2017 IEEE 56th Annual Conference on Decision and Control (CDC)
December 12-15, 2017, Melbourne, Australia

Tuning regularization via scenario optimization

Simone Formentin, Simone Garatti, Marco C. Campi and Sergio M. Savaresi

Abstract— **In learning problems, avoiding to overfit the training data is of fundamental importance in order to achieve good predictive capabilities. Regularization networks have shown to be an effective tool to find reliable models, however their tuning is all but straightforward. In this paper, we consider learning problems that can be formulated as random convex minimization programs, and leverage on recent results established within the *Wait & Judge* theory for scenario optimization. Our main result is that, within this framework, generalization is deeply connected to the number of so-called support points found in optimization. By suitably selecting the regularization parameter, one can adjust the support points set and thereby can tune the trade-off between performance and generalization of the solution on the ground of a rigorous and quantitative theory.**

I. Introduction

Regularization was first introduced in [14] to numerically solve integral equations. After that, there has been a huge amount of work in statistics and machine learning dealing with regularization in a wide spectrum of problems, see, e.g., [9], [6], [10]. In the automatic control community, the interest in regularization has been recently renewed for linear system identification prompted by the novel perspective discussed in [12] and its follow-up works (see, e.g. [11], [5]). In these papers, the main idea is to see the identification of the impulse response of a system as an infinite-dimensional learning problem, instead of considering finite-dimensional parameterizations. In this way, a-priori information like stability can be taken into account and the contributions of bias and variance can be properly balanced.

From a Bayesian perspective, estimation aims at an optimal balance between empirical evidence, i.e., the data, and prior knowledge about the system, expressed in probabilistic terms. As compared to an estimation problem without prior, prior knowledge acts as a regularization term on the estimate. When one releases the assumption that a complete probabilistic description of prior knowledge is available, the tuning of regularization as a penalty term on the data-fitting becomes a non-trivial problem and this important topic has attracted a good deal of research. The conundrum is that there is no clear relationship between the parameter r weighting the regularization term and the ability of the solution to describe new situations. Therefore, cross-validation and extensive simulations are usually needed to properly set the value of r. However, such tools are affected by several drawbacks, [15], and, importantly, they need more data, which are expensive or not available in many application domains.

S. Formentin, S. Garatti and S.M. Savaresi are with the Department of Electronics, Information and Bioengineering, Politecnico di Milano, via G. Ponzio 34/5, 20133 Milano, Italy. M.C. Campi is with the Department of Information Engineering, University of Brescia, via Branze 38, 25123 Brescia, Italy. Email to: simone.formentin@polimi.it.

In this paper, we propose a change of perspective to provide a mathematical interpretation of the relationship between regularization and generalization. Specifically, we consider learning problems that can be formulated as random convex programs and, by using the most recent results of scenario optimization, i.e., the *Wait & Judge* scenario theory of [4], we show that generalization is deeply connected to the number of so-called support points (i.e., those points whose removal yields a change of the solution), while, in turn, the number of support points is regulated by the selection of the regularization parameter r. This result establishes a precise link between r and the generalization properties of the solution and delivers a fundamental insight to perform a suitable trade-off between generalization and performance.

We should remark here that regularized convex programs have already been considered from a scenario perspective in [2]. However, in that paper, the focus was on min-max optimization with a L_1-penalty term to enforce sparsity of the solution. Based on the classical result in scenario optimization that the probability of violation is related to the number of optimization variables, [1], [3], it is shown in [2] that, by suitably selecting the regularization weight which induces sparsity, the generalization properties of the model can be kept under control. Hence, it is crucial in this theory that the L_1-penalty shrinks the number of non-zero variables, thus effectively reducing the size of the optimization vector. Notice that such an approach *cannot* be used in more general regularization frameworks, like L_2 regularization that we shall concentrate on here, since regularization generally does not change the number of the optimization variables in this context, [13].

The key idea to assess the impact of regularization on generalization via the Wait & Judge scenario approach of [4] comes from the fact that such a theory allows us to compute the number of support points *a-posteriori*. Hence, a range of values for r can be tested and the generalization properties for each value of r can be evaluated based on the found number of support points. Other recent uses of scenario optimization for problem robustification can be found in [7], [8], but, to the best knowledge of the authors, this is the first time that scenario optimization is employed to solve inverse problems with L_2-regularization.

