UNCERTAINTY BOUNDS FOR KERNEL-BASED REGRESSION: A BAYESIAN SPS APPROACH

Algo Carè[*] *Gianluigi Pillonetto*[†] *Marco C. Campi*[*]

University of Brescia, Italy
University of Padova, Italy
University of Brescia, Italy

ABSTRACT

This paper shows that kernel-based estimates of unknown input-output maps can be complemented with uncertainty bounds more robust than those commonly derived in the Gaussian regression framework. This is obtained by using the kernel not to define Gaussian priors but a much vaster class of symmetric distributions. Such class is then handled by extending to the Bayesian setting the recently developed *sign-perturbed sums* (SPS) framework.

1. INTRODUCTION

The problem of reconstructing an unknown map from a finite set of input-output examples is central in machine learning [1]. Kernel-based methods have been widely studied and used to solve this task [2]. Such approaches allow to cast in a unified framework many different techniques, like regularization networks [3], and support vector machines [4].

In the analysis of such estimators, a crucial point is to assess their ability of predicting future data. In recent years, many new results have been obtained in regression problems and regularization networks (which involve quadratic losses and penalties). Non-asymptotic error bounds and learning rates can be found e.g. in [5]. Even if of great theoretical interest, in real applications such bounds can be however of limited usefulness since they depend on (the norm of) the unknown function and can also turn out to be somewhat conservative. An alternative route is to exploit the Bayesian interpretation of regularization networks by considering the Gaussian regression framework [6]. In fact, the connection with regularized least squares (ReLS) estimators is obtained by modeling the function and the measurement noise as (independent) Gaussian processes. The posterior becomes so available in closed form and Bayes intervals on the function estimate can be easily extracted [7].

The Gaussian regression approach can however fail in returning reliable bounds in important situations. Gaussian distributions cannot well describe outliers possibly coming from the random sources associated to the function and/or the measurement noise [8]. Difficulties can arise also when the kernel scale factor (regularization parameter) introduced in the model is not well tuned. In this paper, we face these problems in a finite-dimensional linear regression context and show that the ReLS estimate can be complemented with confidence intervals more robust than those commonly adopted. Our work can be cast within the framework of robust Bayesian analysis, see e.g. [9, 10]. We propose to replace the Gaussian prior on the unknown model parameters with a much more general class of distributions, i.e., the class of the distributions that can be generated by a certain linear transformation of symmetrically distributed random variables. We describe a procedure to construct Bayesian confidence regions having a desired and exact probability level for the above mentioned class of prior distributions. Traditionally, Bayesian credible regions are obtained by conditioning on data, that is, from the posterior probability. Instead, our construction is based on a different line of reasoning and aims at guaranteeing that the constructed region includes the true parameter with an exact prior probability level. In other words, we guarantee the algorithm for constructing a region, not the algorithm's outcome for a given output data vector, and our main result gives the probability with which the unknown parameter vector belongs to the region in many repetitions of the experiment. Quoting from [11] "*in certain statistical scenarios a joint frequentist-Bayesian approach is arguably required*" and this is the case in our paper.

From the technical point of view, the main result of the paper is obtained by extending to the Bayesian setting the recently developed sign-perturbed sums (SPS) framework [12, 13, 14]. Numerical experiments show that in many significant circumstances our uncertainty bounds can be much more reliable than those achieved by the Gaussian regression framework.

The paper is organized as follows. Section 2 formulates the problem. After reviewing Gaussian regression in Section 3, we present in Section 4 the new Bayesian SPS approach to recover uncertainty bounds for regularization networks. In Section 5 the new approach is tested using the numerical experiments

[*] A.C. and M.C.C. were partly supported by the H&W program of the University of Brescia under the project CLAFITE.

[†] G.P. has been partially supported by the MIUR FIRB project RBFR12M3AC-Learning meets time: a new computational approach to learning in dynamic systems and by the Progetto di Ateneo CPDA147754/14-New statistical learning approach for multi-agents adaptive estimation and coverage control.

978-1-5386-5477-4/18/$31.00 ©2018 IEEE

introduced in Section 3 as motivating examples. Conclusions then end the paper.

2. PROBLEM STATEMENT

We consider a linear regression problem. Output data are contained in the vector $y \in \mathbb{R}^n$ and the measurements model is

$$y = \Phi\theta^0 + v \tag{1}$$

where Φ is a known (full rank) regression matrix, $\theta^0 \in \mathbb{R}^m$ is the unknown vector while v contains the noise components. We consider a Bayesian setting where both θ^0 and v are (independent) random vectors. Consider a function $\Theta(\cdot)$ that assigns a set $\Theta(y) \subseteq \mathbb{R}^m$ to measurements y. For $\alpha \in [0,1]$, $\Theta(\cdot)$ is said to be of α-level if

$$\mathbb{P}(\theta^0 \in \Theta(y)) = \alpha,$$

where $\mathbb{P}(\theta^0 \in \Theta(y))$ denotes the probability of the event "$\theta^0 \in \Theta(y)$" computed with respect to the probability distribution over the noise realizations ν and the parameter θ^0. Our problem is then to find a function $\Theta(\cdot)$ of α-level to construct regions $\Theta(y)$ that are accurate set estimates of θ^0 for the measurements y. Abusing terminology, if there is no ambiguity, we will not distinguish between the region, $\Theta(y)$, and the procedure to construct the region (which, so far, has been denoted by $\Theta(\cdot)$); for example, we will write "a region $\Theta(y)$ of α-level" instead of "a procedure $\Theta(\cdot)$ of α-level".

