New probabilistic guarantees on the accuracy of Extreme Learning Machines: an application to decision-making in a reliability context

ROBERTO ROCCHETTA[1]

[1]*Department of Mathematics and Computer Science, Technical University of Eindhoven, The Netherlands.*
E-mail: r.rocchetta@tue.nl

This work investigates new generalization error bounds on the predictive accuracy of Extreme Learning Machines (ELMs). Extreme Learning Machines are a special type of neural network that enjoy an extremely fast learning speed thanks to the convexity of the training program. This feature makes ELMs particularly useful to tackle online learning tasks. A new probabilistic bound on the accuracy of ELM is prescribed thanks to scenario decision-making theory. Scenario decision-making theory allows equipping the solutions of data-based decision-making problems with formal certificates of generalization. The resulting certificate bounds the probability of constraint violation for future scenarios (samples). The bounds hold non-asymptotically, distribution-free, and therefore quantify the uncertainty resulting from limited availability of training examples. We test the effectiveness of this new method on reliability-based decision-making problem. A data set of samples from the benchmark problem on robust control design is used for the online training of ELMs and empirical validation of the bound on their accuracy.

Keywords: Extreme Learning Machines, Scenario theory, Generalization bounds, Reliability, Decision-making, Machine learning.

1. Introduction

In the last decade, artificial intelligence and machine learning (ML) techniques grew in popularity within industry, in the medical field, finance and academia. This huge success is mostly due to the growing availability of computational resources and the ability of ML models to tackle complex engineering tasks directly from large volumes of data. ML models have proven to be a valuable asset in many reliability engineering and system safety applications and Xu and Saleh (2021) recently reviewed ML-based approaches to fault detection, remaining useful life estimations, and for maintenance tasks Rocchetta et al. (2019). Despite the many success stories, the use ML model for safety-critical applications remains challenging. Some of the main challenges can be summarized as follows.

Exhaustive testing: Limited coverage of the event space due to lack of data concerning rare failure or other low-probability events. How to quantitatively define 'Exhaustive' for specific safety-critical applications?

Explainable results: Classical ML models are often complex and work as black boxes inducing a lack of trust or practitioners when the model has to be used for safety-critical application.

Regulatory barriers: A lack of standards dedicated to the use of ML on safety-critical applications. Lack of clear regulation concerning liabilities because of failures of ML models. Also, several ML methods required environment-based learning. This may lead to a (seemingly not acceptable) public exposure to risks.

Non-stationarity and adaptability: Data are often assumed generated from a stationary (albeit unknown) probability distribution. There is only a limited number of foundation mathematical works that deal with non-stationary probability and related issues in ML.

Machine learning models are usually trained from static data and the more accurate results are obtained for large datasets and from an exhaustive exploration of the event space. In many reliability applications, however, the data set size may be small and new data not representative of past situations (non-stationary changing environments). Online learning and generalization error analysis methods try to assess some of these issues by continuously updating/re-training ML models and by assessing their reliability under lack of data.

An Extreme Learning Machine is a special type of artificial neural network but with non-tunable input weights, Huang et al. (2006). The architecture of ELM includes (generally) a unique layer of hidden neurons. The tunable output weights can be optimized analytically, and thus very efficiently. This makes ELMs very useful to tackle online learning problems and real-time decision-making tasks. The accuracy and robustness of ELMs are major concerns for ML analysts, especially if these models are used within safety-

Proceedings of the 31st European Safety and Reliability Conference
Edited by Bruno Castanier, Marko Cepin, David Bigaud, and Christophe Berenguer
Published by Research Publishing, Singapore.
ISBN: 978-981-18-2016-8; doi:10.3850/978-981-18-2016-8_597-cd

critical use cases. A lot of research has been devoted to adders the robustness and generalization of ML models, for instance, Luca Oneto Oneto (2018) compared generalization error bounds for ML models derived from re-sampling strategies, complexity-based methods, compression-based methods, and PAC-Bayes bounds and others. Test-based (re-sampling) methodologies are often based on an heuristic estimation of the error probability of the model and typical examples are the k-fold cross-validation method Stone (1974), bootstrapping Isaksson et al. (2008), jackknife method Efron and Stein (1981), and the leave-one-out method Shao and Er (2016). These are well-established and probably some of the most common among data scientists and practitioners because of their ease of interpretation and efficacy. However, these methods can be computationally very intensive, especially for large models and data sets. Moreover, the need to estimate the empirical risk via a test set inevitably reduce the amount of data available for training and cross-validation and bootstrapping method may result in unreliable results for extremely small data sets Isaksson et al. (2008). A similar issue was recently discussed by Gautheron et al. (2020). In contrast to empirical test set-based methods, theoretical generalization error bounds are mathematically derived and do not require an empirical estimation of the model performance on unseen data.

