda}$ is bounded, i.e., $\|\lambda-\hat{\lambda}\|_q \leq \varepsilon_\lambda$, with ε_λ given a priori.

As a consequence of this assumption, the following bound on the process signal variability can be derived, which will be used to further define a class of estimators of f:

$$|\lambda^\top\phi_t - \hat{\lambda}^\top\phi_t| = |(\lambda^\top - \hat{\lambda}^\top)\phi_t| \leq \|\lambda^\top - \hat{\lambda}^\top\|_q\|\phi_t\|_p \leq \varepsilon_\lambda\|\phi_t\|_p. \quad (7)$$

To approximate an upper limit of the mentioned remainder, we define the following variability indexes.

Definition 2. For a given function $g(\cdot)$, let

$$v_g(x,\delta) := \sup_{-\delta\leq\Delta x\leq\delta}|g(x+\Delta x) - [g(x) + g'(x)\Delta x]| \quad (8)$$

be its local index of nonlinear variability. Then, for any domain Ω, one can define a global index of nonlinear variability of $g(\cdot)$ as

$$V_g(\Omega,\delta) := \sup_{x\in\Omega} v_g(x,\delta). \quad (9)$$

To illustrate this concept, we provide the demonstration shown in Fig. 2.

While performing linear approximation, one often assumes the remainder to be negligible. In this work, however, we intend to find strong guarantees and so we are interested in bounding the mentioned remainder. To do so, we are about to use the previously defined global variability index V_d it to measure the estimator's, $\hat{f}$, variability. This comes down to the following assumption.

Assumption 4. For a given attack intensity δ and operational range $D := \bigcup_{t=1}^{N}[\|\phi_t\|_p - \varepsilon_\lambda\|\phi_t\|_p, \|\phi_t\|_p + \varepsilon_\lambda\|\phi_t\|_p]$, we consider a class of estimators $\mathcal{F}$, such that they are differentiable in $\{\hat{\lambda}^\top\phi_t\}_{t=1}^N$ and $r_\mathcal{F} = \sup_{\hat{f}\in\mathcal{F}} V_{\hat{f}}(D,\delta)$ is finite.

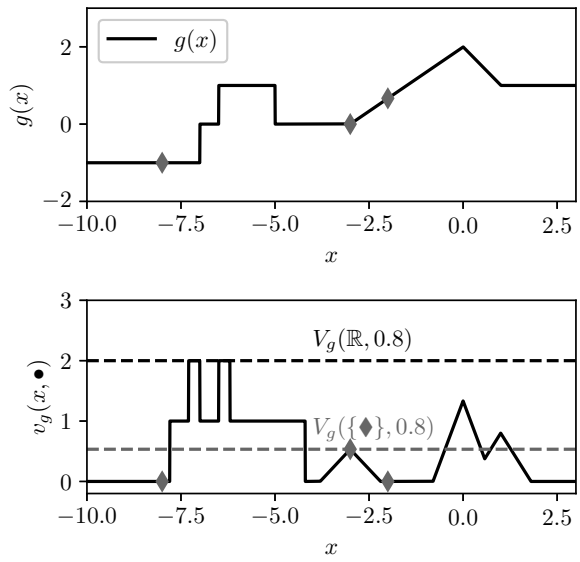

Fig. 2. Demonstration of variability indexes.

In general, one could consider using a convex domain such as $D := [-\max_t\|\phi_t\|_p, \max_t\|\phi_t\|_p]$. However, this approach would result in high values of $r_\mathcal{F}$. Fortunately, thanks to Assumption 3, D can be limited in the abovementioned manner. For smooth $\hat{f}$, $r_\mathcal{F}$ diminishes as δ approaches zero. The rate of this decay depends on the curvature of the estimators in the chosen class as well as on the control signal. In Fig. 3 the values of $r_\mathcal{F}$ were calculated as the functions of δ for several nonlinear functions (note that those results are meant to show common tendencies and, in general, are problem-dependent):

- $PWL(x)$: a continuous piecewise linear function, where the control signal was selected such that the working points were in the middle of linear segments,
- $sin(x)$: a sine function, with the control signal selected such that the working points were in the roots of the function (where it is locally the closest to being linear),
- $cos(x)$: a cosine function, with the control signal selected such that the working points were in the local maxima of the function (where it is locally the furthest to being linear),
- x^2: a quadratic function, where the control signal was selected such that the working points were spaced evenly in some range.