The remainder of the paper is as follows. In Section II, the problem is motivated and mathematically formulated. Sections III and IV present the main theoretical results, including an algorithm to set the value of r. A numerical example illustrates the approach in Section V. The paper is ended by some concluding remarks.

978-1-5090-2873-3/17/$31.00 ©2017 IEEE

II. Problem statement

Consider the learning problem

$$\begin{aligned} \min_{\theta\in\Theta\subseteq\mathbb{R}^d} \quad & J(\theta) \\ \text{subject to} \quad & g(\theta,\delta^{(i)})\leq 0,\ i=1,\dots,N, \end{aligned} \tag{LP}$$

where θ is the optimization vector containing the model parameters, $\Theta\subseteq\mathbb{R}^d$ is a convex set, $J(\theta)$ is a convex function of θ representing the data-fitting loss and $g(\theta,\delta^{(i)})$ is a convex constraint making the solution minimizing J consistent with the observed data $\delta^{(i)}$, $i=1,\dots,N$. Such data are assumed to be independent and identically distributed (i.i.d.) and drawn from an unknown probability $\mathbb{P}_\Delta$ over a set Δ. Notice that (LP) can be solved using standard convex optimization tools.

The aim of this paper is to discuss the effect - in terms of performance and generalization to the unseen data in Δ - of adding to (LP) a regularization term as follows:

$$\begin{aligned} \min_{\theta\in\Theta\subseteq\mathbb{R}} \quad & J(\theta)+r\left\|A\theta-b\right\|_2 \\ \text{subject to} \quad & g(\theta,\delta^{(i)})\leq 0,\ i=1,\dots,N, \end{aligned} \tag{rLP}$$

where A is a $p\times d$ matrix, b is a d-dimensional vector and r is the tunable regularization parameter.

Remark 1 (Prior Model): Commonly, the regularization term in (rLP) reads

$$r\left\|\theta-\bar\theta\right\|_2$$

(that is, A is the identity matrix of dimension d and $b=\bar\theta$), where $\bar\theta$ is a prior on the model parameters coming from preliminary knowledge about the system. For small r, the solution θ^*_{rN} of (rLP) is allowed to be far from $\bar\theta$ to better fit the data. Conversely, high r yields higher fitting error in favor of models in the neighborhood of the prior. ■

The objective of this paper is to provide a methodology to suitably select r in (rLP) such that the corresponding solution θ^*_{rN} satisfies certain user-defined requirements in terms of *both* performance and generalization. In fact, each solution θ^*_{rN} for a fixed value of r is assessed based on the attained value of the cost $J(\theta^*_{rN})$ (performance) and the capability of the solution to satisfy the constraint $g(\theta,\delta)\leq 0$ for instances of δ other than the observed sample $\delta^{(1)},\dots,\delta^{(N)}$. The latter can be more rigorously quantified by the *probability of incorrect description* $V(\theta^*_{rN})$, where $V(\theta)$ for any given feasible point $\theta\in\Theta$ is defined as follows:

$$V(\theta)=\mathbb{P}\left\{\delta\in\Delta : g(\theta,\delta)>0\right\}. \tag{1}$$

The different solutions θ^*_{rN} for varied values of the regularization parameter r attain a different pair of $J(\theta^*_{rN})$ and $V(\theta^*_{rN})$ and typically the two indexes have an opposite trend. In particular, the following Theorem can be stated.

Theorem 1: $J(\theta^*_{rN})$ is monotonically non-decreasing with the regularization parameter r.

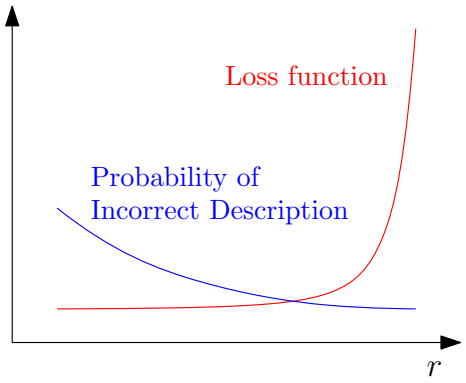

Fig. 1. Trade-off plot to set the regularization parameter r.