Scenario decision-making theory is a new and powerful mathematical framework to formally address the generalization problem concerning data-driven decision-making tasks Campi et al. (2018). Scenario theory finds its roots in Statistical learning theory and shares some similarities with other works on Compression learning and Complexity-based Vapnik–Chervonenkis. Scenario-based generalization error bounds are computed from a statistical measure of the complexity of decisions and with no assumption on the data generating mechanism. If the decision-making problem is convex, a data-independent (a-priori) generalization bound can be obtained independently from the elements within the data set. The term a-priori means that the bounds are obtained before calculating the decision. For instance, the probability distribution error for the solution of convex problems is upper bounded by a beta distribution whose parameters just depend on the number of decision variables and on the size of the data set Campi and Garatti (2008). However, data-independent bounds can be conservative in many cases. To amend for this conservativism, data-dependent (a-posteriori) generalization bounds, have been recently derived and are specifically tuned to the solution of a scenario decision-making problem and the scenarios within the data set, Campi and Garatti (2018). Scenario theory has been extensively studied for convex decision-making problems. Care et al Carè et al. (2015) showed that L_∞ convex minimization problems, the probability of exceeding empirical risk levels follows a Dirichlet distribution with Beta marginals. Recently, a new abstract theory for scenario decision-making was proposed and allows tackling a wider variety of convex problems, oft-constrained problems Garatti and Campi (2019), and also non-convex problems Campi et al. (2018). Rocchetta et al. (2020) applied scenario-based bound to derive bound support vector machine ensembles classifiers for anomaly detection and to tackle non-convex reliability-based design optimization problems, Rocchetta et al. (2020).

In this work, we investigate new generalization error bounds for regularized ELM models. We derive an upper and a lower bound on the probability of prediction errors being greater than a predefined threshold level, i.e., a certificate on the accuracy of the ELM model for future samples. The certificate is derived without any assumptions on the distribution family of the data and works for any number of samples. This makes the bounds particularly useful when a lack of data affects the study or when the distribution of the samples is highly uncertain. We test the new method on a reliability assessment problem of a dynamic controller affected by uncertainty. An ELM model is sequentially trained to emulate the reliability response of the controller and it is used to predict future values of a reliable performance function for the controller. The resulting ELM is used to efficiently assess the failure probability of the controller and the bounds on the accuracy give guarantees on the correct classification of failure and safe labels (up to the desired confidence). We will conclude with a discussion on the applicability of these bounds for reliability-based decision-making and online learning for safety-critical systems context.

2. Preliminaries

Consider a vector of explanatory variable $x \in \mathcal{X} \subseteq \mathbb{R}^{n_x}$ and vector of target variables $y \in \mathcal{Y} \subseteq \mathbb{R}^{n_y}$ from an unknown process $f : \mathcal{X} \to \mathcal{Y}$. We call $\delta = (x, y)$ a *scenario* and its dimension is $n_\delta = n_x + n_y$.
A data set

$$\mathcal{D}_N = \{\delta_i\}_{i=1}^N \in \Omega^N,$$

is available and it contains N independent and identical distributed (iid) scenarios (x, y). We are interest in the identification of a explanatory model $y = \hat{f}(x)$, obtained from $\mathcal{D}_N$, that well-describes the unknown/uncertain process $f : \mathcal{X} \to \mathcal{Y}$. The mechanism generating the data is known as a stationary Data-Generating Mechanism (DGM) and can be regarded as classical probability space $(\Omega, \mathfrak{F}, \mathbb{P})$, comprising the event

space $\Omega = \mathcal{X} \times \mathcal{Y}$, equipped with a σ-algebra $\mathfrak{F}$, and a probability measure $\mathbb{P}$.

3. Scenario decision-making theory

Consider a *scenario decision-making problem* defined as follows:

$$\mathcal{M} : \Omega^N \to \Theta,\ N = 0, 1, 2, ..., \qquad (1)$$

where $\mathcal{M}$ is a generic map from the event space for a multi-sample extraction of size N to a decision space Θ. An optimal decision made according to the rule $\mathcal{M}$ and the set $\mathcal{D}_N$ is given by

$$d^\star = \mathcal{M}_N(\mathcal{D}_N) \in \Theta.$$

Without loss of generality, $\mathcal{M}$ can be an expert-based or heuristic rule, an optimization algorithm or any generic given-data method, like regression or classification methods. Note that an ELM model is essentially a function $f_{ELM}(x, d^\star) : \mathcal{X} \to \mathcal{Y}$ which optimal parameters $d^\star$ can be directly inferred from $\mathcal{D}_N$. Hence, ELM training program, like many other ML methods, are a special class of decision-making problem $\mathcal{M}$. Scenario decision-making theory seek generalization error bounds for the solution $d^\star$, i.e., guarantees on the solution's ability to perform as expected for future scenarios. Even with this high level of generality, it is already possible to prescribe error bounds for $d^\star$. In the next sections, we will review important definitions and assumptions that are needed to derive these grantees.