Note that, while usually the rate of the decay is typically slow, there are particular conditions under which $r_\mathcal{F}$ remains insignificant for relatively high values of δ.

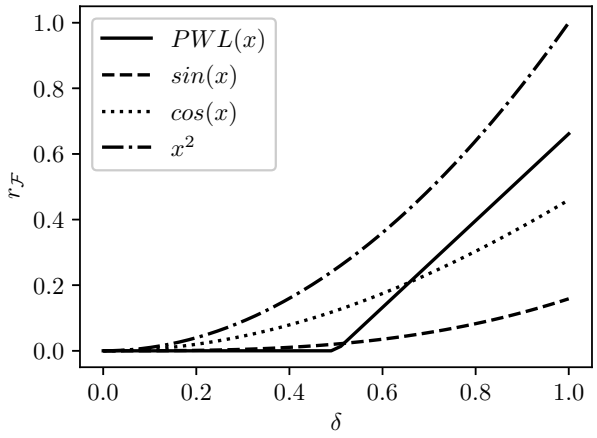

Fig. 3. Values of $r_{\mathcal{F}}$ as a function of δ for different nonlinearities, for $x = 0$.

Such a condition might be related to working points selected such that they appear in almost linear parts of $\hat{f}$.

Under Assumption 4, the right-hand side of (6) is bounded by

$$\frac{1}{N}\sum_{t=1}^{N}\max_{||\delta_t||_p\leq\delta}\left|\left|y_t-\hat{f}\left(\hat{\lambda}^\top\phi_t\right)-\hat{f}'\left(\hat{\lambda}^\top\phi_t\right)\left(\hat{\lambda}^\top\delta_t\right)\right|+r_{\mathcal{F}}\right|^d. \quad (10)$$

To find the exact value of this maximum, we use Hölder's inequality as follows, while remembering that $\|\hat{\lambda}\|_q = 1$:

$$\max_{||\delta_t||_p\leq\delta}|\hat{\lambda}^\top\delta_t| = \delta\|\hat{\lambda}\|_q = \delta. \quad (11)$$

Finally, it can be noticed that the summation terms in (10) are convex with respect to δ_t. Hence, using Eqn. (11), one can write the final upper limit for Q_d,

$$Q_d(\hat{f},\hat{\lambda}) \leq \frac{1}{N}\sum_{t=1}^{N}\left|\left|y_t-\hat{f}\left(\hat{\lambda}^\top\phi_t\right)\right|+\delta|\hat{f}'\left(\hat{\lambda}^\top\phi_t\right)|+r_{\mathcal{F}}\right|^d := \overline{Q}_d(\hat{f},\hat{\lambda}). \quad (12)$$

We will denote the difference between the upper bound and the real value of Q_d as $\xi(\hat{f},\hat{\lambda},r_{\mathcal{F}}) := \overline{Q}_d(\hat{f},\hat{\lambda}) - Q_d(\hat{f},\hat{\lambda})$. Our goal is to minimize ξ to ensure desired tightness. Since the values of $\hat{f}$ and $\hat{\lambda}$ are optimization parameters, tightness of the bound is determined by the value of $r_{\mathcal{F}}$. A decrease in δ leads to a reduction in $r_{\mathcal{F}}$, as can be seen in Fig. 3, yielding a tighter bound. This signifies the importance of selection of a proper estimator class.

Until now, we have not chosen any specific value of d. In further theoretical analysis, however, we consider two specific cases: one with $d = 1$, and the other with $d = 2$. Clearly, for $d = 1$, $R_1 = MAE$ (mean absolute error) and for $d = 2$, $R_2 = MSE$ (mean squared error).

3.1. Mean absolute error. It can be noticed that for $d = 1$ the sum in the right-hand sides of the inequality (12) can be separated into three terms as follows:

$$\begin{aligned}&Q_1(\hat{f},\hat{\lambda})\\&\leq r_{\mathcal{F}}+\frac{1}{N}\sum_{t=1}^{N}\left|\left|y_t-\hat{f}\left(\hat{\lambda}^\top\phi_t\right)\right|+\delta|\hat{f}'\left(\hat{\lambda}^\top\phi_t\right)|\right|\\&= r_{\mathcal{F}}+\underbrace{\frac{1}{N}\sum_{t=1}^{N}\left|y_t-\hat{f}\left(\hat{\lambda}^\top\phi_t\right)\right|}_{Q_T}+\underbrace{\frac{\delta}{N}\sum_{t=1}^{N}\left|\hat{f}'\left(\hat{\lambda}^\top\phi_t\right)\right|}_{Q_R}\\&=: \overline{Q}_1(\hat{f},\hat{\lambda}).\end{aligned} \quad (13)$$