3. GAUSSIAN REGRESSION

If the joint probability of y and θ^0 is specified by means of the density $\mathbf{p}(y,\theta^0)$, a region $\Theta(y)$ of α-level can be obtained by exploiting the Bayes rule. For any possible realization of the measurements y, one can compute the posterior

$$\mathbf{p}(\theta^0|y) = \frac{\mathbf{p}(y|\theta^0)\mathbf{p}(\theta^0)}{\mathbf{p}(y)}$$

and then extract from it a set of probability α for the distribution given by $\mathbf{p}(\theta^0|y)$. This procedure defines $\Theta(y)$ of α-level. In fact, letting $\mathbf{1}(\cdot)$ be the indicator function (equal to 1 when the formula in the argument is true, 0 otherwise), it holds that $\mathbb{P}(\theta^0 \in \Theta(y)) = \int_{\mathbb{R}^n \times \mathbb{R}^m} \mathbf{1}(\theta^0 \in \Theta(y))\mathbf{p}(y,\theta^0) = \int_{\mathbb{R}^n} [\int_{\mathbb{R}^m} \mathbf{1}(\theta^0 \in \Theta(y))\mathbf{p}(\theta^0|y)]\mathbf{p}(y) = \int_{\mathbb{R}^m} \alpha\mathbf{p}(y) = \alpha$.
The building of the Bayesian confidence region is especially simple under Gaussian assumptions since the posterior becomes available in closed form. Let θ^0 and v be independent normal vectors, i.e.

$$\theta^0 \sim \mathcal{N}(\mu, \lambda^2\Sigma), \quad v \sim \mathcal{N}(0, \sigma^2 I_n). \tag{2}$$

with λ^2 and σ^2 positive scale factors. Hence, the mean μ and covariance $\lambda^2\Sigma$ model our expected properties of θ^0. The a posteriori density of θ^0 given y is then

$$\theta^0|y \sim \mathcal{N}\left(\hat{\theta}^{ReLS}, (\frac{\Phi^T\Phi}{\sigma^2} + \frac{\Sigma^{-1}}{\lambda^2})^{-1}\right), \tag{3}$$

where $\hat{\theta}^{ReLS}$ is the posterior mean. Note that we have used the superscript ReLS to stress that $\hat{\theta}^{ReLS}$ is the solution of a regularized least squares problem. In fact, the minimum variance estimator is given by

$$\begin{aligned} \hat{\theta}^{ReLS} &= \arg\min_{\theta} \ \|y - \Phi\theta\|^2 + \eta^2(\theta-\mu)^T\Sigma^{-1}(\theta-\mu) && \text{(4a)} \\ &= \mu + (\Phi^T\Phi + \eta^2\Sigma^{-1})^{-1}\Phi^T(y - \Phi\mu). && \text{(4b)} \end{aligned}$$

In (4), the scalar $\eta^2 = \sigma^2/\lambda^2$ is the regularization parameter which trades off the measurements fit and the prior information on θ^0. Finally, using the factorization $\Sigma = \Sigma^{1/2}(\Sigma^{1/2})^T$, we define

$$\Omega = \begin{pmatrix} \Phi \\ \eta\Sigma^{-1/2} \end{pmatrix}.$$

Then, from (3) the following Bayesian confidence region can be obtained:

$$\Theta(y) = \left\{\theta \ : \ (\theta - \hat{\theta}^{ReLS})^T\Omega^T\Omega(\theta - \hat{\theta}^{ReLS}) \leq \kappa\sigma^2\right\}, \tag{5}$$

where κ determines α, e.g. see subsection 3.B in [12].

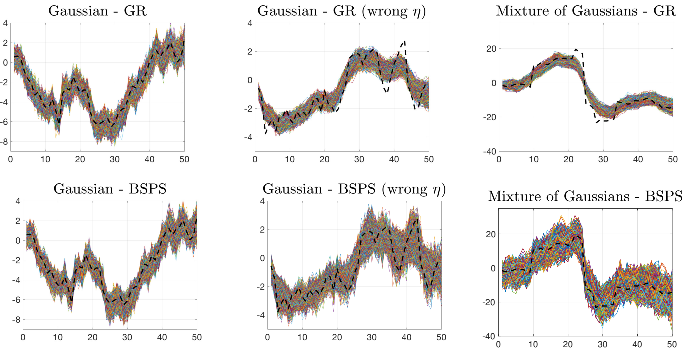

Fig. 1. True function (thick dashed line) and 95% confidence intervals returned by Gaussian regression (GR, top panels), and Bayesian SPS (BSPS, bottom panels).

3.1. Gaussian uncertainty bounds

We test the robustness of the Bayesian confidence region (5), via three simple case studies. In all of them, $\Phi\theta^0$ in (1) represents the discrete convolution between θ^0 and realizations of unit variance white Gaussian noise (independent of θ^0 and v). The random vector θ^0 has dimension 50 and is the output of a discrete integrator fed with white noise ω, i.e. its i-th component is $\theta_i^0 = \sum_{k=1}^{i} \ \omega_k$. The probability density function (pdf) of ω depends on the case study and will be specified later. The noise v is instead always zero-mean Gaussian with a known variance equal to 10 in the first two case studies and to 100 in the third. Data set size (dimension of y) is 100.