3.1. *Feasibility, support element and violation probability*

Definition 3.1 (Feasibility set). *Let us define a feasibility set $\Theta_\delta \subseteq \Theta$ as the collection of decisions $d \in \Theta$ that satisfy given requirements in correspondence of a scenario δ.*

An optimal decision $d^\star$ is acceptable/feasible for $\mathcal{D}_N$ if it lay within the intersection of all the feasibility sets, $d^\star \in \bigcap_{i=1}^N \Theta_{\delta_i}$.

Definition 3.2 (Violation probability). *The error probability, also known as violation probability (or risk), is given by:*

$$V(d^\star) = \mathbb{P}\left[\delta \in \Omega : d^\star \notin \Theta_\delta\right]. \qquad (2)$$

this is the probability that, given a new random $\delta \in \Omega$, the optimized $d^\star$ will fail to comply with the requirements in δ.

Given a reliability parameter $\epsilon \in [0, 1]$, a solution for which $V(d^\star) \leq \epsilon$ is known as $1 - \epsilon$ reliable, i.e., d performs as expected for at least $100 \times (1 - \epsilon)$ percent of the scenarios.

Definition 3.3 (Set of support scenarios). *We further define a support scenario a δ in the data set $\mathcal{D}_N$ that, if removed, leads to different decision. The set of support scenarios, or support set, is a set $\mathcal{S} \subseteq \mathcal{D}_N$ such that $\mathcal{M}(\mathcal{D}_N) \neq \mathcal{M}(\mathcal{D}_N \setminus \delta_s)$ if δ_s is removed from $\mathcal{S}$.*

Consider a problem $\mathcal{M}$ defined by the following convex optimization program with randomized constraints:

$$\min_{\mathbf{d}\in\Theta} \{||d|| : g(d, \delta_i) \leq 0,\ i = 1, ..., N\}, \qquad (3)$$

where $g(d, \delta_i)$ is a convex function (a cost, a reliability requirement or a negative loss) enforcing N sample constraints on the problem. A feasibility set induced by one of the constraints in Eq.(3) is given by, $\Theta_\delta = \{d \in \Theta : g(d, \delta_i) \leq 0\}$, and the violation probability is simply, $V(d^\star) = \mathbb{P}\left[\delta \in \Omega : g(d, \delta_i) > 0\right]$.

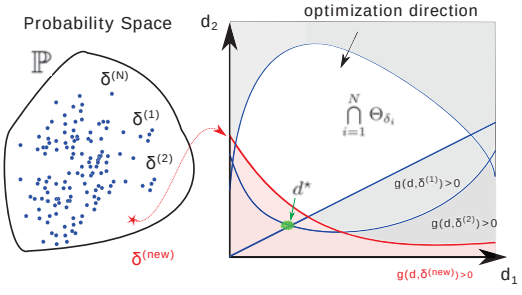

Fig. 1. An example of convex decision-making problem where the loss function g defines convex constraint on a two dimensional decision space.

Figure 1 presents a graphical example of convex scenario decision-making program. The scenarios depicted on the left panel induce N convex constraints, $g \leq 0$, that are depicted in the two-dimensional decision space in the panel on the right. The optimal design presented by a green marker is obtained by following the optimization direction (given by the partial derivative of $|d|$) and enforcing all the non-positive constraints. The optimized $d^\star$ lays within the union of feasibility sets (the non-feasible regions are in grey color). The violation probability is the probability that a new sample (in red), will lead to a failure of failure of the design $d^\star$.

The set $\mathcal{S}$ can be constructed by collecting the scenarios that active the constraints at the optimum[a], i.e., the samples for which $g(d^\star, \delta) = 0$. In scenario theory, the cardinality of the set of

[a]Note that this is a good practice for convex problem like

support elements, $s^\star = |\mathcal{S}|$ is a statistical indicator of the complexity of a decision and it will be used to compute generalization bounds on $V(d^\star)$. If $s^\star$ is computed a-posteriori, after observing $d^\star$, the resulting bounds are data-dependent and random quantity (because depend on the elements in $\mathcal{D}$). In contrast, data-independent (a-priori) bound can be derived by bounding the complexity of the optimization. As example, $s^\star$ is always caped by the dimension of decision variables for convex problems.

We are now ready to present the basic assumption needed on M to obtain generalization bounds on $d^\star$.