Proof: Take r_1 and r_2 as two values of the regularization parameter such that $r_1\leq r_2$. Given that $\theta^*_{r_1N}$ and $\theta^*_{r_2N}$ are the minimizers of the corresponding problems (rLP) with r_1 and r_2 as regularization parameters, it holds that

$$J(\theta^*_{r_1N})+r_1\left\|A\theta^*_{r_1N}-b\right\|_2\leq J(\theta^*_{r_2N})+r_1\left\|A\theta^*_{r_2N}-b\right\|_2,$$
$$J(\theta^*_{r_2N})+r_2\left\|A\theta^*_{r_2N}-b\right\|_2\leq J(\theta^*_{r_1N})+r_2\left\|A\theta^*_{r_1N}-b\right\|_2.$$

Then,

$$J(\theta^*_{r_1N})-J(\theta^*_{r_2N})\leq r_1\left(\left\|A\theta^*_{r_2N}-b\right\|_2-\left\|A\theta^*_{r_1N}-b\right\|_2\right),$$

but also

$$J(\theta^*_{r_1N})-J(\theta^*_{r_2N})\geq r_2\left(\left\|A\theta^*_{r_2N}-b\right\|_2-\left\|A\theta^*_{r_1N}-b\right\|_2\right).$$

The term $\left\|A\theta^*_{r_2N}-b\right\|_2-\left\|A\theta^*_{r_1N}-b\right\|_2$ cannot be positive, as otherwise the two above inequalities would yield a contradiction. Hence, it must be negative, leading to $J(\theta^*_{r_1N})-J(\theta^*_{r_2N})\leq 0$. It follows that, for $r_1\leq r_2$, $J(\theta^*_{r_1N})\leq J(\theta^*_{r_2N})$, which concludes the proof. ■

From the above result, it follows that, as already observed in Remark 1, the smaller r the better the performance $J(\theta^*_{rN})$, while it is an intuitive fact that the smaller r, the more the solution can be adapted to the seen observations and the worst the probability of incorrect description $V(\theta^*_{rN})$.

In order to properly select the regularization parameter r, $J(\theta^*_{rN})$ and $V(\theta^*_{rN})$ should be computed for a grid of values for r and then plotted each against the other so as to obtain a diagram like the one in the illustrative Figure 1. This diagram evidently gives the user all the relevant information to choose the most suitable value of the regularization parameter r for the problem at hand and the main result of this paper is that of offering an effective tool to construct such a diagram.

To this aim, however, a fundamental observation has to be made. Although $J(\theta^*_{rN})$ is revealed once the optimization problem is solved and hence is readily available to the user, on the other hand, $V(\theta^*_{rN})$ is a quantity which is not directly accessible since it depends on the probability with which observations take value. This probability is unknown to the user, who has just partial information of it through the observations $\delta^{(1)},\dots,\delta^{(N)}$. The key problem this paper addresses is the evaluation of $V(\theta^*_{rN})$ from the sole available information represented by $\delta^{(1)},\dots,\delta^{(N)}$.

III. Assessment of the probability of incorrect description

The following assumption is made throughout this section.

Assumption 1 (Existence and uniqueness): For any N and dataset $\delta^{(i)}$, $i = 1, \dots, N$, the problem (rLP) admits a solution, and such a solution is unique. ■

A. A first naive attempt

It has to be noted that the probability of incorrect description $V(\theta^*_{rN})$ is a random variable because it depends on the observations $\delta^{(1)}, \dots, \delta^{(N)}$. On the other hand, the distribution of $V(\theta^*_{rN})$ tends to be supported in intervals of the type $[0, \epsilon)$ as specified by the following theorem taken from [3].

Theorem 2: Fix $\beta \in (0, 1)$ and let ϵ be such that

$$\sum_{i=0}^{d-1} \binom{N}{i} \varepsilon^i (1-\epsilon)^{N-i} = \beta.$$

Then, it holds that

$$\mathbb{P}^N \left\{ V(\theta^*_{rN}) > \epsilon \right\} \leq \beta. \tag{2}$$

Proof: See [3]. ■

By selecting β as a number very close to 0, say e.g. 10^{-7}, the theorem says that $V(\theta^*_{rN})$ will be smaller than ϵ with very high confidence. This suggests that the probabilistic upper bound ϵ can be used as a safe estimate of $V(\theta^*_{rN})$.