First case study: Gaussian pdf Let the components of ω be independent and Gaussian of unit variance, so that θ^0

corresponds to a Gaussian random walk. This is a popular model underlying the Bayesian interpretation of smoothing splines [15]. A realization of θ^0 is displayed in the top left panel of Fig. 1. The Bayesian region (5) with $\alpha = 0.95$ is built using the correct statistics of θ^0, i.e. plugging in (2) the values $\mu = 0$, $\lambda = 1$ and setting the (i, j) entry of Σ to $\min(i, j)$. In this case, the (realization of the) 95% Bayesian region contains (the realization of) θ^0. The uncertainty bound in sampled form is also displayed in the same panel. It is obtained by drawing 2000 independent and uniform realizations of θ^0 from the ellipsoid (5) via the algorithm described in [16].

Second case study: underestimated variance The same data generator described before is used but some misspecification is introduced in the construction of the Bayesian region (5): we now use $\lambda = 1/3$ (in place of $\lambda = 1$). This simulates a situation where the random walk increments variance is underestimated (too large η). Results are displayed in the second top panel of Fig. 1 with the same rationale described above. Now, the 95% confidence interval does not contain the realization of θ^0 and the uncertainty bounds turn out to be too optimistic.

Third case study: mixture of Gaussians In the last case study, ω is a mixture of Gaussians. More specifically, its components are mutually independent with pdf

$$\omega_i \sim \begin{cases} \mathcal{N}(0,1), & \text{with probability } 0.9 \\ \mathcal{N}(0,100), & \text{with probability } 0.1. \end{cases}$$

The 95% Bayesian region is instead built as if $\omega_i \sim \mathcal{N}(0,1)$, i.e. by using in (2) the values $\mu = 0$, $\lambda = 1$ and $\Sigma(i, j) = \min(i, j)$. Hence, with probability 0.1 a rapid and unexpected change between two adjacent components of θ^0 can be present. This is illustrated in the third top panel of Fig. 1. The rapid variation in θ, in the middle of the x-axis, is not contained in the 95% Gaussian confidence region.

4. BAYESIAN SPS

The assumption $\theta^0 \sim \mathcal{N}(\mu, \lambda^2\Sigma)$ is equivalent to the assumption that the unknown vector θ^0 is the sum of μ and the output of a linear system fed with white, stationary and Gaussian noise. In fact, from the factorization $\Sigma = \Sigma^{1/2}(\Sigma^{1/2})^T$ we have

$$\theta^0 = \mu + \Sigma^{1/2}\omega, \tag{6}$$

with $\omega \sim \mathcal{N}(0, \lambda^2 I)$. This is graphically depicted in the top panel of Fig. 2 where $\mu = 0$ to simplify the exposition.

The main idea is now to replace the Gaussian source underlying ω with a much more general one. In particular, we assume that the independent components of ω have distributions which can differ from each other and just need to be symmetric around the origin, i.e., the stochastic source of ω can be nonstationary, as described in the bottom panel of Fig. 2. We also relax the assumptions on the noise v in the same way, by moving from Gaussian to symmetric (possibly non-identical) pdfs. For instance, note that all the three case studies of the previous section, where the linear operator $\Sigma^{1/2}$ was the discrete integrator, fit these mild assumptions.

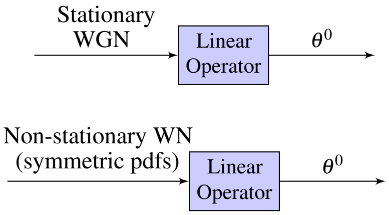

Fig. 2. Gaussian regression (top) and Bayesian SPS (bottom)

We now introduce Algorithms 1 and 2, which allows to construct α-level regions under the above defined mild symmetry properties.

Algorithm 1 Bayesian SPS-initialization

```
1: Factorize the kernel matrix $\Sigma$, by computing $\Sigma^{1/2}$ such that
       $\Sigma = \Sigma^{1/2}(\Sigma^{1/2})^T.$
2: Define $\Omega \in \mathbb{R}^{N \times m}$ as follows
       $\Omega = \begin{pmatrix} \Phi \\ \tilde{\Phi} \end{pmatrix}, \quad \tilde{\Phi} := \eta\Sigma^{-1/2}.$
   Let also $z \in \mathbb{R}^N$ be given by
       $z = \begin{pmatrix} y \\ \tilde{y} \end{pmatrix}, \quad \tilde{y} := \eta\Sigma^{-1/2}\mu.$
3: Define a (rational) confidence probability $\alpha \in (0,1)$ and set integers
   $r > q > 0$ such that $\alpha = 1 - q/r$;
4: Calculate $R_N$ and $R_N^{1/2}$ where
       $R_N = \frac{\Omega^T\Omega}{N}, \quad R_N^{1/2}(R_N^{1/2})^T = R_N;$
5: Generate $N(r-1)$ i.i.d. random signs $\{\alpha_{i,t}\}$ with
       $\mathbb{P}(\alpha_{i,t} = 1) = \mathbb{P}(\alpha_{i,t} = -1) = 1/2,$
   for $i = 1, \dots r-1$ and $t = 1, \dots, N$;
6: Generate a random perturbation $\pi$ of the set $\{0, 1, \dots, r-1\}$, where each
   of the $r!$ possible perturbations has the same probability to be selected.
```