3.2. *Assumptions on $\mathcal{M}$*

We make the following assumptions on the proprieties of $\mathcal{M}$ [b]:
(A1) The problem $\mathcal{M}$ is permutation-invariant, i.e., $\mathcal{M}(\delta_1, ..., \delta_N) = \mathcal{M}(\delta_{k_1}, ..., \delta_{k_N})$ where $\delta_{k1}, ..., \delta_{kn}$ is a permutation of the N samples.
(A2) Consider two data sets $\mathcal{D}_n, \mathcal{D}_m$. For any non-negative integers n, m, if a decision $d^\star = \mathcal{M}(\mathcal{D}_n)$ taken using the first data set belongs to the joint feasibility set induced by the second data set with the new m sample, $\bigcap_{i=1}^{m} \Theta_{\delta_i}$, then $d^\star = \mathcal{M}(\mathcal{D}_n \cup \mathcal{D}_m) \in \bigcap_{i=1}^{m+n} \Theta_{\delta_i}$.
(A3) Instead, if $d^\star \notin \bigcap_{i=1}^{m} \Theta_{\delta_i}$ then the solution of the problem with the additional data set does not coincide with the initial solution, that is,

$$\mathcal{M}(\mathcal{D}_n \cup \mathcal{D}_m) \neq d^\star.$$

(A4) The problem $\mathcal{M}$ and its solution $d^\star$ are non-degenerate. Consider a subset of j samples discarded from $\mathcal{D}_N$ where $j < N$. If the j discarded samples are not support element (are not in $\mathcal{S}$) then, $\mathbb{P}[\mathcal{M}(\mathcal{D}_N) = \mathcal{M}(\mathcal{D}_{N-j})] = 1$. The solution $d^\star$ with all the samples in place coincides with probability one to the solution of $\mathcal{M}$ where only $N - j$ samples are used.

3.3. *Reliability bound for convex $\mathcal{M}$*

An expression for the bounds on R is formally introduced in (Garatti and Campi, 2019, Theorem 4).

Theorem 3.1. *Given a confidence parameter $\beta \in [0, 1]$, and under assumptions **A1-A4**, the risk of $d^\star$ can be bounded as follows:*

$$\mathbb{P}^N \left[\underline{\epsilon}(s^\star) \leq V(d^\star) \leq \overline{\epsilon}(s^\star)\right] \geq 1 - \beta$$

where $\underline{\epsilon}(s^\star)$ and $\overline{\epsilon}(s^\star)$ are a lower and an upper bound on the risk computed as functions of the complexity of the decision.

A formal expression $[\underline{\epsilon}(s^\star), \overline{\epsilon}(s^\star)]$ can be obtained solving the following polynomial equation in the t variable (for any number of samples $k = 0, 1, ..., n-1$),

$$\mathfrak{B}_n(t; k) = \frac{\beta}{2n} \sum_{j=k}^{n-1} \mathfrak{B}_j(t; k) + \frac{\beta}{6n} \sum_{j=n+1}^{4n} \mathfrak{B}_j(t; k)$$

where the factor $\mathfrak{B}_j(t; k) = \binom{j}{k} t^{j-k}$ is a binomial expansion. The upper and lower bounds on $V(z_n^\star)$ are given by $\underline{\epsilon}(k) = \max\{0, 1 - \overline{t}(k)\}$, and, $\overline{\epsilon}(k) = 1 - \underline{t}(k)$, where $\underline{t}(k) \leq \overline{t}(k)$ are the two unique solutions in $[0, +\infty[$ of the polynomial equation. For the special case $k = N$, the upper bound is set to $\overline{\epsilon}(k) = 1$ and the lower bound is obtained solving in t the following equation

$$1 = \frac{\beta}{6n} \sum_{j=n+1}^{4n} \mathfrak{B}_j(t; k).$$

4. Extreme Learning Machines

Extreme Learning Machines are very efficient feed-forward neural networks for which the input weights and parameters of the hidden nodes do not need to be tuned. ELMs are trained to accurately predict future values of y from observations of x. The output weights $d^\star$, defining the ELM, are optimize directly from $\mathcal{D}_N$. In the next sections, we present a classical training method to identify $d^\star$ a regularized version of it. The interested reader is reminded to Huang et al. (2014) for a comprehensive overview of some of the trends and challenges for ELMs.

4.1. *Traditional ELM*

AN ELM is given by a function,

$$f_{ELM}(x; d) = Hd = \sum_{i=1}^{n_h} h_i(x) d_i, \qquad (4)$$

where $d = [d_1, \cdots, d_{n_h}]^T \in \mathbb{R}^{n_h \times n_y}$ is the matrix of weights connecting n_h nodes in the hidden layer to n_y nodes in the output layer. The output matrix from the hidden layer is,

$$H = \begin{bmatrix} h(x_1) \\ \vdots \\ h(x_N) \end{bmatrix} = \begin{bmatrix} h_1(x_1) & \cdots & h_{n_h}(x_1) \\ \vdots & \ddots & \vdots \\ h_1(x_N) & \cdots & h_{n_h}(x_N) \end{bmatrix}, \qquad (5)$$

and $h : \mathcal{X} \to \mathcal{H}$ is an activation function, that is, a map from the physical input space $\mathcal{X}$ to a feature space $\mathcal{H}$. The i^{th} column of $\mathcal{H}$ contains

(3) because support scenarios are always active at the optimum. However, in a general non-convex setting non-active constraints can be support elements.