The above decomposition may look similar to LASSO-style regularization, which in this scenario would look as presented in Eqn. (14); however this similarity is limited to the regularization terms, Q_R. Nevertheless, since both the residual term, Q_T, and Q_R are based on the MAE, unlike LASSO, where Q_T is based on the MSE, bistable behaviour might occur. This means that, up to some threshold value δ, the model that minimizes $\overline{Q}_1$ is the unmodified one, and past a certain threshold, the optimum shifts to a trivialized model (the one where $\hat{f}(x) = 0$):

$$Q_{\text{lasso}}(\hat{f},\hat{\lambda}) = \frac{1}{N}\sum_{t=1}^{N}\left(y_t-\hat{f}\left(\hat{\lambda}^\top\phi_t\right)\right)^2+\frac{\delta}{N}\sum_{t=1}^{N}\left|\hat{f}'\left(\hat{\lambda}^\top\phi_t\right)\right|. \quad (14)$$

3.2. Mean squared error. In the specific case when $d = 2$, one cannot separate the sum as was done for $\overline{Q}_1$. Therefore,

$$\overline{Q}_2(\hat{f},\hat{\lambda}) := \frac{1}{N}\sum_{t=1}^{N}\left(|y_t-\hat{f}\left(\hat{\lambda}^\top\phi_t\right)|+\delta|\hat{f}'\left(\hat{\lambda}^\top\phi_t\right)|+r_{\mathcal{F}}\right)^2. \quad (15)$$

While such a criterion might look too complex to minimize, it can be noticed that, for relatively small δ, some initial $\{\hat{f},\hat{\lambda}\}$ can be easily obtained, and searching for the optimal parameters might be performed by correcting the initial model. Here, one can notice similarity to the Ridge-style regularization, which would

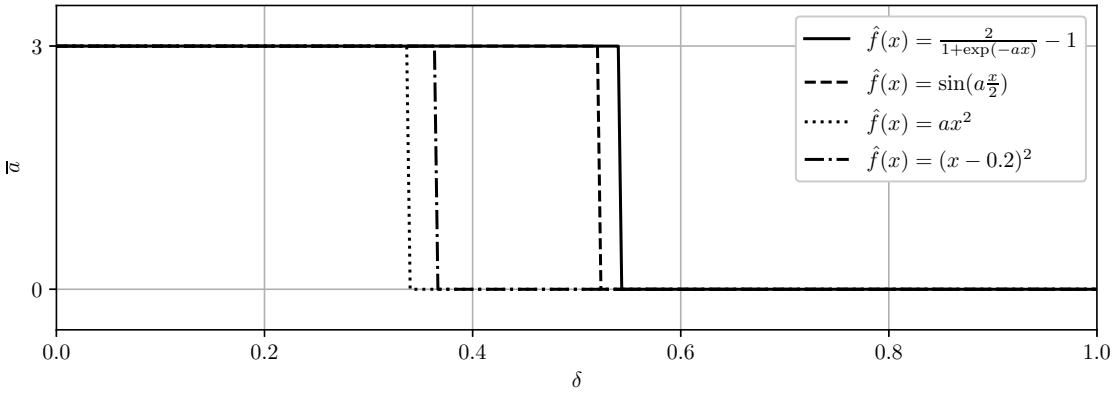

Fig. 4. Discontinuity in optimal $\overline{a}$ as a function of δ for constant $\hat{\lambda}$ and $d = 1$, and for different nonlinearities.

look as shown in Eqn. (16); however, $\overline{Q}_2$ cannot be separated into two terms:

$$Q_{\text{ridge}}(\hat{f}, \hat{\lambda}) = \frac{1}{N}\sum_{t=1}^{N}\left(y_t - \hat{f}\left(\hat{\lambda}^\top \phi_t\right)\right)^2 + \frac{\delta}{N}\sum_{t=1}^{N}\left(\hat{f}'\left(\hat{\lambda}^\top \phi_t\right)\right)^2. \tag{16}$$

4. Adversarial training algorithm

Using the results from Section 3 we are ready to formulate the following theorem about the relation between Q_d and $\overline{Q}_d$.

