Distributed multi-agent Gaussian regression via finite-dimensional approximations

Gianluigi Pillonetto, Luca Schenato, Damiano Varagnolo

***Abstract*—We consider the problem of distributedly estimating Gaussian processes in multi-agent frameworks. Each agent collects few measurements and aims to collaboratively reconstruct a common estimate based on all data. Agents are assumed with limited computational and communication capabilities and to gather M noisy measurements in total on input locations independently drawn from a known common probability density. The optimal solution would require agents to exchange all the M input locations and measurements and then invert an $M \times M$ matrix, a non-scalable task. Differently, we propose two suboptimal approaches using the first E orthonormal eigenfunctions obtained from the Karhunen-Loève (KL) expansion of the chosen kernel, where typically $E \ll M$. The benefits are that the computation and communication complexities scale with E and not with M, and computing the required statistics can be performed via standard average consensus algorithms. We obtain probabilistic non-asymptotic bounds that determine a priori the desired level of estimation accuracy, and new distributed strategies relying on Stein's unbiased risk estimate (SURE) paradigms for tuning the regularization parameters and applicable to generic basis functions (thus not necessarily kernel eigenfunctions) and that can again be implemented via average consensus. The proposed estimators and bounds are finally tested on both synthetic and real field data.**

***Index Terms*—Gaussian processes, sensor networks, distributed estimation, kernel-based regularization, nonparametric estimation, average consensus**

I. Introduction

Many modern engineering problems involve networks containing a large number of agents which have to cooperate to obtain a common goal. Several of these tasks can be seen as problems of function estimation from sparse and noisy data, a central issue in the machine learning field [1], [2]. Examples include the determination of the wind speed and direction field in a wind farm from local measurements of the turbines [3], the reconstruction of the temperature field in a datacenter from local measurements at each server [4], and weather forecasts [5], [6]. Traditional centralized machine-learning estimation approaches are computationally non-scalable when the network is large. Moreover parallelization of computation using client-server architectures, which can alleviate this problem, might not be feasible. This happens, e.g., in applications where communication is peer-to-peer, as in wireless sensor networks or multi-agent robotics, and where each agent is expected to have a common copy of the global estimate. In these cases, fully distributed cooperation approaches are ought [7].

A. State-of-the-art

This paper considers a distributed nonparametric Gaussian regression approach. In this context, the unknown map is modeled as a zero-mean Gaussian process whose covariance (also called kernel in the machine learning literature) has to embed expected properties like smoothness [8], [9]. Other approaches to function estimation could be also adopted, e.g., sparse regression based on the ℓ_1 norm, automatic relevance determination or the elastic net [10], [11], [12], [13], [14]. However, in our framework the implementation of these approaches is not trivial and would require sophisticated distributed optimization algorithms like ADMM [15]. In fact, we consider a scenario where N agents first collect a total of M direct and noisy measurements of the unknown map on input locations drawn from a common and known probability density. The aim is then to obtain a shared function estimate. To simplify the exposition, we assume w.l.o.g. $N = M$, i.e., each agent collects a single measurement. We also assume that computational and data storage capabilities are limited, and that the communication network is peer-to-peer, i.e., agents are able only to communicate with a restricted number of neighbors. As described below, this makes the problem difficult also under Gaussian process assumptions, but we will see that function estimation can be performed using simple average operations.

Assuming that f and the measurements noise are jointly Gaussian, achieving the minimum variance estimate requires knowing all the M measurements and related input locations, plus invert an $M \times M$ matrix with $O(M^3)$ operations, a difficult task in a distributed fashion. When the data set size M is large, the complexity is high also in centralized contexts. Therefore, many alternative approaches have been developed relying, e.g., on the notion of pseudo input locations [16], [17], [18], the use of matrix factorizations [19] and approximations of the kernel function [20], [21] through the Nyström method or greedy techniques [22], [23], [24]. Along this way, KL expansions [25] have been also used to decompose the kernel in terms of eigenfunctions that are orthogonal w.r.t. the input locations probability density. One can then approximate the Gaussian process via the E kernel eigenfunctions associated to the largest eigenvalues, an approximation that corresponds

G. Pillonetto and L. Schenato are with the Department of Information Engineering, University of Padova, Padova, Italy. D. Varagnolo is with the Department of Computer Science, Electrical and Space Engineering, Luleå University of Technology, Luleå, Sweden. Emails: `giapi@dei.unipd.it | schenato@dei.unipd.it | damvar@ltu.se`.

The research leading to these results has received funding from the Swedish research council Norrbottens Forskningsråd.

to perform the best process approximation before seeing the data [25] (see Section III-A for more details). A posteriori, i.e., after seeing the measurements and their input locations, the situation is instead more subtle since there exist E-dimensional subspaces that allow to come closer to the minimum variance estimator [26]. However, the a priori basis given by the KL expansion has important advantages. In fact, as proved in [27], the first E kernel eigenfunctions are asymptotically optimal, i.e., they provide the best E-dimensional approximation of the minimum variance estimator as the data set size M grows to infinity. In addition, differently from the a posteriori basis described in [26], the a priori basis can be computed off-line. Moreover, as detailed in Section III-B, computing the final estimates requires computing sufficient statistics that have the structure of averages of M local matrices and local vectors of dimension respectively $E \times E$ and E. This implies that the basic building block of the estimators involves computing averages over networks which can be more efficient from a memory, computation and communication perspective when $E \ll M$. Such averages can be computed via the so called *average consensus algorithms* [28], [29] which require only mild assumption on network connectivity and communication. In particular, these algorithms require no global topological information, only minimal local coordination and can be implemented also in the context of asynchronous updates and lossy communication [30].

B. Contribution

Our stream of research pairs the ones of other authors focusing on distributed kernel regression. An example is [31], that proposes a distributed regularized kernel Least Squares (LS) regression algorithm that exploits successive orthogonal projections, or [32] that extends [31] by designing strategies to reduce the communication and synchronization needs. Estimators with reduced order model complexity have been proposed in [33], while nonparametric schemes using Nearest-Neighbors interpolation strategies have been studied also in [34]. Another Gaussian estimation approach is considered in [35], with focus on the problem of sequentially predicting the most informative future input locations to minimize simultaneously the prediction error and the uncertainty in the regularization parameters. Other distributed regression algorithms are proposed in [36] with the aim of estimating a dynamic Gaussian process and its gradient, while in [37] authors develop a distributed learning and cooperative control algorithm where agents estimate a static field modeled as a network of radial basis functions whose centers locations are known in advance.

Despite the many research efforts, none of the aforementioned works on distributed regression have addressed the following fundamental issue: *assigned a Gaussian prior (the kernel) and the input locations distribution, how much information does the network need to exchange to obtain, with a probability* $1-\alpha$*, the desired level of estimation accuracy?* In this paper we will answer this question adopting KL-based strategies which exploit E kernel eigenfunctions. In particular, we will study two different estimators denoted by $\widehat{f}_A$ and $\widehat{f}_B$ which have computational and communication complexities of order $O(E^2)$ and $O(E)$, respectively, originally proposed in [38]. Differently from [38] which focused on finding Monte Carlo based strategies for assessing the a posteriori statistical performance of the estimators, in this work the focus is on characterizing their a priori prediction capability on future data by first assigning the kernel and the input locations statistics, and then deriving non-asymptotic error bounds that are functions of E, M and α. This analysis can be also seen as the extension to the Bayesian context of the concept of *effective dimension* developed in deterministic frameworks, e.g., in [39]. There it has been shown that, in the worst case, subspaces of dimension $\sqrt{M}$, i.e., sub-polynomial in the data set size, capture the estimate. Parallel to this, our bound returns information on the *Bayesian effective dimension* revealing which subspace can be really influenced by the measurements.