[b] Refer to Garatti and Campi (2019) for a detailed discussion on these assumptions

the N outputs of hidden node i. For simplicity sake, this work we will focus only on radial basis activation functions, although other options are available Huang et al. (2014). Note that selecting an appropriate $h(x)$ is a problem-dependent tasks and any nonlinear piece-wise continuous functions grantees an universal approximation capability for the ELM. Figure 2 presents a conceptual scheme of ELM model $f_{ELM}(x;d)$.

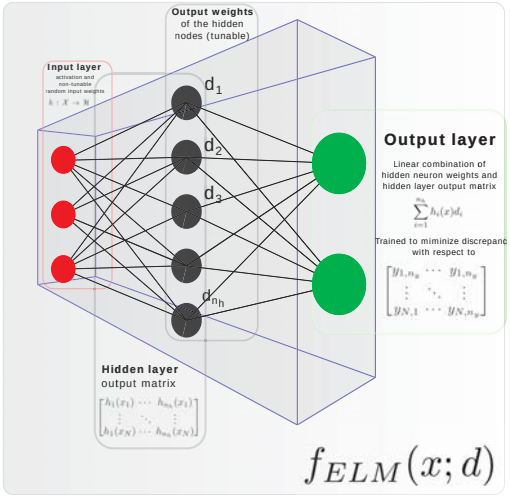

Fig. 2. Conceptual diagram of an extreme learning machine model.

ELM training program seek an optimal matrix of output weights d that minimizes a loss function, for instance, by minimization a loss between prediction and target in the least-square sense:

$$d^{\star} = \arg\min_{d} ||\hat{Y}(d) - Y||^2, \tag{6}$$

where Y is the matrix of output targets,

$$Y = \begin{bmatrix} y_1^T \\ \vdots \\ y_N^T \end{bmatrix} = \begin{bmatrix} y_{1,n_y} & \cdots & y_{1,n_y} \\ \vdots & \ddots & \vdots \\ y_{N,1} & \cdots & y_{N,n_y} \end{bmatrix}, \tag{7}$$

and $\hat{Y} = f_{ELM}(x;d) = Hd$ is the output predicted by the ELM and $||\cdot||$ is the Euclidean norm operator. The optimal solution to (7) can be obtained analytically and very efficiently by pseudo-inversion of the matrix of activation functions:

$$d^{\star} = H^{\dagger}Y, \tag{8}$$

where $H^{\dagger}$ is the Moor-Penrose generalized inverse of the hidden layer output matrix.

4.2. *Regularized ELM*

An alternative and popular method to optimize an ELM is by regularized version of thee least square method defined as follows:

$$\min_{d,\zeta} \frac{1}{2}||d||^2 + \frac{C}{2}\sum_{i=1}^{N}||\zeta_i||^2$$
$$s.t.\ |h(x_i)d - y_i| \leq \zeta_i\ i = 1, ..., N$$

where $C \in \mathbb{R}^+$ is a non-negative regularization parameter defining the cost of constraint violation, $h(x_i)d - y_i$ is a scenario constraint quantifying the discrepancy between the output $f_{ELM}(x_i;d)$ of the ELM model and the true value y_i. The vector $\zeta \in \mathbb{R}^{N,+}$ comprises N non-negative slack variables softening the sample constraints.
An equivalent formulation of (9) is given by the following unconstrained minimization program:

$$\min_{d} \frac{1}{2}||d||^2 + \frac{C}{2}||Hd - Y||^2, \tag{9}$$

which is known in the literature as the ridge regression (or regularized least squares) and admits a closed form solution for $n_h < N$ Huang et al. (2014):

$$d^{\star} = \left(H^T H + \frac{I_{n_h}}{C}\right)^{-1} H^T Y$$

and for $n_h > N$:

$$d^{\star} = H^T \left(HH^T + \frac{I_N}{C}\right)^{-1} Y$$

where I_{n_h} and I_N are the identity matrices of size n_h and N, respectively. Note that the term $H^T H$ is kernel matrix of an ELM and its elements $h(x_i) \cdot h(x_j)$ are the dot products of the activation functions.