Algorithm 1 is an initialization procedure. The inputs are the regression matrix Φ, the measurements y, the mean μ, the kernel matrix Σ, the regularization parameter η and the desired confidence probability $\alpha \in (0,1)$ defined by the integers $r > q > 0$. Algorithm 1 then returns the matrix Ω and the vector z, which are instrumental to the use of Algorithm 2 whose inputs also include candidate values of the unknown vector. Algorithm 2 then accepts or refuse any proposed value of θ, and this implicitly defines a region $\Theta(y)$: for any possible realization of y, Θ maps y in the subset of $\mathbb{R}^m$ formed by all the accepted values of θ. Algorithm 2 extends to the Bayesian setting the SPS approach discussed in [12], and can be explained as follows. A reference function $S_0(\theta)$ is defined in step 2, and it can be easily verified that $2R_N^{1/2}S_0(\theta)$ is the gradient of $\|y - \Phi\theta\|^2 + \eta^2(\theta-\mu)^T\Sigma^{-1}(\theta-\mu)$, that is, of the cost function that is minimized in (4). Thus, by construction $\|S_0(\theta)\|$ tends to grow as θ departs from θ^0. Other functions

Algorithm 2 Bayesian SPS-indicator(θ) given a matrix $\Omega \in \mathbb{R}^{N\times m}$ and a vector $z \in \mathbb{R}^N$

```
1: For the given $\theta$, compute the prediction errors
       $$\varepsilon_t(\theta) = z_t - \Omega(t,:)\theta, \quad t = 1,\dots,N$$
   where $\Omega(t,:)$ is the $t$-th row of $\Omega$;
2: Evaluate for $i = 1,2,\dots,r-1$
       $$S_0(\theta) = R_N^{-1/2}\frac{1}{N}\sum_{t=1}^{N} \Omega(t,:)^T\varepsilon_t(\theta)$$
   and
       $$S_i(\theta) = R_N^{-1/2}\frac{1}{N}\sum_{t=1}^{N} \alpha_{i,t}\Omega(t,:)^T\varepsilon_t(\theta);$$
3: Order the scalars $\{\|S_i(\theta)\|\}$ in increasing order. If $\|S_a(\theta)\| = \|S_b(\theta)\|$,
   $\|S_a(\theta)\|$ precedes $\|S_b(\theta)\|$ iff $\pi(a) < \pi(b)$;
4: Compute the rank $\mathscr{R}(\theta)$ of $\|S_0(\theta)\|$ in the ordering, e.g. $\mathscr{R}(\theta) = 1$ if
   $\|S_0(\theta)\|$ is the smallest one;
5: Return "accept" if $\mathscr{R}(\theta) \le r-q$.
```

$S_1(\theta),\dots,S_{r-1}(\theta)$ are constructed by applying random sign perturbations to the terms that define $S_0(\theta)$. On the one hand, these sign perturbations tame the growth of $\|S_i(\theta)\|$ as θ departs from θ^0, so that, the more θ and θ^0 differ, the larger is $\|S_0(\theta)\|$ with respect to any other $\|S_i(\theta)\|$. On the other hand, these sign perturbations affect the random variables ν and ω in such a way that $S_0(\theta)$ is statistically indistinguishable from $S_i(\theta)$ when $\theta = \theta^0$ (more details are provided in the proof of Theorem 1 below). Algorithm 2, in step 5, refuses a given value of θ when $\|S_0(\theta)\|$ is sufficiently large as compared with $\|S_i(\theta)\|$ $i = 1,\dots,r-1$, an event that has low probability when θ is the correct parameter. Indeed, the following result holds true.

Theorem 1 *Consider model* (1) *and assume that*

- *the components of the noise v are independent random variables with a symmetric probability distribution around zero;*
- *$\theta^0 = \mu + \Sigma^{1/2}\omega$ where the components of ω are independent random variables with a symmetric probability distribution around zero.*

Then, independently of the adopted regularization parameter η, the Bayesian region defined by Algorithm 2 satisfies

$$\mathbb{P}(\theta^0 \in \Theta(y)) = \alpha \;\; \text{with} \;\; \alpha = 1 - q/r \tag{7}$$

(where the probability $\mathbb{P}$ is over ω, ν and the randomly generated quantities $\{\alpha_{i,t}\}$ and π in Algorithm 1). Moreover, for any y, the region $\Theta(y)$ is star convex with the estimate $\hat{\theta}^{ReLS}$, (4)*, as a star center, that is,*

$$\forall\theta \in \Theta(y)\,\forall\beta \in [0,1] : \beta\theta + (1-\beta)\hat{\theta}^{ReLS} \in \Theta(y). \tag{8}$$

Proof. Referring to Algorithm 2, point 5, equality (7) is true iff

$$\mathbb{P}(\mathscr{R}(\theta^0) > r-q) = q/r. \tag{9}$$

In what follows, we prove (9) by showing that the rank $\mathscr{R}(\theta^0)$ of $\|S_0(\theta^0)\|$ is uniformly distributed over $\{1,\dots,r\}$. Note that $\mathscr{R}(\theta^0)$ can be written as $\mathscr{R}(\theta^0) = f_{\mathscr{R}}(\|S_0(\theta^0)\|, \|S_1(\theta^0)\|,\dots,\|S_{r-1}(\theta^0)\|,\pi)$, where $f_{\mathscr{R}}(\cdot)$ is the deterministic function of $\|S_0(\theta^0)\|,\dots,\|S_{r-1}(\theta^0)\|$ and of the random permutation π that is defined by points 3 and 4 of Algorithm 2.