Another major contribution provided in this work is to show that both $\widehat{f}_A$ and $\widehat{f}_B$ are asymptotically optimal, i.e., for fixed E, as M grows to infinity there is no other estimator which can perform better in the mean squared error sense. We will also see that, while $\widehat{f}_A$ is always consistent, i.e., convergent in probability to the true function as $E, M \to \infty$, consistency of $\widehat{f}_B$ requires E to grow slower than M. In some sense, such result clarifies the price to pay when adopting a estimator parsimonious in the information exchange.

Finally, in many applications the kernel scale factor is unknown and its tuning is critical since it strongly affects the performance of the Bayesian estimator. In addition, the kernel expansion could be hard to be obtained and one would rather use a different set of basis functions. In the terminal part of the paper, we address these problems by proposing a novel distributed tuning strategy based on the SURE criterion [40]. Standard approaches proposed in the literature in the context of a centralized framework (like cross-validation and maximum likelihood [41], [42], [43], [44], MAP estimation [45], expected improvement [46] and Markov chain Monte Carlo [47], [48]) require high computation and communication overhead, and are therefore not suited for distributed implementations. Instead, our strategy allows for simultaneous hyperparameter tuning and function estimation via a single average consensus algorithm over a vector of size $O(E^2)$ when $\widehat{f}_A$ is employed, and via only two averages of size $O(E)$ when $\widehat{f}_B$ is employed. Very importantly, the SURE criterion be used also for generic basis functions, such as kernel sections or Nyström bases, thus not necessarily restricted to be kernel eigenfunctions and defined.

C. Paper outline

The paper is organized as follows. Section II formulates the Bayesian estimation problem while Section III describes the KL expansion of the Gaussian process and the distributed estimators. Section IV provides the statistical characterization of our distributed estimators, also deriving error bounds which are then tested via some numerical experiments. Section V proposes distributed strategies to tune the possibly unknown regularization parameter entering our estimators for generic basis functions and discusses practical implementation issues. These strategies are also tested on both synthetic and real data. Section VI collects conclusions and future research directions while proofs are collected in the Appendix.

II. Bayesian estimation

A. The measurements model

We consider the measurements model

$$y_m = f(x_m) + \nu_m, \qquad m = 1, \dots, M \tag{1}$$

with the input locations x_m following the stochastic generation scheme

$$x_m \sim \mu(\mathcal{X}) \text{ i.i.d.}, \qquad m = 1, \dots, M, \tag{2}$$

with μ a non-degenerate probability measure on the compact $\mathcal{X}$. The unknown function $f : \mathcal{X} \to \mathbb{R}$ is a zero-mean Gaussian process with continuous covariance $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, i.e.,

$$f \sim \mathcal{N}(0, K). \tag{3}$$

The measurement noise is also Gaussian of known variance σ_ν^2:

$$\nu_m \sim \mathcal{N}\left(0, \sigma_\nu^2\right).$$

Finally, $\{\nu_m\}_{m=1}^M$, $\{x_m\}_{m=1}^M$ and f are all assumed mutually independent.

B. The Bayesian estimator

The Gaussian assumptions of Section II-A imply that the posterior of f given the dataset $\{x_m, y_m\}_{m=1}^M$ is still Gaussian. Also, the Maximum A Posteriori (MAP) estimator coincides with the minimum variance estimator and is given by

$$\widehat{f}_{\text{MAP}}(x) = \begin{bmatrix} K(x, x_1) & \dots & K(x, x_M) \end{bmatrix} H_{\text{MAP}} \begin{bmatrix} y_1 \\ \vdots \\ y_M \end{bmatrix}$$

with

$$H_{\text{MAP}} := \left(\begin{bmatrix} K(x_1, x_1) & \cdots & K(x_1, x_M) \\ \vdots & & \vdots \\ K(x_M, x_1) & \cdots & K(x_M, x_M) \end{bmatrix} + \sigma_\nu^2 I \right)^{-1}.$$

The storage and computational requirements needed to compute $\widehat{f}_{\text{MAP}}$ are thus $O\left(M^2\right)$ and $O\left(M^3\right)$, respectively. The communication complexity is either $O(\dim(\mathcal{X})M)$ if agents share the input locations x_m or $O\left(M^2\right)$ if they share the covariances $K(x_m, x_{m'})$. Thus, storage, computational and communication complexities do not scale favorably with the dataset size M. Our aim is thus to find good approximators of $\widehat{f}_{\text{MAP}}$ that are suitable for distributed implementations.

III. Finite-dimensional approximations of the Bayesian estimator

A. KL expansion: kernel

The kernel (3) can be expanded in terms of eigenfunctions ϕ_e orthonormal w.r.t. the measure μ in (2) and related eigenfunctions λ_e [25]. They are defined by

$$\lambda_e \phi_e(x) = \int_{\mathcal{X}} K(x, x') \phi_e(x') d\mu(x'), \tag{4}$$

$$K(x, x') = \sum_{e=1}^{+\infty} \lambda_e \phi_e(x) \phi_e(x') \qquad \lambda_1 \geq \lambda_2 \ldots > 0, \tag{5}$$

and, using δ_{ij} for the Kronecker delta,

$$\int_{\mathcal{X}} \phi_i(x) \phi_j(x) d\mu(x) = \delta_{ij}. \tag{6}$$

Let E be a positive integer. Then (4), (5) and (6) allow us to reformulate the process f via the following KL expansion

$$f(x) = \underbrace{\sum_{e=1}^{E} a_e \phi_e(x)}_{=:\, f_a(x)} + \underbrace{\sum_{e=1}^{+\infty} b_e \phi_{E+e}(x)}_{=:\, f_b(x)}. \tag{7}$$

The expansion coefficients have been thus divided into two sets: a finite one composed by the E random variables a_e, and an infinite one given by the remaining variables b_e. The elements in these two sets are all mutually independent, and satisfy

$$a_e \sim \mathcal{N}(0, \lambda_e), \ e = 1, \dots, E \tag{8a}$$

$$b_e \sim \mathcal{N}(0, \lambda_{E+e}), \ e = 1, 2, \dots \tag{8b}$$

It is well known that

$$\mathcal{S} := \text{span}\langle \phi_1(\cdot), \dots, \phi_E(\cdot) \rangle \tag{9}$$

is that E-dimensional subspace that captures the biggest part of the statistical energy of f as measured by $\mathbb{E}\left[\int f^2 d\mu\right]$. In other words, f_a is the best E-dimensional approximation of f in the mean square sense [27].

In what follows, it is always assumed that all the kernel eigenfunctions are contained in a ball of finite radius in the space of continuous functions, i.e.,

Assumption 1 *There exists a $k < +\infty$ s.t.*

$$\sup_{x \in \mathcal{X}} |\phi_e(x)| \leq \sqrt{k} < +\infty \qquad e = 1, 2, \dots. \tag{10}$$

Assumption 1 is satisfied by all the finite-dimensional kernels and also by classical covariances like the spline kernels, e.g., see [49] for the case of uniform μ. In practice, if the KL expansion is not available in closed form, it can be obtained numerically with arbitrary accuracy, as for example described in [50], also permitting to compute the constant k.

B. KL expansion: measurement model

Our next step is to search for finite-dimensional estimators of f suitable for distributed implementations. Below, we introduce two different estimators, denoted by $\widehat{f}_A$ and $\widehat{f}_B$, which assume values in the finite-dimensional subspace $\mathcal{S}$ defined in (9). First, it is useful to rewrite model (1) in a more compact form.