4.3. *The proposed probabilistic error guarantees*

Regularized ELM training program are convex optimization method because the optimization parameters enter linearly in the constraint (quadratic in the objective function). Furthermore, we consider assumptions A1- A4 hold true for regularized ELM programs. Thus, theorem 3.1 can be used to prescribed the following certificate of generalization for the ELM predictions (with a confidence level $1 - \beta$):

$$\underline{\epsilon}(s_{\gamma}^{\star}, N, \beta) \leq V(d^{\star}) \leq \overline{\epsilon}(s_{\gamma}^{\star}, N, \beta) \tag{10}$$

The risk of error for the optimized ELM is given by:

$$V(d^{\star}) = \mathbb{P}[|h(x)d^{\star} - y| > \gamma],$$

and γ is an error level, i.e., a metric of discrepancy between the prediction and the true value of y, and $s^\star_\gamma$ is the number of support elements corresponding to the level γ. For the regularized ELM, the number $s^\star_\gamma$ is given by $s^\star_\gamma = \sum_{i=1}^{N} 1_{|h(x_i)d^\star - y_i| > \gamma}$, where $1_{|h(x_i)d^\star - y_i| > \gamma}$ is an indicator function for the condition $|h(x)d^\star - y| > \gamma$.

5. Numerical example on a reliability problem

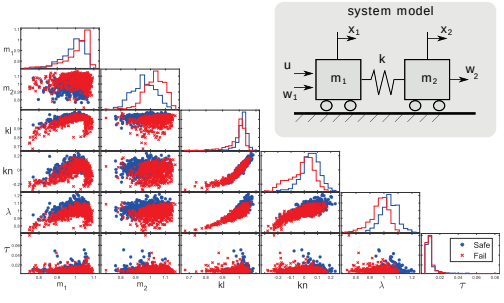

Fig. 3. The two-mass spring system and scatter of 10^3 samples of the six uncertain quantities labeled as failure (red cross markers) and safe (blue markers).

We test the proposed method on a modified version of the benchmark problem for robust control design, see Rocchetta et al. (2019). A data set contains 1000 samples of six uncertain input factors $x = [m_1, m_2, k_1, k_n, \lambda, \tau]$ where m_1 the mass of the first body, m_2 the mass of the second body, k_1 and k_n the linear and non linear spring constants, τ a time delay and λ an uncertainty factor on the loop-gain.

Each input vector x has three reliability performance scores $[g_1, g_2, g_3]$ associated to. The reliability performance scores model three reliability requirements (closed-loop stability, setting time and control effort). A $g_i(x) \geq 0$ means that the system fails to satisfy requirement i when the sample x is observed and a system failure occurs if at least one of the requirements is not satisfied. The worst-case reliability performance $w(x) = \max(g_1(x), g_2(x), g_3(x))$ offers a metric to quantify the reliability of the system and if $w(x) \geq 0$ the system fails at least one of the requirement. The probability of failure is thus given by,

$$P_f = \mathbb{P}[x \in \Omega : w(x) \geq 0].$$

Figure 3 presents the system structure and the available data set. Samples leading to a system failure are presented by red markers whilst the safe scenarios are displayed in blue colour. Note that computing $g_2(x)$ and $g_3(x)$ entails solving a set of state-space equations governing the system dynamics response in time. This is generally done via numerical integration and can be time-consuming. The goal of this work is to train an ELM capable of predicting values of w from samples of x. The resulting model $w = f_{ELM}(x; d^\star)$, will replace the numerically expensive state-space model of the two-mass spring system and it will be used to assess the reliability of the system and to assign a class to new scenarios. The optimized ELM will be equipped with a certificate of robustness that guarantees the surrogate model ability to predict new w accurately.

5.1. *Results*

As preliminary analysis we compare ELM models optimized by the Moor-Penrose pseudo inverse method or by using regularized ridge regression problem presented in Eq. (9). We use a subset of 500 samples from the data set, $n_h = 20$ hidden nodes and a parameter $C = 10^3$ weighting the cost of violations in the regularized regression problem. The top panel of Figure 4 displays the values for the optimized nodes weifhts $d^\star$ for the MP case (black dotted line) and the regularized case (red solid line). The bottom panel compares the goodness of fit of the two models, the 500 test values of w are displayed on the y-axes whilst the predicted values are presented on the x-axis.

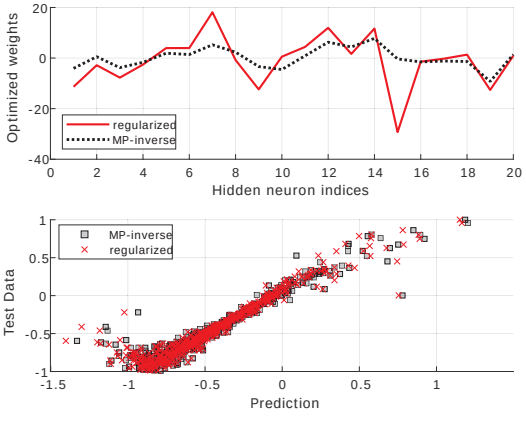

Fig. 4. The optimized weights $d^\star$ of a $n_h = 20$ ELM trained with MP and regularization method. In the bottom panel, a regression scatter plot of the test values for w and the predictions.