The problem with this assessment of $V(\theta^*_{rN})$ is that it is typically not able to discern between different regularization levels. Indeed, as ϵ is determined as a function of β, d, N, which are the same to all problems (rLP), the same bound, adapted to the worst generalization level, is obtained irrespective of the regularization parameter r.

The main issue here is that this bound can be loose and therefore little informative for the problem at hand.

B. A Wait & Judge perspective

To introduce the new perspective that will make it possible to obtain tight evaluations of $V(\theta^*_N)$, we need to introduce the following definition.

Definition 1: Consider (rLP) for a given r. An observation $\delta^{(i)}$ is said to be a *support point* for (rLP) if its removal changes the solution to (rLP).[1] ■

In a sense, support points are those points that actively concur in the determination of θ^*_{rN}, and they can be easily found by applying the definition and solving N optimization problems in succession, where one point at a time is removed. The number of support points for a solution θ^*_{rN} is denoted by s^*_{rN}.

[1] In other words, $\delta^{(i)}$ is a support point if the solution to a problem like (rLP), where the constraint corresponding to $\delta^{(i)}$ is removed, is not equal to θ^*_{rN}.

It can be observed that the choice of r affects s^*_{rN} in a way that resembles how r affects $V(\theta^*_{rN})$. In fact, increasing r in (rLP) tends to yield a shrinkage of the number of support points s^*_{rN}. At the limit (see again Theorem 1), when $r \to \infty$, the regularization term obtains more and more importance to the detriment of the other element of the problem, and it is then clear that few $\delta^{(i)}$'s will be of support as the solution is mainly dictated by the regularization term irrespective of the $\delta^{(i)}$'s[2]. On the other hand, when $r \to 0$, problem (rLP) tends to (LP) where there is no regularization, and the solution is free to adapt as much as possible to the observations. In this case, s^*_{rN} will coincide with the usually big number of $\delta^{(i)}$'s determining the solution θ^*_N of the non-regularized problem (LP).

The intuition that a good evaluation of $V(\theta^*_N)$ can be obtained based on the assessment of s^*_{rN} is put on a solid ground by the recently introduced Wait & Judge Scenario approach of [4], where bounds to the probability of incorrect description adapted to s^*_{rN} are computed.

The main idea of [4] to be used here is based on the following assumption.

Assumption 2 (Non-degeneracy): For any N and with probability 1 with respect to the dataset $\delta^{(i)}$, $i = 1, \dots, N$, the solution to problem (rLP) corresponds to the solution obtained if the sole support points are in place. ■

This assumption is mild and normally follows if, e.g., data points take value according to a distribution with no-concentrated mass. See [4] for further discussion.

The result of [4] is as follows:

Theorem 3 (Wait & Judge Scenario Optimization): Given a confidence level $\beta \in (0, 1)$, it holds that

$$\mathbb{P}^N \left\{ V(\theta^*_{rN}, \delta) > \varepsilon(s^*_{rN}) \right\} \leq \beta, \tag{3}$$

where θ^*_{rN} is the minimizer of (rLP), s^*_{rN} is the number of support points of (rLP), $\varepsilon(s) = 1 - t(s)$ and $t(s)$ is the unique solution in $(0, 1)$ of the polynomial equation

$$\frac{\beta}{N+1} \sum_{m=s}^{N} \left(\begin{array}{c} m \\ s \end{array} \right) t^{m-s} - \left(\begin{array}{c} N \\ s \end{array} \right) t^{N-s} = 0.$$

Proof: See [4]. ■

As before, Theorem 3 says that the probability of incorrect description is below a given upper bound with very high confidence, and by selecting β to be very small, one can safely use the upper bound as an estimate of the probability of incorrect description. However, differently from before, the upper bound $\varepsilon(s^*_{rN})$ is revealed only *a posteriori*, as it depends on the actually seen number of support points and, as such, is representative of the situation at hand: different

[2] To observe this fact, take a simple example of (rLP) with the simple regularization term of Remark 1. At limit, for $r \to \infty$, the solution θ^*_{rN} tends to $\bar{\theta}$, provided this is a feasible value. Such a solution represents some prior knowledge on θ and, since it is independent of the measurements, the corresponding number of support points is zero.

values for the regularization parameter r lead to different solutions θ^*_{rN} and different probabilities of incorrect description $V(\theta^*_{rN})$; correspondingly, different numbers of support points s^*_{rN} and different evaluations of the probability of incorrect description as given by $\varepsilon(s^*_{rN})$ will be made. We should remark here that it has been shown in [4] that these evaluations are indeed tight.