Let

$$\boldsymbol{x} := [x_1, \dots, x_M]^T$$

$$\boldsymbol{y} := [y_1, \dots, y_M]^T \qquad \boldsymbol{\nu} := [\nu_1, \dots, \nu_M]^T \tag{11}$$

$$\boldsymbol{a} := [a_1, \ldots, a_E]^T \qquad \boldsymbol{b} := [b_1, b_2, \ldots]^T \tag{12}$$

$$G := \begin{bmatrix} G_{11} & \ldots & G_{1E} \\ \vdots & & \vdots \\ G_{M1} & \ldots & G_{ME} \end{bmatrix} \qquad Z := \begin{bmatrix} Z_{11} & Z_{12} & \ldots \\ \vdots & & \vdots \\ Z_{M1} & Z_{M2} & \ldots \end{bmatrix} \tag{13}$$

$$G_{me} := \phi_e(x_m), \quad m = 1, \ldots, M,\ e = 1, \ldots, E,$$

$$Z_{me} := \phi_{E+e}(x_m), \quad m = 1, \ldots, M,\ e = 1, 2, \ldots \tag{14}$$

Considering decomposition (7), definitions (11)-(14) and using classical algebraic notation to handle infinite-dimensional objects, the measurements model (1) becomes

$$\boldsymbol{y} = G\boldsymbol{a} + Z\boldsymbol{b} + \boldsymbol{\nu}. \tag{15}$$

With this novel notation $G\boldsymbol{a}$ accounts for the contribution from f_a while $Z\boldsymbol{b}$ accounts for the contribution from f_b.

C. The E-dimensional estimator $\widehat{f}_A$

Let

$$\widehat{f}_A(x) := \left[\phi_1(x) \ \cdots \ \phi_E(x)\right] H_A \boldsymbol{y} \tag{16}$$

where

$$H_A := \left(\frac{G^T G}{M} + \frac{\sigma_\nu^2}{M} \Lambda_E^{-1} \right)^{-1} \frac{G^T}{M} \tag{17}$$

and $\Lambda_E := \operatorname{diag}(\lambda_1, \ldots, \lambda_E)$. The estimator $\widehat{f}_A$ is suitable for distributed computations. In fact, defining

$$G_m := [\phi_1(x_m), \ldots, \phi_E(x_m)]$$

one has

$$\frac{G^T G}{M} = \frac{1}{M} \sum_{m=1}^{M} G_m^T G_m, \quad \frac{G^T \boldsymbol{y}}{M} = \frac{1}{M} \sum_{m=1}^{M} G_m^T y_m. \tag{18}$$

Since $G_m^T G_m \in \mathbb{R}^{E \times E}$ and $G_m^T y_m \in \mathbb{R}^E$ are local quantities, (18) points out that $\widehat{f}_A$ can be distributedly computed through the parallelization of two average consensus strategies: one on the $G_m^T G_m$'s and one on the $G_m^T y_m$'s, for a total of $E^2 + E$ scalars. This estimator would correspond to the Minimum Variance Unbiased Estimator (MVUE) estimator if the process f in (7) were truncated just to f_a.

D. The E-dimensional estimator $\widehat{f}_B$

As stated in (6), one has

$$\mathbb{E}\left[\left[\frac{G^T G}{M} \right]_{e,e'} \right] = \int_{\mathcal{X}} \phi_e(x)\, \phi_{e'}(x)\, d\mu(x) = \delta_{e,e'}.$$

and, given the assumptions in Section II-A and Assumption 1, the following convergence in probability holds

$$\frac{G^T G}{M} = \frac{1}{M} \sum_{m=1}^{M} G_m^T G_m \xrightarrow{M \to +\infty} \mathbb{E}\left[\frac{G^T G}{M} \right] = I.$$

Thus, it is tempting to use the approximation

$$\frac{G^T G}{M} \approx I \tag{19}$$

and use, in place of H_A in (16), the matrix

$$H_B := \left(I + \frac{\sigma_\nu^2}{M} \Lambda_E^{-1} \right)^{-1} \frac{G^T}{M}. \tag{20}$$

In turn, this approach approximates $\widehat{f}_A$ with

$$\widehat{f}_B(x) := \left[\phi_1(x) \ \cdots \ \phi_E(x)\right] H_B \boldsymbol{y}.$$

The estimator $\widehat{f}_B$ is more advantageous than $\widehat{f}_A$ for distributed computations. In fact, it requires an average consensus on just the column vectors $G_m^T y_m$'s, for a total of E scalars (differently from the $E^2 + E$ ones required by $\widehat{f}_A$), and does not require any expensive matrix inversion since $I + \frac{\sigma_\nu^2}{M} \Lambda_E^{-1}$ is diagonal.

IV. Statistical analysis of $\widehat{f}_A$ and $\widehat{f}_B$

Ideally one would like to compute $\mathbb{E}\left[\|f - \widehat{f}_A\|^2 \right]$ and $\mathbb{E}\left[\|f - \widehat{f}_B\|^2 \right]$, or at least some bounds that quantify the performance of the estimator for any specified E and M a priori. However, the computation of such quantities is intractable or, at least, requires an expensive Monte Carlo analysis, possibly to be repeated for many different design variables like, e.g., M, E, σ_ν^2. To circumvent this challenge, we will exploit the assumption that the input locations are randomly drawn from a known distribution μ and the orthonormality of the eigenfunctions to find bounds on $\mathbb{E}\left[\|f - \widehat{f}_A\|^2 \right]$ that hold with arbitrarily high probability. More specifically, the key idea is to find an event $\mathcal{E}$ that occurs with arbitrarily high probability such that informative bounds on $\mathbb{E}\left[\|f - \widehat{f}_A\|^2 \mid \mathcal{E} \right]$ can be computed. This is formally described in the next sections.

A. Performance indexes and lower bound

Two important performance indexes we consider for $\widehat{f}_A$ and $\widehat{f}_B$ are the errors defined by the conditional expectations

$$\begin{aligned} \mathrm{Err}_A(\boldsymbol{x}) &:= \mathbb{E}\left[\left\| f - \widehat{f}_A \right\|^2 \mid \boldsymbol{x} \right] \\ \mathrm{Err}_B(\boldsymbol{x}) &:= \mathbb{E}\left[\left\| f - \widehat{f}_B \right\|^2 \mid \boldsymbol{x} \right] \end{aligned} \tag{21}$$

where

$$\|g\|^2 := \int_{\mathcal{X}} g^2(x) d\mu(x).$$

The variables $\mathrm{Err}_A(\boldsymbol{x})$ and $\mathrm{Err}_B(\boldsymbol{x})$ are stochastic, since they are functions of the random input locations $\boldsymbol{x}$ that in our settings are assumed random as described in (2). Hence, the crux of our analysis will be how to account for the randomness coming from $\boldsymbol{x}$. Note also that $\|\cdot\|$ depends on μ so that Err_A and Err_B quantify the prediction errors on future data independently drawn from the same training set distribution.

Exploiting the KL expansion introduced in Section III-A a lower bound on the errors $\mathrm{Err}_A(\boldsymbol{x})$ and $\mathrm{Err}_B(\boldsymbol{x})$ can be also easily obtained. More generally, the following result bounds the performance achievable by any generic E-dimensional estimator of f.

Theorem 2 *Let $\widehat{f}_\star$ be any generic estimator of f, function of $\boldsymbol{x}$ and $\boldsymbol{y}$ and assuming values in any generic E-dimensional* *space fixed a priori. Then*

$$\min_{\widehat{f}_\star} \mathbb{E}\left[\|f-\widehat{f}_\star\|^2 \mid \boldsymbol{x}\right] \geq \sum_{e=E+1}^{+\infty} \lambda_e. \tag{22}$$

The following definition will be especially important for our future developments.