5.1.1. *Probabilistic error guarantees*

In this second analysis we present a numerical study the generalization of the proposed ELM

for reliability predictions. The generalization error bounds are computed using theorem 3.1. The following probabilistic certificate is obtained for the ELM $\mathbb{P}^N\left[V(d^\star) \in [\underline{\epsilon}, \overline{\epsilon}]\right] \geq 1-\beta$, where a high confidence level $1-\beta = 1-10^{-4}$ and the risk of inaccurate estimation is given by

$$V(d^\star) = \mathbb{P}[|f_{ELM}(x; d^\star) - w| > \gamma],$$

where γ measure the discrepancy between the prediction $\hat{w}$ and the round truth w. Figure 5 presents the generalization bounds $[\underline{\epsilon}, \overline{\epsilon}]$ for an increasing value of γ. The results refer to a regularized ELM trained with 20 hidden neurons and for a subset of $N = 500$ training samples from the available 10^3. The remaining samples are used to estimate the empirical probability of exceeding the discrepancy level γ. The true violation probability (risk) is presented by the red marked curve in the figure. WE can see that the bounds prescribed by scenario theory always hold and contain the estimate of the true violation probability.

For a $\gamma = 0$, the number of support elements is $s^\star_{\gamma=0} = 500$ and theorem 3.1 can be used to derive a generalization interval [0.96,1] bounding the risk of exceeding the given error margin. Clearly, these bounds are not very useful in practice because, i.e., not surprisingly the probability that the prediction discrepancy is non-null error ($\gamma \geq 0$) is quite high. In contrast, selecting a non-zero discrepancy $\gamma = 0.02$ the number of support elements result $s^\star_{0.02} = 90$ and the risk results bounded by $V(d^\star) \in [0.106, 0.272]$. In words, the probability that the lack of accuracy in a prediction of w from a new sample x will be less than 0.02 is at worst 72.8%.

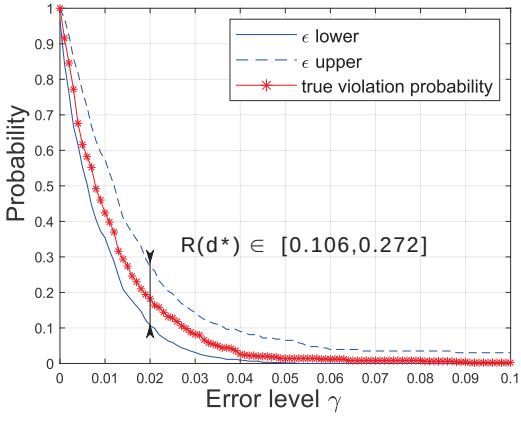

Fig. 5. Generalization error bounds $[\underline{\epsilon}, \overline{\epsilon}]$ for increasing discrepancy level γ. The red marked line presents the out-of-sample risk estimate.

5.1.2. *Online reliability assessment*

In this test example, we assume that only a very limited number of samples of x are initially available to optimize the ELM model and when a new observation is collected the model is re-trained using Eq. (9). At each iteration, the model is used to estimate the system failure probability. We monitor the improvement in generalization with scenario-based generalization bounds. Figure 6 presents the results of this analysis where only 5 samples were initially available and 300 were collected at the end of the procedure. The top panel presents an ELM-based estimator of the failure probability (red dashed line), the true failure probability (black constant line) and upper and lower predictions. The upper and lower bound on P_f were obtained, respectively, by evaluation of the probabilities $[\mathbb{P}[w > -\gamma]$ and $\mathbb{P}[w > \gamma]]$ where an accuracy level $\gamma = 0.01$ was selected. In other words, we consider a strip with half-width ± 0.01 around the limit state function $w = 0$ and we assume that a prediction falling within this strip can not be classified as failure or safe by the ELM. The bottom panel presents the probability bounds $[\underline{\epsilon}, \overline{\epsilon}]$ for the level of accuracy $\gamma = 0.01$. If only 5 samples are available to train the ELM, we can not guarantee anything on the prescribed level of accuracy $\gamma = 0.01$ because the generalization interval results vacuous $[\underline{\epsilon}, \overline{\epsilon}] = [0, 1]$. Conversely, when more data is collected the epistemic uncertainty in the ELM probabilistic accuracy substantially reduces, $[\underline{\epsilon}, \overline{\epsilon}]$ shrinks to approx [0.4,0.7], and the ELM-based estimate of the failure probability results quite accurate.