IV. Tuning of the regularization

A. The algorithm

The results reported in Section III suggest the following algorithm, from now on referred to as Regularized random convex program (RRCP), to compute the tradeoff diagram discussed in Section II and to secure a desired level of performance and generalization through r.

Regularized random convex program (RRCP)
(inputs: N, β, A, b, d, $J(\cdot)$, r_{max}, M)

1) **[r-sampling]** Select M grid points $r^{(k)}$, $k = 1, \ldots, M$, of r within the interval $\mathcal{I}_r = [0, r_{max}]$.
2) **[Optimization]** Compute the solution $\theta^{*(k)}_{rN}$ of (rLP) corresponding to each $r^{(k)}$ and yielding the cost $J_k = J(\theta^{*(k)}_{rN})$; then, compute the number of support constraints $s^{*(k)}_{rN}$ corresponding to each $\theta^{*(k)}_{rN}$.
3) **[Assessment]** Assess the generalization level $\varepsilon_k = \varepsilon(s^{*(k)}_{rN})$ of each solution $\theta^{*(k)}_{rN}$, $k = 1, \ldots, M$ via Theorem 3. Such a generalization level corresponds to an upper bound on the probability of incorrect description $V(\theta^{*(k)}_{rN})$.
4) **[Selection]** Select the solution corresponding to the desired couple (J_k, ε_k).

B. An additional theoretical analysis

If the solution θ^*_{rN} is selected via the above algorithm, it should be noticed that several solutions (corresponding to M values of r) are evaluated before the choice is made. It follows that the confidence with which the probability of incorrect description of the chosen solution is below the computed threshold is slightly smaller than $1 - \beta$. The reason is that we have to secure that all the solutions corresponding to various grid points $r^{(k)}$, $k = 1, \ldots, M$, have simultaneously a probability of incorrect description smaller than ε_k. The following simple result formalizes this observation.

Theorem 4 (Confidence level of RRCP): Referring to the RRCP Algorithm, under the same assumptions of Theorem 3, it holds that

$$\mathbb{P}^N \left\{ \exists k : V(\theta^{*(k)}_{rN}, \delta) > \varepsilon_k \right\} \leq M\beta. \tag{4}$$

Proof:

$$\mathbb{P}^N \left\{ \exists k : V(\theta^{*(k)}_{rN}, \delta) > \varepsilon_k \right\} \leq \sum_{\kappa=1}^{M} \mathbb{P}^N \left\{ V(\theta^{*(\kappa)}_{rN}, \delta) > \varepsilon_k \right\} \leq M\beta, \tag{5}$$

where the last inequality follows because (3) holds for each k thanks to Theorem 3. ■

As a consequence of Theorem 4, the level of confidence, which was previously equal to $1 - \beta$, is now decreased to $1 - M\beta$. However, it has to be remarked that, from a practical point of view, this is not an issue, since M cannot be high due to computational issues and β is usually very small. For instance, if $\beta = 10^{-7}$, for $M = 100$ evaluations of r, the level of confidence for the bound to the probability of incorrect description would be 0.99999, which is still very close to 1.

V. A simulation example

In this section, we consider a learning example to numerically show the effectiveness of the proposed approach. We should remark that the proposed method can be applied to a regularized optimization problem of any dimension (e.g., any size of inputs and outputs, any number of data, etc.). However, a simple scalar learning problem is presented here to facilitate the visualization of the results.

A bi-variate phenomenon takes place in the x-y space, where the range for the x variable is $[0, 1]$. The goal is to find a curve that upper bounds the variability of y as a function of x. To this end, we have at our disposal $N = 1000$ samples (x_i, y_i), $i = 1, \ldots, 1000$. Moreover, an a-priori guess on the upper bound curve is also available. We set as objective to be minimized the area behind the upper bound curve while all seen points (x_i, y_i) remains below this curve.