Theorem 1. *Consider a non-linear dynamic object that can be modeled by a SISO Wiener system with finite memory. Let Assumptions 1–4 be in force and $\{\hat{f}^*, \hat{\lambda}^*\}$ denote a model of interest, optimal under an adversarial attack of a given intensity $\delta > 0$ in the minimax sense, i.e.,*

$$Q_d(\hat{f}, \hat{\lambda}) = \max_{\{\Delta u_t\} \in PLS_p(\delta)} R_d(\hat{f}, \hat{\lambda}; \{\Delta u_t\}), \tag{17}$$

$$\hat{f}^*, \hat{\lambda}^* = \arg\min_{\hat{f}, \hat{\lambda}} Q_d(\hat{f}, \hat{\lambda}). \tag{18}$$

Then, for the corresponding model

$$\bar{f}, \{\bar{\lambda}\} = \arg\min_{\hat{f}, \hat{\lambda}} \overline{Q}_d(\hat{f}, \hat{\lambda}), \tag{19}$$

it holds that

$$Q_d\left(\hat{f}^*, \hat{\lambda^*}\right) \leq \overline{Q}_d\left(\bar{f}, \bar{\lambda}\right). \tag{20}$$

Proof. The inequality (20) is a consequence of the reasoning presented in the formulas (5)–(12) in Section 3. ■

From the above theorem, one can see that, instead of direct optimization of Q_d, which will usually be computationally complex due to its minimax nature, one can obtain $\{\bar{f}, \bar{\lambda}\}$ by minimizing $\overline{Q}_d$. While such optimization might still be complex in general, one avoids calculating the maximum directly.

5. Numerical experiments

This section investigates the applicability of the method for an exemplary Wiener system. The reference nonlinear system is defined as follows:

$$\lambda = \frac{1}{38}[5, 3, 2]^\top, \tag{21}$$

$$f(x) = \frac{2}{1 + \exp(-ax)} - 1, \quad a = 3, \tag{22}$$

where λ and a are tunable parameters. The length of λ might seem short; however, one can imagine a scenario where the input data was compressed in some way, or where the signals themselves were undersampled. Initially, we focus on the problem with $d = 1$; cf. Eqn. (1). To investigate the impact of the intensity attack, in the first experiment δ was uniformly increased, $\delta \in \{0.000, 0.001, \ldots, 1.000\}$, and the values of adversarially robust model $\{\bar{f}, \bar{\lambda}\}$ were obtained according to Eqn. (19). Input sequence $\{u_t\}$ of $N = 1000$ points was sampled from the uniform distribution $\mathcal{U}(-1, 1)$ and output sequence $\{y_t\}$ was generated in noise-free conditions. The attack's intensity was chosen to be $\delta = 0.5$. The results of simulations for nonlinearity (Eqn. (22)) and selected alternative characteristics are presented in Fig. 4. An expected bistable behaviour was observed, as mentioned in Section 3.1; however, the exact threshold value varies for each nonlinearity.

To investigate whether such behaviour occurs only for tasks with $d = 1$, the experiment was conducted for $d = 2$. The comparison of adversarially robust model $\{\bar{f}, \bar{\lambda}\}$ and the original one $\{f, \lambda\}$ is shown in Fig. 5, where it can be clearly seen that in this scenario neither initial nor trivialized models are the most adversarially robust. The upper bound for the initial model was $\overline{Q}(f, \lambda) = 2.01$ and $\overline{Q}\left(\bar{f}, \bar{\lambda}\right) = 1.97$ for the robust one, so there is a merit of employing $\{\bar{f}, \bar{\lambda}\}$. We introduce the following measures, used to check on how much correction is needed in $\{f, \lambda\}$ to achieve the mentioned