Using the fact that $y_t - \Phi(t,:)\theta^0 = v_t$, see (1), and that $\omega_k = \Sigma^{-1/2}(k,:)(\theta^0 - \mu)$, see (6), we can write

$$S_0(\theta^0) = R_N^{-1/2}\frac{1}{N}\left[\sum_{t=1}^{n}\Phi(t,:)^T v_t + \sum_{k=1}^{m}\eta^2\Sigma^{-1/2}(k,:)^T\omega_k\right]. \tag{10}$$

Note that $S_0(\theta^0)$ is a random variable through its dependence on v and ω. Defining $\mathbf{r} := (v_1,\dots,v_n,\omega_1,\dots,\omega_m)$, we can then define $Z(\mathbf{r}) := \|S_0(\theta^0)\|$, that is, $Z(\cdot)$ is the deterministic function that, given the values of $\mathbf{r}$, computes $\|S_0(\theta^0)\|$ according to (10). For $i = 1,\dots,r-1$, we also define $\mathbf{a}_i$ as the sequence of random signs $(\alpha_{i,1},\dots,\alpha_{i,n},\alpha_{i,n+1},\dots,\alpha_{i,N})$ that are generated in Algorithm 1. Finally, we denote by $\mathbf{a}_i \circ \mathbf{r}$ the sequence $(\alpha_{i,1}v_1,\dots,\alpha_{i,n}v_n,\alpha_{i,n+1}\omega_1,\dots,\alpha_{i,N}\omega_m)$, that is, the element-wise product of $\mathbf{a}_i$ and $\mathbf{r}$. With this notation, it is immediate to check that $\|S_i(\theta^0)\|$, $i = 1,\dots,r-1$, can be written as $\|S_i(\theta^0)\| = Z(\mathbf{a}_i \circ \mathbf{r})$, where $Z(\cdot)$ is the same function as before. Define now $\mathbf{r}' := \mathbf{a}_0 \circ \mathbf{r}$, where $\mathbf{a}_0$ is a new independent sequence of random signs. Define also $\mathscr{R}'(\theta^0) := f_{\mathscr{R}}(Z(\mathbf{r}'), Z(\mathbf{a}_1 \circ \mathbf{r}'),\dots,Z(\mathbf{a}_N \circ \mathbf{r}'),\pi)$. Note that the sequence $\mathbf{r}'$ is equal *in distribution* to $\mathbf{r}$ (short notation: $\mathbf{r}' \overset{d}{=} \mathbf{r}$). Hence, also the sequence $(Z(\mathbf{r}'), Z(\mathbf{a}_1 \circ \mathbf{r}'),\dots,Z(\mathbf{a}_N \circ \mathbf{r}'))$ is distributed as $(Z(\mathbf{r}), Z(\mathbf{a}_1 \circ \mathbf{r}),\dots,Z(\mathbf{a}_N \circ \mathbf{r}))$, and therefore

$$\mathscr{R}(\theta^0) \overset{d}{=} \mathscr{R}'(\theta^0). \tag{11}$$

Conditioning on a given value of $\mathbf{r}$, say $\bar{\mathbf{r}}$, one can write $(Z(\mathbf{r}'), Z(\mathbf{a}_1 \circ \mathbf{r}'),\dots,Z(\mathbf{a}_N \circ \mathbf{r}')) = (Z(\mathbf{a}_0 \circ \bar{\mathbf{r}}), Z((\mathbf{a}_0 \circ \mathbf{a}_1) \circ \bar{\mathbf{r}}),\dots,Z((\mathbf{a}_0 \circ \mathbf{a}_N) \circ \bar{\mathbf{r}}))$. Since $\mathbf{a}_0, \mathbf{a}_1,\dots,\mathbf{a}_N$ are independent sequences of random signs, $\mathbf{a}_0, \mathbf{a}_0 \circ \mathbf{a}_1,\dots,\mathbf{a}_0 \circ \mathbf{a}_N$ are also independent sequences of random signs, and therefore the sequence $(Z(\mathbf{a}_0 \circ \bar{\mathbf{r}}), Z((\mathbf{a}_0 \circ \mathbf{a}_1) \circ \bar{\mathbf{r}}),\dots,Z((\mathbf{a}_0 \circ \mathbf{a}_N) \circ \bar{\mathbf{r}}))$ is exchangeable, that is, every permutation of its components has the same probability. In the absence of ties, $\mathscr{R}'(\theta^0)$ computes the rank of $Z(\mathbf{a}_0 \circ \bar{\mathbf{r}})$ in $(Z(\mathbf{a}_0 \circ \bar{\mathbf{r}}), Z((\mathbf{a}_0 \circ \mathbf{a}_1) \circ \bar{\mathbf{r}}),\dots,Z((\mathbf{a}_0 \circ \mathbf{a}_N) \circ \bar{\mathbf{r}}))$ and is therefore uniformly distributed over $\{1,\dots,r\}$. In the presence of ties, ties are broken by resorting to π, which is uniformly distributed over all the permutations, so that, conditional on $\mathbf{r} = \bar{\mathbf{r}}$, $\mathscr{R}'(\theta^0)$ is also uniformly distributed. Note now that $\mathbb{P}(\mathscr{R}'(\theta^0) > r-q|\mathbf{r} = \bar{\mathbf{r}}) = q/r$ for every $\bar{\mathbf{r}}$ implies that $\mathbb{P}(\mathscr{R}'(\theta^0) > r-q) = q/r$, and (9) follows from (11).