Formally, we consider the model class

$$f_\theta(x) = b + \sum_{j=1}^{(d-1)/2} \left(\rho^{(j)} \sin(2\pi j x) + \eta^{(j)} \cos(2\pi j x) \right)$$

for the curve, where d is an odd number to adequately divide the degrees of freedom in sinusoids and co-sinusoids. In this example, we take $d = 299$. Hence, the set of optimization variables is

$$\theta = \left[b \ \rho^{(1)} \eta^{(1)} \ldots \ \rho^{(149)} \ \eta^{(149)} \right]^T$$

and the regularized estimation problem is written as

$$\min_{\theta \in \mathbb{R}^d} \int_0^1 f_\theta(x) dx + r \left\| \theta - \bar{\theta} \right\|_2 \tag{6}$$

$$\text{subject to} \quad y_i \leq f_\theta(x_i), \quad i = 1, \ldots, 1000, \tag{7}$$

where $\bar{\theta}$ denotes a prior estimate of θ given by a rough preliminary knowledge of the curve shape (e.g., obtained via previous experimental identification). Specifically, in the simulation, the domain for (x, y) is upper bounded by a damped oscillating function and $\bar{\theta}$ has been taken as

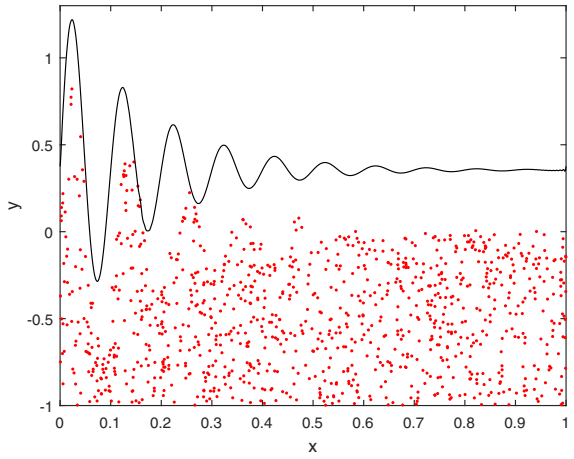

Fig. 2. Model output for $r = 10$.

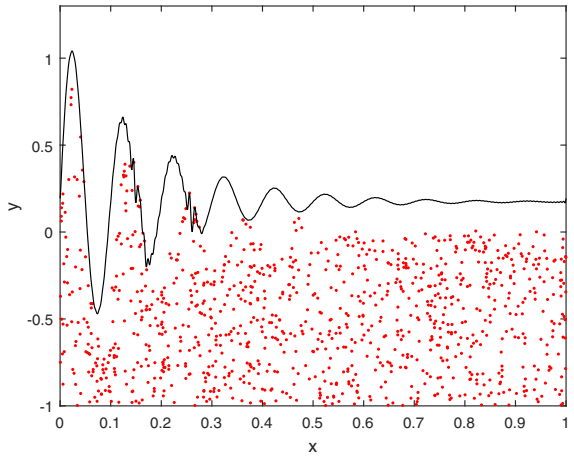

Fig. 3. Model output for $r = 4$.

an approximation to the real bound with an approximated frequency and damping factor and overestimated in hight. To run the RRCP algorithm for the problem at hand in (6), we took $M = 20$ values of r from 0.5 to 10 with step size 0.5. Thus, with $\beta = 10^{-6}$ we can apply the result of Theorem 4 to obtain a (satisfactory) level of confidence of $1 - M\beta = 0.99998$.

The models for 5 different values of r are given in Figures 2-6. While low values of r, e.g. $r = 0.5$, lead to an evident overfitting, a large weight on the regularization term, e.g. $r = 10$ disregards the information content of the data and makes the model too close to the prior choice.