Definition 3 *We say that $\overline{\mathrm{Err}}_A \leq q$ or $\overline{\mathrm{Err}}_B \leq q$ with probability $1-\alpha$ if there exists an event $\mathcal{E}$ in the σ-algebra induced by $\boldsymbol{x}$ of probability at least $1-\alpha$ such that, respectively,*

$$\mathbb{E}\left[\mathrm{Err}_A(\boldsymbol{x}) \mid \mathcal{E}\right] \leq q \tag{23}$$

or

$$\mathbb{E}\left[\mathrm{Err}_B(\boldsymbol{x}) \mid \mathcal{E}\right] \leq q. \tag{24}$$

Thus, if α is close to zero saying that $\overline{\mathrm{Err}}_A \leq q$ with probability $1-\alpha$ is equivalent to saying that the average error associated to $\widehat{f}_A$ is smaller than q with high probability. Finally, note that setting $\mathcal{E}$ to the entire sample space, the conditional expectations in the Left Hand Side (LHS) of (23) and (24) become unconditional ones, and actually correspond to the Mean Square Errors (MSEs) of $\widehat{f}_A$ and $\widehat{f}_B$, i.e.,

$$\mathrm{MSE}_{\widehat{f}_A} = \int_{\mathcal{X}} \mathrm{Err}_A(x) d\mu(x), \tag{25}$$

$$\mathrm{MSE}_{\widehat{f}_B} = \int_{\mathcal{X}} \mathrm{Err}_B(x) d\mu(x). \tag{26}$$

B. Non asymptotic error bounds

The key issue is to bound the performance indexes Err_A and Err_B for any finite number of measurements M and eigenfunctions E. The following theorem provides the desired bounds. It depends on the input locations distribution μ, the kernel eigenvalues λ_e and constant k defined in (4) and (10), the number of eigenfunctions E and measurements M. In addition the bound is also function of a parameter $\varepsilon \in (0,1]$ connected to maximal and minimal (stochastic) eigenvalue of $\frac{G^T G}{M}$, as detailed in the proof contained in the Appendix.

Theorem 4 *Let the assumptions in Section II-A and Assumption 1 hold, $\alpha \in (0,1)$ be a desired confidence level (e.g., 0.01 or 0.05), and $\varepsilon \in (0,1]$ be given. If E, M and k satisfy*

$$1-\varepsilon+\varepsilon\log(\varepsilon) \geq \frac{Ek}{M}\log\left(\frac{E}{\alpha}\right) \tag{27}$$

then with probability at least $1-\alpha$ it holds that

$$\overline{\mathrm{Err}}_A \leq \mathrm{Bnd}_A \tag{28}$$

with

$$\mathrm{Bnd}_A := \frac{kM}{1-\alpha}\left(\sum_{e=1}^{E}\frac{\lambda_e^2}{(\varepsilon M\lambda_e+\sigma_\nu^2)^2}\right)\left(\sum_{e=E+1}^{+\infty}\lambda_e\right) + \frac{\sigma_\nu^2}{1-\alpha}\left(\sum_{e=1}^{E}\frac{\lambda_e}{\varepsilon M\lambda_e+\sigma_\nu^2}\right) + \left(\sum_{e=E+1}^{+\infty}\lambda_e\right). \tag{29}$$

Under the same assumption but with E, M and k now satisfying

$$1-\varepsilon+\varepsilon\log(\varepsilon) \geq \frac{Ek}{M}\log\left(\frac{2E}{\alpha}\right), \tag{30}$$

then with probability at least $1-\alpha$ it holds that

$$\overline{\mathrm{Err}}_B \leq \mathrm{Bnd}_B \tag{31}$$

with

$$\mathrm{Bnd}_B := \frac{kM}{1-\alpha}\left(\sum_{e=1}^{E}\frac{\lambda_e^2}{(M\lambda_e+\sigma_\nu^2)^2}\right)\left(\sum_{e=E+1}^{+\infty}\lambda_e\right) + \frac{\sigma_\nu^2}{1-\alpha}\left(\sum_{e=1}^{E}\frac{\lambda_e}{\varepsilon M\lambda_e+\sigma_\nu^2}\right) + \left(\sum_{e=E+1}^{+\infty}\lambda_e\right) + \kappa\left(\frac{E}{M}\sigma_\nu^2+\sum_{e=1}^{E}\lambda_e\right) \tag{32}$$

where

$$\kappa = \frac{1}{1-\alpha}\left(\varepsilon+\frac{\lambda_1^{-1}\sigma_\nu^2}{M}\right)^{-4}(1-\varepsilon)^2(2-\varepsilon)^2 \tag{33}$$

The obtained bounds are now tested via a numerical example.

C. Numerical study

Consider the first-order spline kernel [51] which corresponds to the Brownian motion covariance, i.e.,

$$K(x,x') = \min(x,x') = \sum_{e=1}^{\infty}\lambda_e\phi_e(x)\phi_e(x')$$

with the input locations probability measure μ in (2) set to the uniform distribution on $[0,1]$. With these settings

$$\phi_e(x) = \sqrt{2}\sin\left(x(e\pi-\pi/2)\right), \quad \lambda_e = \frac{1}{(e\pi-\pi/2)^2}$$

and $k=2$. To make the bounds only depend on E we set $M=10000$, $1-\alpha=0.95$, the noise variance $\sigma_\nu^2=0.1^2$, and $\varepsilon\in(0,1]$ that minimizes the bound while satisfying (27) or (30) accordingly.

The thick lines in the two top panels of Figure 1 show how Bnd_A (left) and Bnd_B (right) vary with E (bounds are normalized using the prior process variance $\sum_{e=1}^{\infty}\lambda_e$). For the sake of comparison we also display the true (normalized) MSEs (dashed line) as defined in (25) and (26), calculated via a Monte Carlo of 1000 runs, and its lower bound (thin line), i.e., $\sum_{e=E+1}^{\infty}\lambda_e/\sum_{e=1}^{\infty}\lambda_e$ as illustrated in Theorem 2.

As for Bnd_A, it is interesting to notice that just 20 eigenfunctions are needed to obtain an high estimation accuracy in both the cases. In addition, the curve is very close to the true error profile (which in turn is close to the lower bound) and is monotonically decreasing. Indeed, as discussed in the proof of Theorem 7 contained in the next subsection, when one adopts $\widehat{f}_A$ one should set E as large as possible (compatibly with communication capabilities) since, at the limit, convergence to the minimum variance estimator holds.

The profile of Bnd_B is instead different and exhibit a clear minimum at $E=7$. The reason is that $\widehat{f}_B$ relies on the asymptotic matrix approximation (19). The bound Bnd_B then points out that if E is too large then the quality of this approximation can worsen, hence leading to an increment of the corresponding MSE. One can see that also the true error profile is not monotonically decreasing (indeed, we will see in the next subsection that for M fixed and E going to infinity $\widehat{f}_B$ is not guaranteed to converge to the minimum variance