To prove star convexity, (8), observe first that the ranking function $\mathscr{R}(\theta)$ can be written more explicitly as $\mathscr{R}(\theta) = r - \sum_{i=1}^{r-1}\mathbf{1}(\|S_0(\theta)\| \le_\pi \|S_i(\theta)\|)$, where $\mathbf{1}(\cdot)$ is the indicator function (equal to 1 when the formula in the argument is true, 0 otherwise), and "$\le_\pi$" stands for "$\le$" if $\pi(0) < \pi(i)$, "$<$" otherwise (see points 3 and 4 in Algorithm 2). Thus, the region $\Theta(y)$ is the set where $\mathbf{1}(\|S_0(\theta)\| \le_\pi \|S_j(\theta)\|) \ge q$. Thus, by defining for all $i = 1,\dots,r-1$ the sets $\mathscr{E}_i^{\le} := \{\theta \in \mathbb{R}^m : \|S_0(\theta)\|^2 - \|S_i(\theta)\|^2 \le 0\}$, $\mathscr{E}_i^{<} := \{\theta \in \mathbb{R}^m : \|S_0(\theta)\|^2 - \|S_i(\theta)\|^2 < 0\}$, and $\mathscr{E}_i^{\pi}$ as equal to $\mathscr{E}_i^{\le}$ if $\pi(0) < \pi(i)$ and equal to $\mathscr{E}_i^{<}$ otherwise, we can write $\Theta(y)$ as the union of all the sets of the kind $\bigcap_{i=i_1,\dots,i_q}\mathscr{E}_i^{\pi}$, for all the possible choices of q indexes from the set $\{1,\dots,r-1\}$. The star convexity of the region follows from the fact, which is shown in what follows, that each set $\mathscr{E}_i^{\pi}$ is either convex and includes $\hat{\theta}^{ReLS}$ or it is empty. Function $\|S_0(\theta)\|^2$ is quadratic in θ and can be written as $S_0(\theta) = (\theta - \hat{\theta}^{ReLS})^T R_N(\theta - \hat{\theta}^{ReLS})$, which reveals that

$$\|S_0(\hat{\theta}^{ReLS})\|^2 = 0, \tag{12}$$

and that the Hessian is $2R_N$. On the other hand, it is straightforward to check that each function $\|S_i(\theta)\|^2$, $i = 1, \dots, r-1$, is a quadratic function in θ with Hessian equal to $2P_iR_N^{-1}P_i$, where $P_i = \frac{1}{N}\Phi^T D_{n,i}\Phi + \frac{\eta^2}{N}(\Sigma^{-1/2})^T D_{m,i}\Sigma^{-1/2}$, with $D_{n,i} = \mathrm{diag}(\alpha_{i,1}, \dots, \alpha_{i,n})$ and $D_{m,i} = \mathrm{diag}(\alpha_{i,n+1}, \dots, \alpha_{i,n+m})$. It is also easy to show (e.g. by an argument like the one in [12], Appendix B) that

$$R_N \succeq P_iR_N^{-1}P_i \tag{13}$$

in the Löwner partial ordering, i.e., $R_N - P_iR_N^{-1}P_i$ is positive semidefinite. From (13), we can conclude that $\|S_0(\theta)\|^2 - \|S_i(\theta)\|^2$ is a convex function in θ, so that the set $\mathscr{E}_i^{\leq}$ is a convex set, and $\hat{\theta}^{ReLS} \in \mathscr{E}_i^{\leq}$ because of (12). Likewise, the set $\mathscr{E}_i^{<}$ is either convex or empty, and, if it is not empty, then it must include $\hat{\theta}^{ReLS}$ because of the following: assume by contradiction that there is a $\bar{\theta} \in \mathscr{E}_i^{<}$ and $\hat{\theta}^{ReLS} \notin \mathscr{E}_i^{<}$. Then, $\|S_0(\bar{\theta})\|^2 < \|S_i(\bar{\theta})\|^2$. On the other hand, $\hat{\theta}^{ReLS} \notin \mathscr{E}_i^{<}$ implies that $\|S_i(\hat{\theta}^{ReLS})\|^2 = 0$, so that $\hat{\theta}^{ReLS}$ must be a minimum point also for the quadratic function $\|S_i(\theta)\|^2$. Thus, combining the fact that $\|S_0(\theta)\|^2$ and $\|S_i(\theta)\|^2$ have both a minimum point at $\hat{\theta}^{ReLS}$ with value 0 with the fact that the Hessian of $\|S_i(\theta)\|^2$ is no larger than the Hessian of $\|S_0(\theta)\|^2$, we must have that $\|S_0(\theta)\|^2 \geq \|S_i(\theta)\|^2$ for every θ, which contradicts $\|S_0(\bar{\theta})\|^2 < \|S_i(\bar{\theta})\|^2$. □