To perform a choice of the best trade-off on a rigorous basis, we apply the scenario theory. Figures 7 and 8 display the loss function and the probability of incorrect description for the considered r. Based on these graphs, the choice was made to take the solution given by $r = 2.5$. For this value, the loss function becomes around 8% of the worst case, while the probability of incorrect description is 7.15%, which we assume is acceptable for the application at hand. The number

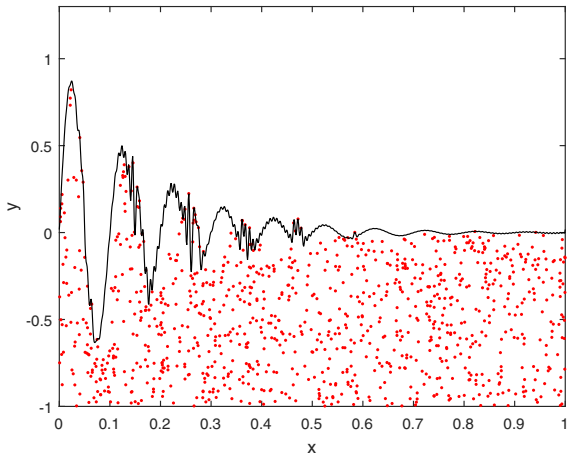

Fig. 4. Model output for $r = 2.5$.

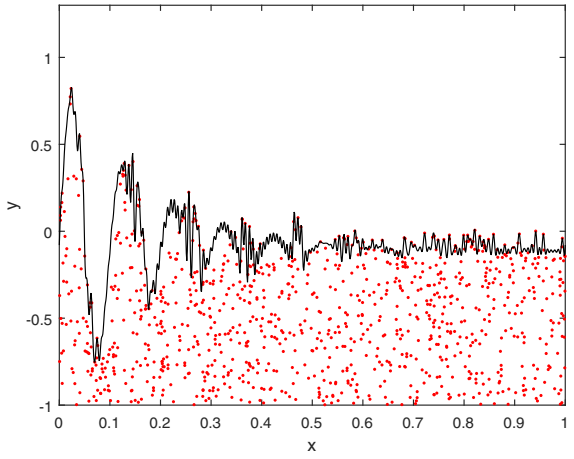

Fig. 5. Model output for $r = 1.5$.

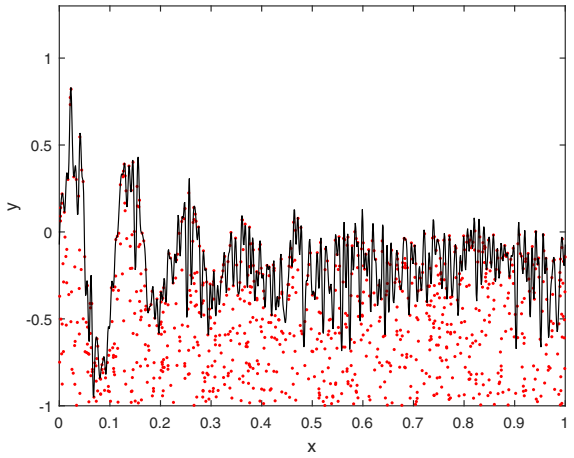

Fig. 6. Model output for $r = 0.5$.

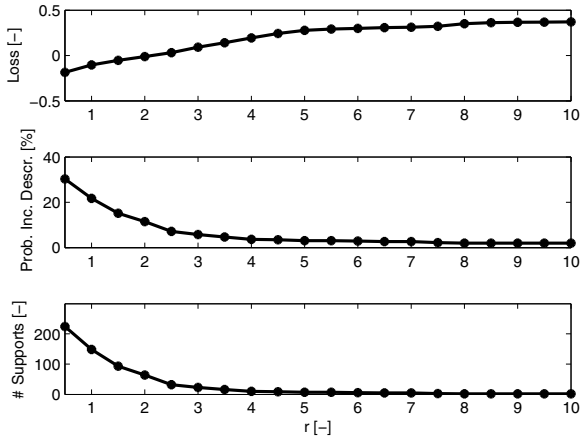

Fig. 7. Loss function and probability of incorrect description as functions of the regularization parameter r. The number of support constraints is also shown.