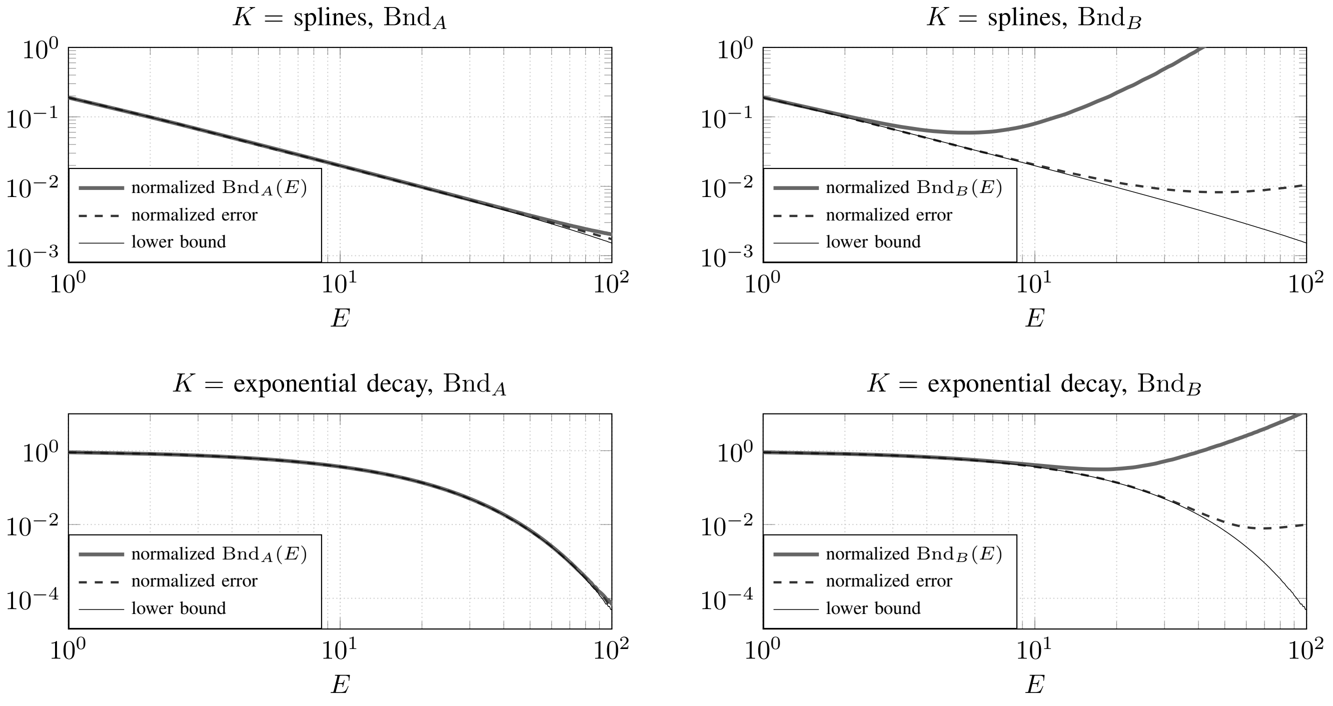

Figure 1. Bnd_A and Bnd_B (normalized by the a priori function variance) as a function of E, with $\alpha = 0.05, M = 10000$ and for different eigenvalues decay rates.

estimator). Note that, in this case, Bnd_B is close to truth only for low values of E and that the Monte Carlo analysis suggests the best E to be around 50. Overall, this indicates that the eigenfunctions number has to be seen as an important design parameter for $\widehat{f}_B$ to optimize the performance. This point will be the focus of Section V.

Finally, the two bottom panels of Figure 1 display the same bounds except that the kernel eigenvalues now decay exponentially to zero as $\lambda_e = \exp(-0.1e)$. Exponentially decaying eigenvalues are typical for Gaussian kernels, and therefore of practical relevance. The shapes of the curves change but the same comments hold true.

D. Asymptotic behaviors of the estimators and of the bounds

Now, we start investigating the asymptotic properties of our estimators considering a situation where their dimension E is fixed while the number of measurements M grows to infinity. The next result then shows that $\widehat{f}_A$ and $\widehat{f}_B$ asymptotically reach the lower bound (22).

Theorem 5 *Given the assumptions in Section II-A and Assumption 1,*

$$\lim_{M\to+\infty} \mathrm{Err}_A = \sum_{e=E+1}^{+\infty} \lambda_e \qquad \textit{in probability}$$

$$\lim_{M\to+\infty} \mathrm{Err}_B = \sum_{e=E+1}^{+\infty} \lambda_e \qquad \textit{in probability.}$$

We now discuss the statistical consistency of our estimators. In this case, the conditions under which $\widehat{f}_A$ and $\widehat{f}_B$ converge to f as both E and M grow to infinity are different, as illustrated in the following two results.

Theorem 6 *Given the assumptions in Section II-A and Assumption 1,*

$$\lim_{M\to+\infty} \lim_{E\to+\infty} \mathrm{Err}_A = 0 \qquad \textit{in probability} \tag{34}$$

Theorem 7 *Let $E = E(M)$ such that*

$$E(M)\log E^{\delta}(M) \leq M^{\delta}, \qquad \lim_{M\to+\infty} E(M) = +\infty$$

for some $\delta \in (0,1)$. Given the assumptions in Section II-A and Assumption 1, then

$$\lim_{\substack{M\to+\infty \\ E=E(M)}} \mathrm{Err}_B = 0 \qquad \textit{in probability} \tag{35}$$

Remark 8 *The sufficient condition required in the theorem in terms of the growth rate of $E(M)$ as a function of M is tight according to the Chernoff's bound. In fact, our requirement is that M grows up a bit more slowly w.r.t. the relationship $E\log E = M$. Now, assume instead that $E\log E = M$, i.e., $\delta = 1$ and fix any rule such that $\varepsilon \to 1$ and $\alpha \to 0$. Recall that*

$$1 - \varepsilon + \varepsilon\log(\varepsilon) \geq \frac{Ek}{M}\log\left(\frac{2E}{\alpha}\right)$$

must be satisfied. Asymptotically, the lhs tends to 0^+ while the second term becomes $k - \frac{Ek}{M}\log(\alpha/2)$ and is larger than k when α is sufficiently close to zero. One would thus need $0 \geq k$ but this is not possible. Also note that the previous theorem implies that any sublinear power growth of $E(M) = M^a$, for *any $a \in (0,1)$, satisfies the consistency condition, which can be readily verified by choosing $\delta = \frac{1+a}{2}$.*

The consistency properties of $\widehat{f}_A$ and $\widehat{f}_B$ are thus remarkably different. For what regards $\widehat{f}_A$, as M goes to infinity its consistency is guaranteed without any control on the growth rate of the dimension E. Indeed, as E increases such estimator can approximate arbitrarily well the optimal $\widehat{f}_{\text{MAP}}$. This agrees with what already discussed in the previous subsection: when using $\widehat{f}_A$ it is convenient for the network to use a dimension E as large as possible, just compatible with its communication constraints. Differently, the estimator $\widehat{f}_B$ is instead consistent only if M augments sufficiently faster than E.

V. Distributed tuning of the regularization parameter

The statistical bounds obtained in the previous section quantify the performance of $\widehat{f}_A$ and $\widehat{f}_B$ assuming that the prior function model is correct. Beyond their theoretical interest, in real applications these bounds can give useful guidelines to select the amount of information that agents need to exchange. However, the covariance K is often defined only except for a scalar factor γ. In addition, the prior is never perfect and the tuning of γ could also hinder possible undermodeling. So, in place of (3), in practical applications it is beneficial to consider

$$f \sim \mathcal{N}\left(0, \gamma^{-1}K\right)$$

with γ to be estimated from the observed noisy outputs and related input locations. Furthermore, when $\widehat{f}_B$ is considered, it has been shown that also the parameter E plays an important role since, for a fixed number of samples M, its performance degrades if E is too small or too large. Hence, it could be desirable to adjust also the number of eigenfunctions forming the estimate after seeing the data.

In the following we will follow the SURE approach for tuning the free parameters. Although alternative approaches are possible, such as cross validation and marginal likelihood optimization, we will see that SURE has the advantage to require less communication and computation processing, and also to be suitable for distributed implementations. We start by reporting a result obtained through a simple generalization of the arguments in [44][Section 7.4].