4.1. Constructing Bayesian SPS regions in sampled form

In this section, we describe an efficient Markov chain Monte Carlo (MCMC) scheme [17] to reconstruct in sampled form the Bayesian SPS region. We start by considering two subproblems which will be building blocks of the MCMC algorithm. In what follows, just to simplify the exposition, the SPS region is assumed to be bounded and closed. First, assume we are given a direction $\theta - \hat{\theta}^{ReLS}$ in the parameter space. Since the region is star convex with center $\hat{\theta}^{ReLS}$ (as seen in Theorem 1) there exist scalars x_{min} and x_{max} such that $\hat{\theta}^{ReLS} + x(\theta - \hat{\theta}^{ReLS}) \subseteq \Theta(y) \iff x \in [x_{min}, x_{max}]$. Our aim is to compute x_{min} and x_{max}, and, hence, also the two associated θ vectors belonging to the region's boundary. To this purpose, note that, if $\theta = \hat{\theta}^{ReLS} + x(\theta - \hat{\theta}^{ReLS})$, the squared norms $\|S_i(\theta)\|^2$ boil down to second-order polynomials, here denoted by $P_i(x)$. One then easily obtains that x_{min} and x_{max} belong to the set of the real roots of the polynomials $P_i(x) - P_0(x)$, $i \neq 0$, so allowing their efficient determination using Algorithm 2. This numerical procedure is denoted by $\mathscr{B}(\theta)$: it maps θ into the two boundary vectors that stay on the line $\hat{\theta}^{ReLS} + x(\theta - \hat{\theta}^{ReLS})$. As for the second subproblem, let $\mathbf{u}(\theta)$ denote the uniform probability density whose support is the SPS region $\Theta(y)$, and let θ_{-j} be the subvector obtained by removing from θ (in the SPS region) its j-th component θ_j. The aim is to draw the j-th component θ_j according to $\mathbf{u}(\theta_j|\theta_{-j})$, that is, uniformly from the SPS region and conditional on θ_{-j}. Note that the squared norms $\|S_i(\theta)\|^2$ now boil down to second-order polynomial functions of θ_j, which we denote by $P_i(\theta_j)$. It is then easy to see that the probability density function $\mathbf{u}(\theta_j|\theta_{-j})$ is uniform with (possibly non-connected) support that can be represented as $\bigcup_k [a_k, b_k]$, where the scalars a_k and b_k are contained in the set R of the real roots of the functions $P_i(\theta_j) - P_0(\theta_j)$, $i \neq 0$. The support of $\mathbf{u}(\theta_j|\theta_{-j})$ can then be efficiently determined e.g. by ordering the values in R ($r_1 \leq r_2 \leq \cdots$) and checking which θ vectors obtained by setting the j-th component equal to $\frac{r_k + r_{k+1}}{2}$ are accepted by Algorithm 2. The two above-described subproblems are part of the sampling strategy now summarized in Algorithm 3.

Some comments about Algorithm 3 are now in order. With probability p each iteration uses Gibbs sampling to update ℓ (randomly drawn) components of the Markov chain state. This step exploits the region's structure that allows to easily sample the conditional densities. Only a subvector of size ℓ is updated so that new boundary vectors can be frequently stored along new directions. The Gibbs sampling is fast and guaranteed to generate samples always inside the SPS region. However, it can turn out even more efficient to move all the components at once using Algorithm 2 only one time. For this reason, with probability $1-p$ a random walk Metropolis is exploited. Increments' covariance is proportional to the posterior covariance in (3) since $(\frac{\Phi^T\Phi}{\sigma^2} + \frac{\Sigma^{-1}}{\lambda^2})^{-1}$ influences the shape of the SPS region. In our implementation, the scale factor ξ in (14) is updated every 100 iterations during the first 1000 MCMC steps in order to ensure an acceptance rate of the random walk proposals around $30-40\%$.

Algorithm 3 Construction of the Bayesian SPS region in sampled form

```
Set the initial vector $\theta^{(1)}$ to $\hat{\theta}^{ReLS}$, let $p \in [0,1]$ and $\ell \in [1, \dots, m]$ where $m = \mathrm{dim}(\theta^{ReLS})$. Then, for $i = 2, \dots, M$, with $M$ the prescribed number of iterations, repeat the procedure described below;
Draw a realization $U$ uniformly from the unit interval;
If $U \leq p$, select uniformly $\ell$ components of $\theta^{(i-1)}$. Then, update them sequentially by drawing samples from $\mathbf{u}(\theta_j|\theta_{-j})$ where $\theta_{-j}$ are subvectors from the updated versions of $\theta^{(i-1)}$. Denote the vector so obtained by $\theta^{(i)}$ and store it together with the two boundary vectors $\mathscr{B}(\theta^{(i)})$.
If $U > p$, draw $\theta$ from the following Gaussian distribution
    $$\theta \sim \mathscr{N}\left(\theta^{(i-1)}, \xi(\frac{\Phi^T\Phi}{\sigma^2} + \frac{\Sigma^{-1}}{\lambda^2})^{-1}\right), \tag{14}$$
    where $\xi$ is a suitable scale factor (see discussion below). If $\theta$ is accepted by Algorithm 2, set $\theta^{(i)} = \theta$ and store it together with the two boundary vectors $\mathscr{B}(\theta^{(i)})$. Otherwise, define $\theta^{(i)} = \theta^{(i-1)}$.
```

Algorithm 3 returns two kinds of samples. The first are boundary vectors, denoted by $\theta_B^{(i)}$, which reconstruct the region's shell. The others are interior samples, denoted by $\theta_I^{(i)}$, which (after a burn in) are drawn uniformly from the SPS region in accordance with MCMC theory [17]. Let the dispersion index $\mathscr{D}_\Theta$ of the region be defined by the average distance from its centroid as follows:

$$\mathscr{D}_\Theta = \int_{\theta \in \Theta} \|\theta - c_\theta\| d\mathbf{u}(\theta), \quad c_\theta = \int_{\theta \in \Theta} \theta d\mathbf{u}(\theta). \tag{15}$$

Then, the $\theta_B^{(i)}$ allow to obtain the following Monte Carlo estimate of $\mathscr{D}_\Theta$:

$$\mathscr{D}_\Theta \approx \frac{1}{M}\sum_{i=1}^{M} \|\theta_I^{(i)} - \hat{c}_\theta\|, \quad \hat{c}_\theta = \frac{1}{M}\sum_{i=1}^{M} \theta_I^{(i)}. \tag{16}$$

Algorithm 3 has been used to solve the three case studies described in Section 3 setting $p = 0.1$ and $\ell = 10$. For this kind of problems our procedure (implemented in MATLAB on an iMAC 2.8 GHz IntelCore i7) is able to draw 1000 vectors from the SPS region in around 1 second.