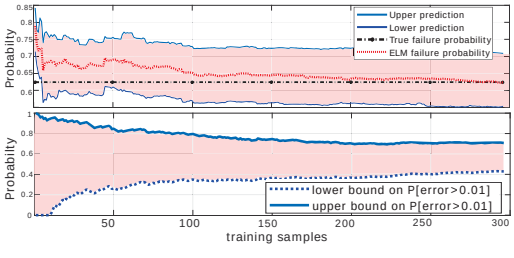

Fig. 6. Safe, failure and undecided domain according to the trained ELM and a predefined target $\gamma = 0.1$.

6. Conclusions

Scenario decision-making theory offers a formal mathematical framework to assess the robustness and generalization of data-driven decisions. In this work, scenario theory has been used to derive a probabilistic certificate of guarantees on the predictive accuracy of regularized extreme learning machine models, i.e., bounds on the probability

of exceeding a predefined error level. The only assumption on the data is on the samples being independent and identically distributed and this makes the resulting certificate distribution-free and non-asymptotic (for any number of samples). This makes scenario-based generalization bounds particularly useful when the databases are limited in size and/or severe uncertainty affects the distribution of the data. The proposed method has been tested on a data set obtained from the benchmark problem for robust control design. ELM models have been trained to predict the reliability of the system from a small data set of observations and sequentially re-trained when new data are collected The results prove the validity of the proposed ELM method, especially in a reliability context where high confidence in the prediction is needed.

Acknowledgement

This research was funded by the Digital Lifecycle Twins for predictive maintenance "ITEA3-2018-17030-DayTiMe" grant.

References

Campi, M. and S. Garatti (2008). The exact feasibility of randomized solutions of uncertain convex programs. *SIAM Journal on Optimization 19*(3), 1211–1230.

Campi, M. C. and S. Garatti (2018, Jan). Wait-and-judge scenario optimization. *Mathematical Programming 167*(1), 155–189.

Campi, M. C., S. Garatti, and F. A. Ramponi (2018, Dec). A general scenario theory for nonconvex optimization and decision making. *IEEE Transactions on Automatic Control 63*(12), 4067–4078.

Carè, A., S. Garatti, and M. C. Campi (2015). Scenario min-max optimization and the risk of empirical costs. *SIAM Journal on Optimization 25*(4), 2061–2080.

Efron, B. and C. Stein (1981). The Jackknife Estimate of Variance. *The Annals of Statistics 9*(3), 586 – 596.

Garatti, S. and M. C. Campi (2019, Nov). Risk and complexity in scenario optimization. *Mathematical Programming*.

Gautheron, L., A. Habrard, E. Morvant, and M. Sebban (2020). Metric learning from imbalanced data with generalization guarantees. *Pattern Recognition Letters 133*, 298–304.

Huang, G., G.-B. Huang, S. Song, and K. You (2014, 10). Trends in extreme learning machines: A review. *Neural Networks 61*.

Huang, G.-B., Q.-Y. Zhu, and C.-K. Siew (2006). Extreme learning machine: Theory and applications. *Neurocomputing 70*(1), 489–501. Neural Networks.

Isaksson, A., M. Wallman, H. Göransson, and M. Gustafsson (2008). Cross-validation and bootstrapping are unreliable in small sample classification. *Pattern Recognition Letters 29*(14), 1960–1965.

Oneto, L. (2018). Model selection and error estimation without the agonizing pain. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 8*(4), e1252.

Rocchetta, R., L. Bellani, M. Compare, E. Zio, and E. Patelli (2019). A reinforcement learning framework for optimal operation and maintenance of power grids. *Applied Energy 241*, 291–301.

Rocchetta, R., L. Crespo, and S. Kenny (2019). Solution of the benchmark control problem by scenario optimization. *Proceedings of the ASME Dynamic Systems and Control Conference, DSCC, October,*.

Rocchetta, R., L. G. Crespo, and S. P. Kenny (2020). A scenario optimization approach to reliability-based design. *Reliability Engineering & System Safety 196*, 106755.

Rocchetta, R., Q. Gao, and M. Petkovic (2020). Scenario-based generalization bound for anomaly detection support vector machine ensembles. In P. Baraldi, F. Di Maio, and E. Zio (Eds.), *Proceedings of the 30th European Safety and Reliability Conference and the 15th Probabilistic Safety Assessment and Management Conference*.

Shao, Z. and M. J. Er (2016). Efficient leave-one-out cross-validation-based regularized extreme learning machine. *Neurocomputing 194*, 260–270.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B (Methodological) 36*(2), 111–133.

Xu, Z. and J. H. Saleh (2021). Machine learning for reliability engineering and safety applications: Review of current status and future opportunities. *Reliability Engineering & System Safety 211*, 107530.