Theorem 9 *Let $\boldsymbol{\eta}$ be a deterministic unknown parameter vector. Assume that the measurements model is*

$$\boldsymbol{z} = \boldsymbol{\eta} + \boldsymbol{e}$$

and consider also future measurements

$$\boldsymbol{z}^* = \boldsymbol{\eta} + \boldsymbol{e}^*$$

where the noises $\boldsymbol{e}$ and $\boldsymbol{e}^*$ are uncorrelated, zero mean with covariance Σ. Then, given the linear estimator $\widehat{\boldsymbol{z}} = S\boldsymbol{z}$, an unbiased estimator of the risk $\mathbb{E}\left[\|\boldsymbol{z}^* - \widehat{\boldsymbol{z}}\|^2\right]$ is given by:

$$\|\boldsymbol{z} - \widehat{\boldsymbol{z}}\|^2 + 2\text{tr}\left(S\Sigma\right). \tag{36}$$

The quantity $\text{tr}\left(S\Sigma\right)$ entering the second part of the objective (36) is connected to the concept of *equivalent degrees of freedom* [52], [53].

In what follows, we assume that γ is unknown but belongs to the finite set Γ which is known in advance to the network. In addition, let us assume that the estimation step has been performed adopting a certain value E. Hence, if $\widehat{f}_A$ has been used, each agent has stored $\frac{G^TG}{M}$ and $\frac{G^T}{M}y$ so that, letting

$$H_A(\gamma) := \left(\frac{G^TG}{M} + \frac{\gamma\sigma_\nu^2}{M}\Lambda_E^{-1}\right)^{-1}\frac{G^T}{M},$$

it can compute H_Ay for any $\gamma \in \Gamma$.

If $\widehat{f}_B$ has been adopted, then also the optimal number of eigenbases E' has to be found within the set $E' \in \Omega$. In this case, each agent knows only $\frac{G^T}{M}y$ and, letting

$$H_B(\gamma, E') := \mathcal{I}_{E'}\left(I + \frac{\gamma\sigma_\nu^2}{M}\Lambda_E^{-1}\right)^{-1}\frac{G^T}{M}, \tag{37}$$

where

$$\mathcal{I}_{E'} := \begin{bmatrix} I_{E'} & \\ & \mathbf{0}_{E-E'} \end{bmatrix}, \tag{38}$$

it can compute H_By for any $\gamma \in \Gamma$ and integer $E' \in \Omega$.

A. Distributed SURE for $\widehat{f}_A$: tuning of γ

The first strategy is suited for $\widehat{f}_A$. Surprisingly, we will see that the tuning of γ can be performed by the network using only local operations, without the need of performing any additional consensus operation. Now, let us reconsider our measurements model

$$\boldsymbol{y} = G\boldsymbol{a} + Z\boldsymbol{b} + \boldsymbol{\nu}$$

where $\boldsymbol{a}$ is E-dimensional. Hereby, we break away from the assumptions on prior correctness by thinking of $G\boldsymbol{a} + Z\boldsymbol{b}$ as a deterministic vector. It thus corresponds to the deterministic function f sampled on the realizations of the input locations.

We then create a (projected) measurement model via premultiplication by G^T/M, i.e.,

$$\underbrace{\frac{G^T\boldsymbol{y}}{M}}_{\boldsymbol{z}} = \underbrace{\frac{G^TG}{M}\boldsymbol{a} + \frac{G^TZ}{M}\boldsymbol{b}}_{\boldsymbol{\eta}} + \underbrace{\frac{G^T\boldsymbol{\nu}}{M}}_{\boldsymbol{e}} \tag{39}$$

where the correspondences with the key quantities defining the risk estimator (36) have been pointed out. From such definitions, we also obtain $\widehat{\boldsymbol{z}} = \frac{G^TG}{M}H_A\boldsymbol{y} = S\boldsymbol{z}$ where

$$S := \frac{G^TG}{M}\left(\frac{G^TG}{M} + \frac{\gamma\sigma_\nu^2}{M}\Lambda_E^{-1}\right)^{-1}$$

and

$$\Sigma = \sigma_\nu^2\frac{G^TG}{M^2}.$$

Recall that the matrix $V = \frac{G^TG}{M} = \frac{1}{M}\sum_{m=1}^M G_m^TG_m$ and the vector $\boldsymbol{z} = \frac{1}{M}\sum_{m=1}^M G_m^Ty_m$ have been already computed by each agent via a distributed consensus algorithm [29] to implement $\widehat{f}_A$. Then, since the network cardinality M is known, each agent can tune γ by optimizing the SURE score (36) connected with the prediction risk on the future data $\boldsymbol{z}^* = \frac{G^T\boldsymbol{y}^*}{M}$, i.e.,

$$\widehat{\gamma}_A = \arg\min_{\gamma\in\Gamma} J_A\left(\gamma\right) \tag{40}$$

with

$$\begin{aligned} J_A(\gamma) := {} & \|(I-S)\boldsymbol{z}\|^2 + 2\mathrm{tr}\,(S\Sigma) \\ = {} & \left\|\frac{\gamma\sigma_\nu^2}{M}\left(V\Lambda_E + \frac{\gamma\sigma_\nu^2}{M}I\right)^{-1}\boldsymbol{z}\right\|^2 + \\ & + \frac{2\sigma_\nu^2}{M}\mathrm{tr}\left(V^2\left(V+\frac{\gamma\sigma_\nu^2}{M}\Lambda_E^{-1}\right)^{-1}\right). \end{aligned}$$

To understand the rationale underlying this strategy we have just to consider that the novel process (39) is formed by E measurements, each corresponding to the projection of the original ones on the space of the sampled eigenfunctions $[\phi_e(x_1)\ \cdots\ \phi_e(x_M)]$. For large M, the quantity $G^T Z\boldsymbol{b}$ vanishes so that $\boldsymbol{\eta} \approx \boldsymbol{a}$. This means that the SURE score becomes an unbiased estimator of those signal components which are expected to capture the most part of the energy.

Remark 10 *Based on the previous analysis, it is straightforward to observe that the SURE strategy described above is not suited for $\widehat{f}_B$. In fact, it requires each agent to know $\frac{G^TG}{M}$. But if this quantity were known, each agent could implement $\widehat{f}_A$, an estimator that has more favorable features than $\widehat{f}_B$.*

B. Distributed SURE for $\widehat{f}_B$: tuning of E and γ

The second strategy is designed for $\widehat{f}_B$. It tunes $\gamma \in \Gamma$ and $E' \in \Omega$ just using an additional average consensus on a vector of size $E\dim(\Omega)\dim(\Gamma)$. Our starting point is still (39), i.e., the E-dimensional projected measurement space, where $\boldsymbol{z} = \frac{G^T\boldsymbol{y}}{M} = \frac{1}{M}\sum_{m=1}^{M} G_m^T y_i \in \mathbb{R}^E$ has been computed to implement $\widehat{f}_B$ via a standard distributed consensus algorithm and is therefore known to each agent. Let us define $\widehat{a}(\gamma, E') = H_B(\gamma, E')\boldsymbol{y} \in \mathbb{R}^E$. Clearly $\widehat{a}(\gamma, E')$ for $E' < E$ is simply the truncated version of $\widehat{a}(\gamma, E)$ where the last $E - E'$ components are set to zero. Moreover, the vectors $\widehat{a}(\gamma, E')$ can be independently computed by each agent for each value of $\gamma \in \Gamma, E' \in \Omega$ once $\boldsymbol{z}$ is available. The output prediction can be written as

$$\widehat{\boldsymbol{z}}(\gamma, E') = \frac{G^TG}{M}\widehat{a}(\gamma, E') = \frac{1}{M}\sum_{m=1}^{M} G_m^T G_m \widehat{a}(\gamma, E') \in \mathbb{R}^E.$$

Hence, each agent can compute the vectors $\widehat{\boldsymbol{z}}(\gamma, E')$ for each $\gamma \in \Gamma$ and $E' \in \Omega$ by running an additional consensus of size $O(\dim(\Omega)\dim(\Gamma)E)$. As so, the first part of the SURE score $\|\boldsymbol{z} - \widehat{\boldsymbol{z}}(\gamma, E')\|^2$ can be readily computed by each agent.