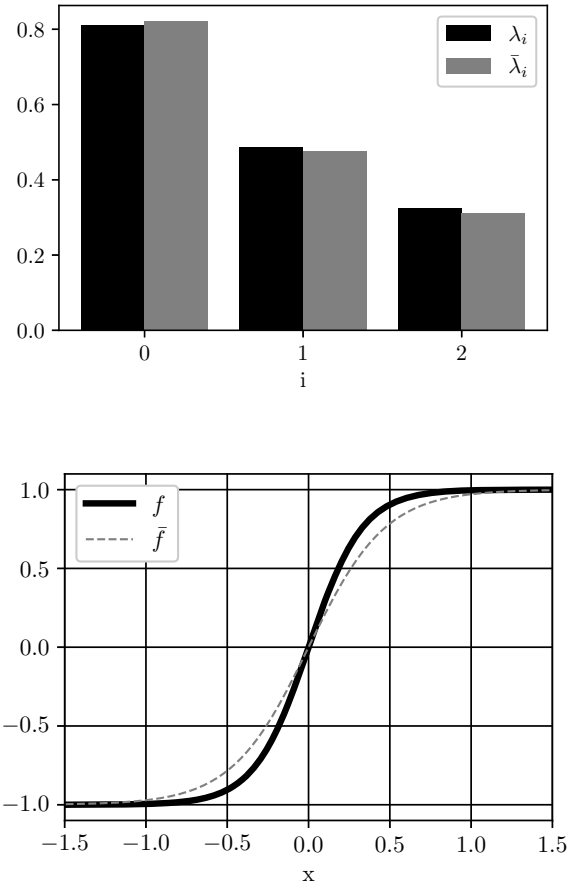

Fig. 5. Comparison of $\{f, \lambda\}$ and $\{\bar{f}, \bar{\lambda}\}$ for $\delta = 0.5, d = 2$.

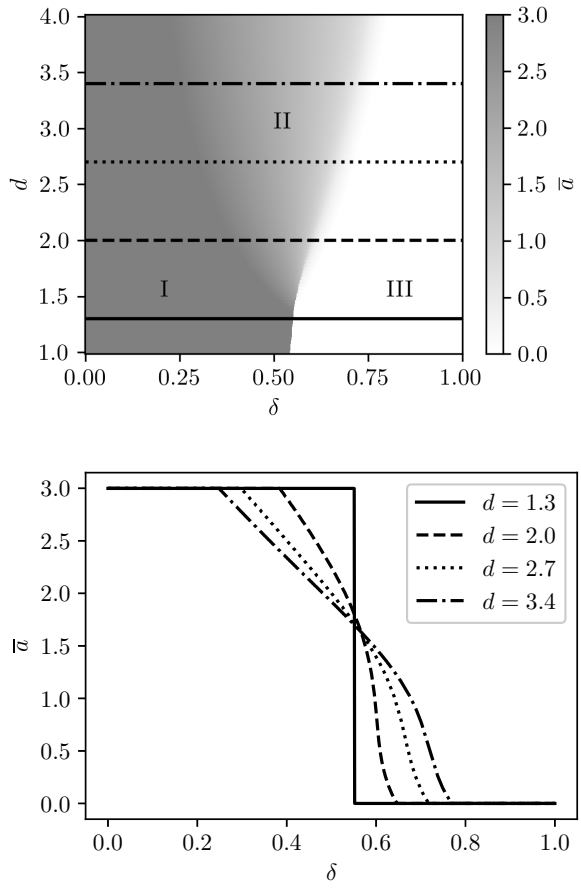

Fig. 6. Analysis for values of $d \in [1, 4]$, for constant $\hat{\lambda}$ and $\hat{f}$ as in Eqn. (22).

resilience:

$$R_\lambda = \max_i \left| \frac{\lambda_i - \bar{\lambda}_i}{\lambda_i} \right| 100\%, \tag{23}$$

$$R_f = \sup_{x \in (\ker(f))^\complement} \left| \frac{f(x) - \bar{f}(x)}{f(x)} \right| 100\%, \tag{24}$$

where $\ker(f)$ stands for a kernel, i.e., a null space of f. The values of the above indicators for the current experiment were $R_\lambda = 3.9\%, R_f = 29.7\%$. One can thus notice that, to increase adversarial robustness in this scenario, one should focus on correcting the nonlinear block of the Wiener system; thus, in further experiments, the value of $\bar{\lambda}$ is assumed to be λ.

To investigate the impact that the value of d has on $\hat{f}$, additional experiments were performed with the nonlinearity (22) for fixed $\hat{\lambda} = \lambda$, $\delta \in \{0.000, 0.001, \ldots, 1.000\}$, and $d \in \{1.000, 1.003, \ldots, 4.000\}$. The analogous Results are shown in Fig. 6. One can notice that, around $d = 2$, discontinuity disappears, and a smooth change is visible instead. Furthermore, three different regions can be distinguished:

- *passive region*, solid grey (I), relatively small δ, where the optimal model is the unchanged one—it can be seen as a situation where the attack will not have a greater impact than any noise (but one that is still of class PLS_p bounded by δ);
- *degenerated region*, white (III), relatively large δ, where the optimal model is the trivialized one–it can be seen as a situation where the adversarial one is actually a spoofing attack and defending against it through simple model correction is impossible, thus a different approach shall be taken;
- *active region*, (II), where the optimal model is different than the initial one and $\{\overline{\lambda}, \overline{f}\}$ shall be used to minimize the impact of adversarial attacks.