5. SOLUTION OF THE THREE CASE STUDIES USING BAYESIAN SPS

The three bottom panels of Fig. 1 report the Bayesian SPS regions of level $\alpha = 0.95$ in sampled form. Each of them contains around 30000 samples drawn according to the stochastic simulation scheme detailed in Section 4.1. The estimates (16) of the dispersion indexes (15) corresponding to the Gaussian and Bayesian SPS regions are reported in Table 1. In all the three cases the Bayesian SPS region now contains the realization of θ^0. Note also that in the first case study, where the Gaussian model was correct, the SPS region dispersion is somewhat similar to that reported in the top panel. Conversely, in the other cases, where the Gaussian region dispersion is underestimated, the Bayesian SPS region is significantly larger and provides a much better picture about the uncertainty around the estimate.
We have also performed three Monte Carlo studies using the data generators associated to the three case studies. At any run new realizations of θ^0 and of the measurement noise are generated. Table 2 then reports the frequencies with which the 95% Gaussian and SPS regions contain the realization of θ^0 after 1000 runs. Differently from the Gaussian regions, which e.g. never contain the realization of θ^0 in the second case, for Bayesian SPS the value is always very close to 95%, confirming the theory previously illustrated.

Table 1. Dispersion index of the 95% regions in Fig. 1 achieved by Gaussian regression and Bayesian SPS.

Case study	#1	#2	#3
GR	4.1	2.1	9.3
BSPS	4.3	4.2	29.1

Table 2. 95% Bayesian regions accuracy for the three Monte Carlo studies of 1000 runs

Case study	#1	#2	#3
GR	94.6%	0%	2.1%
BSPS	94.7%	95.3%	94.7%

6. CONCLUSIONS

A new robust Bayesian framework based on the SPS technique has been introduced, focusing on finite-dimensional linear learning machines. The approach permits to equip regularization networks with uncertainty bounds that can be much more reliable than those returned by the classical Gaussian regression. In the future, we plan to extend the work in several directions. These include theoretical studies to better understand limits and potentials of this approach, also assessing the effect of the regularization parameters when estimated from the same data used to build the Bayesian regions.

7. REFERENCES

[1] F. Cucker and S. Smale, "On the mathematical foundations of learning," *Bulletin of the American mathematical society*, vol. 39, pp. 1–49, 2001.

[2] B. Schölkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*, (Adaptive Computation and Machine Learning). MIT Press, 2001.

[3] T. Poggio and F. Girosi, "Networks for approximation and learning," in *Proceedings of the IEEE*, 1990, vol. 78, pp. 1481–1497.

[4] H. Drucker, C.J.C. Burges, L. Kaufman, A. Smola, and V. Vapnik, "Support vector regression machines," in *Advances in Neural Information Processing Systems*, 1997.

[5] S. Smale and D.X. Zhou, "Learning theory estimates via integral operators and their approximations," *Constructive Approximation*, vol. 26, pp. 153–172, 2007.

[6] C.E. Rasmussen and C.K.I. Williams, *Gaussian Processes for Machine Learning*, The MIT Press, 2006.

[7] M.P. Deisenroth, D. Fox, and C.E. Rasmussen, "Gaussian processes for data-efficient learning in robotics and control," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 2, pp. 408–423, 2015.

[8] T.J. Hastie, R.J. Tibshirani, and J. Friedman, *The Elements of Statistical Learning. Data Mining, Inference and Prediction*, Springer, Canada, 2001.

[9] J.O. Berger, "A robust generalized Bayes estimator and confidence region for a multivariate normal mean," *The Annals of Statistics*, vol. 8, pp. 716–761, 1980.

[10] J.O. Berger, "An overview of robust Bayesian analysis," *Test*, vol. 3, pp. 5–124, 1994.

[11] M.J. Bayarri and J.O. Berger, "The interplay of Bayesian and frequentist analysis," *Stat. Science*, vol. 19, pp. 58–80, 2004.

[12] B.Cs. Csáji, M.C. Campi, and E. Weyer, "SPS: A new system identification approach for constructing exact non-asymptotic confidence regions in linear regression models," *IEEE Trans. Signal Processing*, vol. 63, no. 1, pp. 169–181, 2015.

[13] G. Pillonetto, A. Carè, and M.C. Campi, "Kernel-based SPS," in *Proceedings of the 18th IFAC Symposium on System Identification (SYSID), Stockholm*, 2018.

[14] A. Carè, B.Cs. Csáji, M.C. Campi, and E. Weyer, "Finite-sample system identification: An overview and a new correlation method," *IEEE Control Systems Letters*, vol. 2, no. 1, pp. 61–66, Jan 2018.

[15] G. Wahba, *Spline models for observational data*, SIAM, Philadelphia, 1990.

[16] J. Dezert and C. Musso, "An efficient method for generating points uniformly distributed in hyperellipsoids," in *Proceedings of the Workshop on Estimation, Tracking and Fusion, Monterey*, 2011.

[17] W. Gilks, S. Richardson, and D. Spiegehalter, *Markov Chain Monte Carlo in Practice*, London: Chapman and Hall, 1996.