As for the second part of SURE related to the equivalent degrees of freedom, we need to compute $\mathrm{tr}\,(S(\gamma, E')\Sigma)$ where

$$S(\gamma, E') = \frac{G^TG}{M}\mathcal{I}_{E'}\left(I + \frac{\gamma\sigma_\nu^2}{M}\Lambda_E^{-1}\right)^{-1}, \quad \Sigma = \frac{\sigma_\nu^2}{M}\frac{G^TG}{M}.$$

Obviously, this would not make too much sense in the context of $\widehat{f}_B$ since the computation of $V = \frac{G^TG}{M}$ would allow us to compute $\widehat{f}_A$ which has better performance anyways. Therefore we will approximate such matrix V (similarly to what we did to obtain $\widehat{f}_B$) by replacing it with an identity matrix. This corresponds to use a sort of *expected equivalent degrees of freedom*:

$$\begin{aligned} \mathrm{tr}\left(S(\gamma, E')\Sigma\right) & \approx \frac{\sigma_\nu^2}{M}\mathrm{tr}\left(\mathcal{I}_{E'}\left(I + \frac{\gamma\sigma_\nu^2}{M}\Lambda_E^{-1}\right)^{-1}\right) \\ & = \frac{\sigma_\nu^2}{M}\sum_{e=1}^{E'}\frac{\lambda_e}{\lambda_e + \gamma\sigma_\nu^2/M}. \end{aligned}$$

The optimal tuning of the parameter is then obtained as

$$\left(\widehat{\gamma}_B, \widehat{E}_B\right) = \underset{\gamma\in\Gamma, E'\in\Omega}{\operatorname{argmin}}\ J_B(\gamma, E') \tag{41}$$

with

$$J_B(\gamma, E') := \|\boldsymbol{z} - \widehat{\boldsymbol{z}}(\gamma, E')\|^2 + 2\frac{\sigma_\nu^2}{M}\sum_{e=1}^{E'}\frac{\lambda_e}{\lambda_e + \gamma\sigma_\nu^2/M}.$$

Note that this strategy for tuning $\widehat{f}_B$ is more efficient from a communication and computational point of view than $\widehat{f}_A$ only if $\dim(\Omega)\dim(\Gamma) < E$.

C. Practical implementation issues

We now illustrate how to implement the proposed distributed estimators, also in connection with the properties of the SURE tuning strategies described above. We discuss first the use and the derivation of the KL expansion and then how f can be estimated in a distributed way also adopting basis functions different from the kernel eigenfunctions. All the code developed for implementing the algorithms below is publicly available in the repository github.com/damianovar/Gaussian-regression-via-finite-dimensional-approximations.

1) Computing the KL expansions: Assume that the prior on f is correct and that the input locations distribution μ is known. Then, according to Theorem 5, at least for large data set size M, the use of the eigenfunctions in (5) is statistically optimal. Obtaining the kernel expansion in closed form is in general difficult but important exceptions are the popular spline and Gaussian kernel. In particular, for uniform μ the expansion of the linear and cubic smoothing spline kernel is reported in [49]. For Gaussian μ on the real line, the Gaussian kernel expansion is given via Hermite polynomials, as reported in [27][Section 4]. Such result then immediately generalizes to multi-dimensional domains: if $\mu(\cdot)$ and $K(\cdot,\cdot)$ are tensor products of one-dimensional distributions and kernels, respectively, the expansion involves tensor products of the one-dimensional eigenfunctions.

Assume then that the kernel expansion is not available in closed form. It is worth pointing out that in many relevant distributed problems the dimension of the function domain $\mathcal{X}$ is limited to 2 or 3, and this makes the numerical determination of the eigenfunctions and eigenvalues viable. More specifically, let $\{\widetilde{x}_e\}_{e=1}^{q}$ be independent samples from μ, and let $\boldsymbol{K}$ be the $q \times q$ kernel matrix whose (i,j)-entry is

$$[\boldsymbol{K}]_{ij} = K\left(\widetilde{x}_i, \widetilde{x}_j\right), \quad i = 1, \ldots, q, \quad j = 1, \ldots, q \tag{42}$$

Then, according to [50][Lemma 9 and Corollary 10], the eigenvalues and (normalized) eigenvectors from the Singular Values Decomposition (SVD) of $\boldsymbol{K}$ converge to the eigenvalues and eigenfunctions of $K(\cdot,\cdot)$ as $q \to +\infty$. Hence, the agents can be equipped with arbitrarily accurate approximations of the KL expansion.

2) Generic basis functions: Kernel sections: As discussed above, in some circumstances the kernel eigenfunctions could be not available in closed form, or have a complex functional form that makes storing them in the agents' memory unpractical. In such cases, one would rather use basis functions which admit simple closed-form expressions, possibly also non orthonormal. Even if the bounds developed in Section IV cannot be used anymore, we will see that the SURE strategies for hyperparameters tuning generalize well also to this situation.

We limit our discussion to the use of the kernel sections as basis (an important case also in view of their connections with the representer theorem [54], [55]). This basis is associated to a set $\{\widetilde{x}_e\}_{e=1}^E$ of input locations[1] which could be drawn from μ or selected in a deterministic way to cover sufficiently well $\mathcal{X}$. We then define our E basis functions as

$$\phi_1(\cdot) = K\left(\widetilde{x}_1, \cdot\right) \qquad \dots \qquad \phi_E(\cdot) = K\left(\widetilde{x}_E, \cdot\right).$$

Using the kernel sections in the decomposition (7), we can think of f_a as

$$f_a(x) = \sum_{e=1}^{E} a_e K\left(\widetilde{x}_e, x\right)$$

where the vector $\boldsymbol{a} := \left[a_1, \dots, a_E\right]^T$ is now zero-mean Gaussian with covariance proportional to the inverse of the kernel matrix

$$[\boldsymbol{K}]_{ij} = K\left(\widetilde{x}_i, \widetilde{x}_j\right), \quad i = 1, \dots, E, \quad j = 1, \dots, E$$

i.e.,

$$\boldsymbol{a} \sim \mathcal{N}\left(0, \gamma^{-1}\boldsymbol{K}^{-1}\right).$$

In fact, if the prior were correct, this would indeed correspond to see f_a sampled on $\{\widetilde{x}_e\}_{e=1}^E$ as zero-mean Gaussian with covariance $\gamma^{-1}\boldsymbol{K}^{-1}$.