It can also be noticed that, for large d, the problem becomes one of minimizing the maximum cost for any t (instead of minimizing the average cost for all t).

Until now, all the experiments were done for approximately odd or even nonlinearities. In Fig. 7, the results analogous to those from Fig. 6 are presented, but

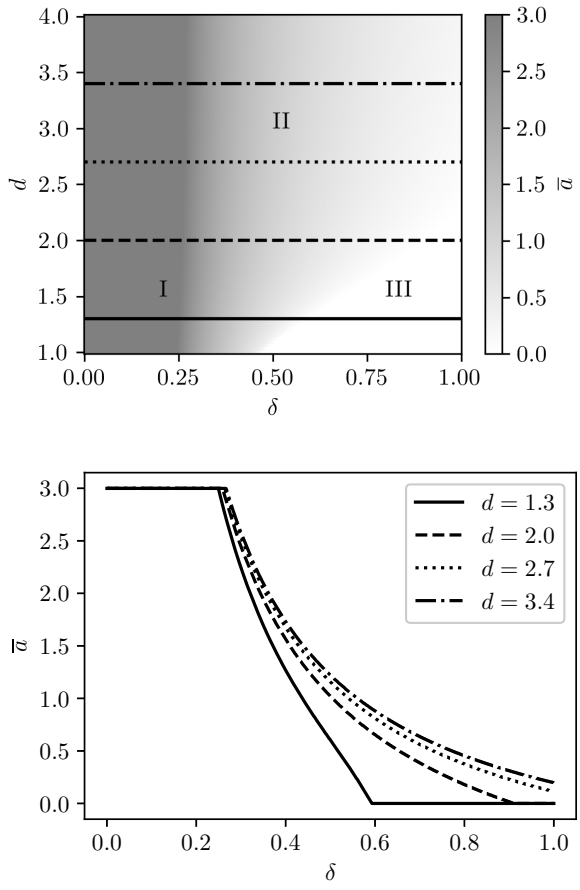

Fig. 7. Results for $\hat{f}(x) = \exp(ax)$.

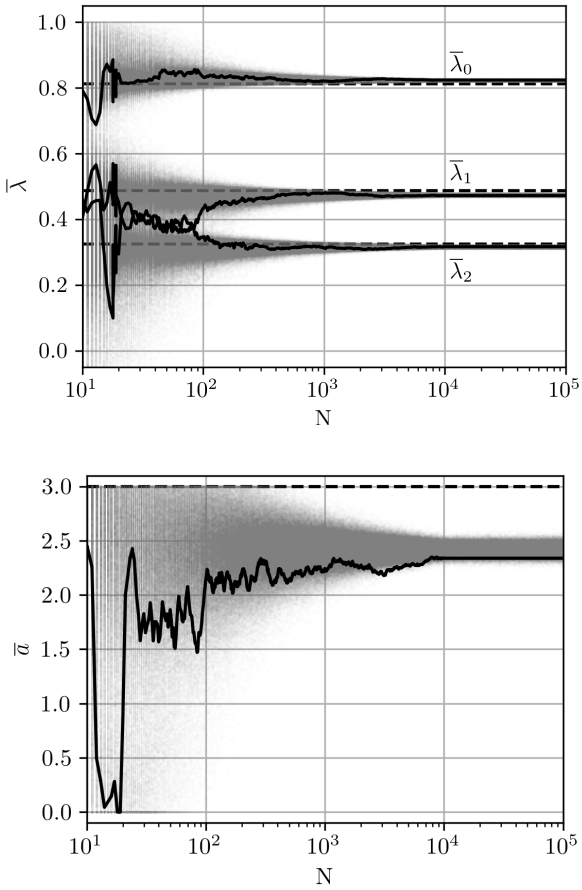

Fig. 8. Values of $\{\overline{\lambda}, \overline{f}\}$ as a function of N: dashed lines are initial values, i.e., $\{\lambda, f\}$, solid lines represent sample trajectories of corresponding values.

for $\hat{f} = \exp(ax)$. One can see that a continuous change occurs even for $d = 1$; however, the aforementioned three regions can still be distinguished. It can be suspected that whether or not the mentioned bistability occurs depends on specific nonlinearity; nevertheless, further investigation is needed to uncover the exact conditions for its occurrence.