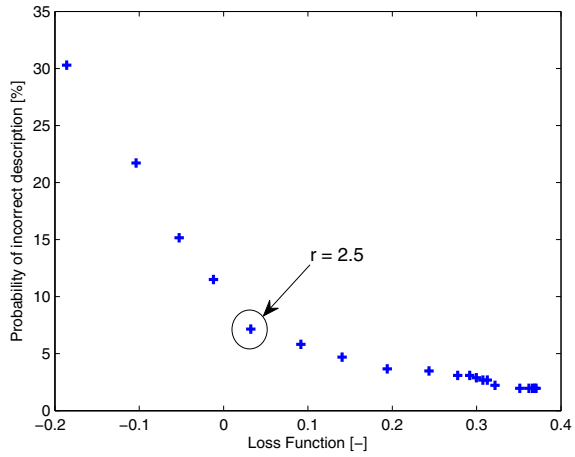

Fig. 8. Trade-off between loss function and probability of incorrect description. The solution corresponding to $r = 2.5$ turns out to be a good balance between the two performance indexes.

of support points for this choice was 32.

VI. Conclusions

In this paper, we used the Wait & Judge scenario theory of [4] to mathematically analyze learning problems to which a regularization term is added. The analysis showed that the regularization parameter r affects the number of support points, thus changing the probability of incorrect description of unseen data. Thanks to this link between the number of support points and the probability of incorrect description, we then provided an algorithm to select the value of r that is most suited for a given application, in terms of the trade-off between generalization and data-fitting.

Future research will be devoted to the optimal selection of the M values of r, as well as the application of the proposed approach to different learning problems.

References

[1] Giuseppe Calafiore and Marco C. Campi. Uncertain convex programs: randomized solutions and confidence levels. *Mathematical Programming*, 102(1):25–46, 2005.

[2] Marco C. Campi and Algo Caré. Random convex programs with L_1-regularization: sparsity and generalization. *SIAM Journal on Control and Optimization*, 51(5):3532–3557, 2013.

[3] Marco C. Campi and Simone Garatti. The exact feasibility of randomized solutions of uncertain convex programs. *SIAM Journal on Optimization*, 19(3):1211–1230, 2008.

[4] Marco C. Campi and Simone Garatti. Wait-and-judge scenario optimization. *Mathematical Programming*, Published online, DOI: 10.1007/s10107-016-1056-9, 2016.

[5] Alessandro Chiuso. Regularization and bayesian learning in dynamical systems: Past, present and future. *Annual Reviews in Control*, 41:24–38, 2016.

[6] Heinz Werner Engl, Martin Hanke, and Andreas Neubauer. *Regularization of inverse problems*, volume 375. Springer Science & Business Media, 1996.

[7] Simone Formentin, Fabrizio Dabbene, Roberto Tempo, Luca Zaccarian, and Sergio M Savaresi. Robust linear static anti-windup with probabilistic certificates. *IEEE Transactions on Automatic Control*, 62(4):1575–1589, 2017.

[8] Simone Formentin, Gianmarco Rallo, Simone Garatti, and Sergio M. Savaresi. Robust direct data-driven controller tuning with an application to vehicle stability control. *International Journal of Robust and Nonlinear Control*, Published online, DOI: 10.1002/rnc.3782, 2017.

[9] Federico Girosi, Michael Jones, and Tomaso Poggio. Regularization theory and neural networks architectures. *Neural computation*, 7(2):219–269, 1995.

[10] Patricia K Lamm. A survey of regularization methods for first-kind volterra equations. In *Surveys on solution methods for inverse problems*, pages 53–82. Springer, 2000.

[11] Gianluigi Pillonetto, Alessandro Chiuso, and Giuseppe De Nicolao. Prediction error identification of linear systems: a nonparametric gaussian regression approach. *Automatica*, 47(2):291–305, 2011.

[12] Gianluigi Pillonetto and Giuseppe De Nicolao. A new kernel-based approach for linear system identification. *Automatica*, 46(1):81–93, 2010.

[13] Suvrit Sra, Sebastian Nowozin, and Stephen J. Wright. *Optimization for machine learning*. Mit Press, 2012.

[14] Andrey Nikolayevich Tikhonov. On the stability of inverse problems. In *Dokl. Akad. Nauk SSSR*, volume 39, pages 195–198, 1943.

[15] Sudhir Varma and Richard Simon. Bias in error estimation when using cross-validation for model selection. *BMC bioinformatics*, 7(1):91, 2006.