The following experiment was conducted to investigate the impact of the number of measurements N (used for the optimization of $\overline{Q}_d$) on $\overline{\lambda}$. The reference structure (described by Eqns. (21) and (22)) was investigated for $N \in [10, 100000]$. The model of the dynamic part, $\hat{\lambda}$, was no longer fixed. The results for minimizing $\overline{Q}_2$ with $\delta = 0.5$ are presented in Fig. 8. The solid lines represent sample trajectories, where longer control sequences were created by adding samples to shorter ones. Vertical artifacts visible for the smallest values of N are caused by the use of a logarithmic scale on the plots. It can be seen that, as the value of the measurements used increases, that of $\overline{\lambda}$ becomes closer to the initial λ. One can suspect that this is the result of the sequence $\{u_t\}$ being sampled for uniform distribution, and as the value of N increases, the distribution of the values in sequence $\{\overline{\lambda}^\top \phi_t\}_{t=1}^N$ becomes more symmetric (as $N \to \infty$, the distribution would become an even function), although further investigation is required.

Noticeably, the value of $\overline{a}$ also converges to some constant, although with greater variance.

6. Conclusions

In the paper, we showed that for selected problems in modelling the nonlinear dynamic systems, particularly if an adversary might have manipulated the input, tuning the preexisting model to increase its robustness is beneficial. The concept proposed in this work can be understood as a specific type of regularization technique generalized to be applied in the modelling of Wiener-class systems. Pursuit of a perfect representation of the phenomenon observed in laboratory conditions may result in excessive sensitivity of the model to malicious disturbances on the input side. Thanks to the proposed approach, assuming a limiting intensity of the attack on the input signal,

we are able to design a model that is more resilient against adversarial attacks. Moreover, the procedure does not require performing complex operations related to the optimization of the minimax criterion and, therefore, can be used in an adaptive manner. It can be noticed that the presented approach may lead to biased models, especially for relatively high intensities of the attack, δ. In recent years it has been shown, however, that such models can shine in scenarios where there is a limited amount of data available (Ribeiro and Schön, 2023; Ribeiro *et al.*, 2022; 2023), which will be explored in further research. The above advantages seem promising from the point of view of many areas of application, such as safe drug dosing in medicine (Baayen and Hougaard, 2015), control of unmanned aerial vehicles exposed to intentional interference (Zhang *et al*., 2018), etc.

References

Avram, M.V., Mishra, S., Parulian, N.N. and Diesner, J. (2019). Adversarial perturbations to manipulate the perception of power and influence in networks, *2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), Vancouver, Canada*, pp. 986–994.

Baayen, C. and Hougaard, P. (2015). Confidence bounds for nonlinear dose–response relationships, *Statistics in Medicine* **34**(27): 3546–3562.

Bucur, D. and Holme, P. (2020). Beyond ranking nodes: Predicting epidemic outbreak sizes by network centralities, *PLOS Computational Biology* **16**(7): 1–20.

Ceccato, M., Formaggio, F. and Tomasin, S. (2020). Spatial GNSS spoofing against drone swarms with multiple antennas and Wiener filter, *IEEE Transactions on Signal Processing* **68**(10): 5782–5794.

Chen, H.F. and Zhao, W. (2014). *Recursive Identification and Parameter Estimation*, CRC Press, Boca Raton, USA.

Diaz Ruiz, C. and Nilsson, T. (2023). Disinformation and echo chambers: How disinformation circulates on social media through identity-driven controversies, *Journal of Public Policy & Marketing* **42**(1): 18–35.

Frąszczak, D. (2023). Detecting rumor outbreaks in online social networks, *Social Network Analysis and Mining* **13**(1): 91.

Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y. (2020). Generative adversarial networks, *Communications of the ACM* **63**(11): 139–144.

Hoerl, A.E. and Kennard, R.W. (1970). Ridge regression: Biased estimation for nonorthogonal problems, *Technometrics* **12**(1): 55–67.

Janczak, A. and Korbicz, J. (2019). Two-stage instrumental variables identification of polynomial Wiener systems with invertible nonlinearities, *International Journal of Applied Mathematics and Computer Science* **29**(3): 571–580, DOI: 10.2478/amcs-2019-0042.

Khoei, T.T., Ismail, S. and Kaabouch, N. (2022). Dynamic selection techniques for detecting GPS spoofing attacks on uavs, *Sensors* **22**(2): 1–18.

Lemieszewski, Ł., Radomska-Zalas, A., Perec, A., Dobryakova, L. and Ochin, E. (2021). The spoofing detection of dynamic underwater positioning systems (DUPS) based on vehicles retrofitted with aacoustic speakers, *Electronics* **10**(17): 1–11.

Li, Y., Cheng, M., Hsieh, C.J. and Lee, T.C. (2022). A review of adversarial attack and defense for classification methods, *The American Statistician* **76**(4): 329–345.

Mzyk, G. (2013). Nonparametric instrumental variables for identification of block-oriented systems, *International Journal of Applied Mathematics and Computer Science* **23**(3): 521–537, DOI: 10.2478/amcs-2013-0040.

Mzyk, G. (2014). *Combined Parametric-Nonparametric Identification of Block-Oriented Systems*, Lecture Notes in Control and Information Sciences, Vol. 454, Springer, Berlin.

Norquay, S.J., Palazoglu, A. and Romagnoli, J. (1998). Model predictive control based on Wiener models, *Chemical Engineering Science* **53**(1): 75–84.

Reily, B., Coniff, C., Rogers, J.G. and Reardon, C. (2022). Disruption of connectivity graphs in uncertain multi-agent systems, *IEEE International Conference on Omni-layer Intelligent Systems (COINS), Barcelona, Spain*, pp. 1–6.

Ribeiro, A.H. and Schön, T.B. (2023). Overparameterized linear regression under adversarial attacks, *IEEE Transactions on Signal Processing* **71**(2): 601–614.

Ribeiro, A.H., Zachariah, D. and Schön, T. (2022). Surprises in adversarially-trained linear regression, *arXiv* 2205.12695.

Ribeiro, A.L.P., Zachariah, D., Bach, F. and Schön, T. (2023). Regularization properties of adversarially-trained linear regression, *arXiv* 2310.10807.

Siyu, W. and Jian-Qiao, S. (2012). A physics-based linear parametric model of room temperature in office buildings, *Building and Environment* **50**: 1–9.

Söderström, T. (2018). *Errors-in-Variables Methods in System Identification*, Springer, Cham.

Sunusi, A.B., Abdulkarim, I.H., Auwal, A.B. and Abhiwat, K. (2022). Optimizing Hammerstein–Wiener model for forecasting confirmed cases of COVID-19, *International Journal of Applied Mathematics* **52**(1): 1–10.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society: Series B (Methodological)* **58**(1): 267–288.

Wing, O. (2009). Circuit dynamics, *Classical Circuit Theory*, Springer, Boston, pp. 1–24.

Xu, H., Caramanis, C. and Mannor, S. (2008). *Robust Regression and Lasso*, Advances in Neural Information Processing Systems, Vol. 21, Curran Associates, Inc., Vancouver.

Zhang, Y., Chen, Z., Zhang, X., Sun, Q. and Sun, M. (2018). A novel control scheme for quadrotor UAV based upon active disturbance rejection control, *Aerospace Science and Technology* **79**: 601–609.

Wojciech Sopot was born in 1997 in Poland. He received his MSc degree in control engineering and robotics in 2021 and his MSc degree in technical physics in 2024, both from the Wrocław University of Science and Technology (WUST). Currently he is pursuing a PhD in information and communication technology at the WUST. His research interests include nonlinear system modelling, estimation theory and machine learning.

Paweł Wachel (IEEE member) was born in 1980. He received his MSc degree in 2004 and his PhD degree in 2008, both in control engineering and robotics from the Wrocław University of Technology (WUST), Poland. Since 2008, he has been with that university, where he is currently an associate professor of computer science at the Faculty of Information and Communication Technology. During his career, Prof. Wachel has been a postdoctoral researcher with the neuroengineering lab (BrainLab) that is part of the Biomedical Signal Processing Group at the Faculty of Fundamental Problems of Technology, WUST. In his research, he focuses on system identification under limited prior knowledge, nonasymptotic aspects of machine learning, and signal processing.

Grzegorz Mzyk was born in 1973 in Poland. He received his MSc, PhD, and DSc degrees in control engineering from the Wrocław University of Science and Technology in 1998, 2002, and 2015, respectively. He is currently an associate professor at that university and teaches courses in control theory and system identification. His research interests include identification of Hammerstein and Wiener systems. He is the author of the book *Combined Parametric-Nonparametric Identification of Block-Oriented Systems* (Springer, 2015). Currently, he is implementing projects related to the use of mathematical models in production organization and error detection in databases.

Received: 24 December 2024
Revised: 3 June 2025
Re-revised: 5 June 2025
Accepted: 4 July 2025