Since the kernel sections are generally not orthonormal w.r.t. μ, even if $M \to \infty$ the projected measurements $\frac{G^T\boldsymbol{y}}{M}$ do not converge to the expansion coefficients a_e. However, these can be still used to tune the regularization parameters. In particular, for what concerns $\widehat{f}_A$, the distributed SURE estimator introduced in Section V-A can estimate f and γ with a single consensus just replacing Λ_E^{-1} with $\boldsymbol{K}$. Thus, this estimator does not even need the knowledge of μ and the agents can implement it once they know the function $K(\cdot, \cdot)$ and the expansion grid $\{\widetilde{x}_e\}_{e=1}^E$. The estimator $\widehat{f}_A$ is thus given by:

$$\widehat{\boldsymbol{a}}(\gamma) := \left(\frac{G^TG}{M} + \frac{\sigma_\nu^2}{M}\boldsymbol{K}\right)^{-1}\frac{G^T\boldsymbol{y}}{M}$$

$$S(\gamma) := \frac{G^TG}{M}\left(\frac{G^TG}{M} + \frac{\gamma\sigma_\nu^2}{M}\boldsymbol{K}\right)^{-1}$$

Consider now the implementation of $\widehat{f}_B$ through the kernel sections with the estimator defined by the set of potential $E' \in \Omega$. In particular, for the sake of simplicity, assume that each E' is associated to the kernel sections induced by the first E' input locations in the (ordered) set $\{\widetilde{x}_e\}_{e=1}^E$. Given a generic matrix A, the submatrix obtained by retaining its first E' rows and columns is denoted by $[A]_{E'}$. Assume moreover that the same notation applies to vectors to retain only their first E' elements. Then, the same SURE strategy developed in Section V-B can be adopted by setting

$$\widehat{\boldsymbol{a}}\left(\gamma, E'\right) = \begin{bmatrix} I_{E'} \\ \mathbf{0} \end{bmatrix}\left(\left[\mathbb{E}\frac{G^TG}{M}\right]_{E'} + \frac{\gamma\sigma_\nu^2}{M}[\boldsymbol{K}]_{E'}\right)^{-1}\left[\frac{G^T\boldsymbol{y}}{M}\right]_{E'}$$

and

$$S(\gamma, E') = \left[\mathbb{E}\frac{G^TG}{M}\right]_{E'}\begin{bmatrix} I_{E'} \\ \mathbf{0} \end{bmatrix}\left(\left[\mathbb{E}\frac{G^TG}{M}\right]_{E'} + \frac{\gamma\sigma_\nu^2}{M}[\boldsymbol{K}]_{E'}\right)^{-1}$$

where, instead of using $\mathcal{I}_{E'}$ defined in (38), we use $I_{E'}$ and $\mathbf{0} \in \mathbb{R}^{(E-E')\times E'}$ to account for the non-diagonal nature of the matrices now at stake, with the trace of S given by the sum of the its (i,i) entries with $i = 1, \dots, E'$, and where $\mathbb{E}\left[\frac{G^TG}{M}\right]$ substitutes I in (37), since the kernel sections are generally not orthonormal. The exact expectation of $\mathbb{E}\left[\frac{G^TG}{M}\right]$ can be explicitly computed in some special cases as in the example below, or can be approximated via its sampled version, i.e., $\mathbb{E}\left[\frac{G^TG}{M}\right] = \mathbb{E}\left[\sum_{m=1}^M \frac{G_m^TG_m}{M}\right] \approx \frac{1}{E}\sum_{e=1}^E \frac{\widetilde{G}_e^T\widetilde{G}_e}{E}$ where $\widetilde{G}_e$ are computed on the input locations $\{\widetilde{x}_e\}_{e=1}^E$ that shall be used for computing this empirical expectation, not to be confused with the set of a-posteriori input locations $\{x_m\}_{m=1}^M$ used in the actual coefficients estimation step.

Example 11 *An interesting case, relevant for many applications, arises when one wants to use the Gaussian kernel and its kernel sections as basis with μ a mixture of Gaussians. In this case $\mathbb{E}\left[\frac{G^TG}{M}\right]$ can be obtained in closed form. In fact, consider first a scalar scenario, i.e., $x \in \mathbb{R}$, with a mixture of Gaussians made of a single component:*

$$K(x, x') = \exp\left(-\frac{(x-x')^2}{\eta}\right), \ \mu \sim \mathcal{N}(\mu_0, a^2).$$

After simple computations, one finds that the (e, e')-entry of $\mathbb{E}\left[\frac{G^TG}{M}\right]$ is

$$\int_{-\infty}^{+\infty} \frac{\exp\left(-\frac{(x-x_e)^2}{\eta} - \frac{(x-x_{e'})^2}{\eta} - \frac{(x-\mu_0)^2}{2a^2}\right)}{\sqrt{2\pi}a}dx =$$

$$= \frac{\sqrt{\eta}}{\sqrt{\eta + 4a^2}}\exp\left(-\frac{\star}{\eta^2 + 4\eta a^2}\right)$$

with $\star = \eta\left(x_e^2 - 2\mu_0 x_e + x_{e'}^2 - 2\mu_0 x_{e'} + 2\mu^2\right) + 2a^2\left(x_e - x_{e'}\right)^2$. In the multivariate case, assume that $K(x, x') = e^{-\frac{\|x-x'\|^2}{\eta}}$ while μ is given by tensor products of one-dimensional Gaussian densities. Then, the result is still available in closed form: $\mathbb{E}\left[\frac{G^TG}{M}\right]$ corresponds to convex combinations of Hadamard products of the matrices obtained in the scalar case.

3) Generic basis functions: Nyström method: The analysis provided in the previous section can also be applied to the popular Nyström method [56], [57]. The idea is to find a basis for f of dimension E which has almost the same performance of the basis composed of $q \gg E$ kernel sections with q an arbitrary number as in Section V-C1. More specifically,

[1] As in (42), the set $\{\widetilde{x}_e\}_{e=1}^q$ is available a priori and has not to be confused with the input locations $\{x_m\}_{m=1}^M$ then visited by the agents.

let $\{\widetilde{x}_n\}_{n=1}^q$ be q input location defined a-priori[2]he Nyström method closely resembles the eigenfunctions/eigenvalues numerical computation method presented in Section V-C1, the difference being that in Nyström the parameter q is in general not extremely large and the input locations $\{\widetilde{x}_n\}_{n=1}^q$ are not generated by μ but are randomly extracted from the training set. from μ, and consider both the corresponding kernel matrix $\boldsymbol{K} \in \mathbb{R}^{q \times q}$ defined in (42) and its SVD decomposition $\boldsymbol{K} = V D_q V^T$ with $V := [v_1, \dots, v_q]$ the orthonormal eigenvectors of $\boldsymbol{K}$ and D_q the diagonal matrix formed by the corresponding eigenvalues of $\boldsymbol{K}$ sorted in non-increasing order. If $V_E := [v_1, \dots, v_E]$ and $D_E \in \mathbb{R}^{E \times E}$ is the diagonal matrix with the first E sorted eigenvalues of $\boldsymbol{K}$, then $\boldsymbol{K}_E = V_E D_E V_E^T$ is the best rank-E approximation of $\boldsymbol{K}$. The a priori basis

$$\phi_e(x) := \sum_{n=1}^{q} v_e(n) K(\widetilde{x}_n, x), \quad e = 1, \dots, E$$

with $v_e(n)$ the n-th element of the vector v_e, can then be used to define

$$f_a(x) = \sum_{e=1}^{E} a_e \phi_e(x)$$

where $\boldsymbol{a} := [a_1, \dots, a_E]^T$ is a zero-mean Gaussian vector with

$$\boldsymbol{a} \sim \mathcal{N}\left(0, \gamma^{-1}\left(V_E^T \boldsymbol{K} V_E\right)^{-1}\right) = \mathcal{N}\left(0, \gamma^{-1} D_E^{-1}\right).$$

We can then use once again (13) to build G using the ϕ_e's above, and exploit the same strategies considered in Section V-C2 just replacing $\boldsymbol{K}$ with D_E.

D. Numerical study on synthetic data

Let us consider the same data generators based on the spline and the exponentially decaying kernels described in Section IV-C. The unknown function has to be reconstructed from $M = 10000$ measurements by $\widehat{f}_A$ and $\widehat{f}_B$. The errors are still the MSEs defined in (25) and (26) normalized by the prior variance (the same definition was used to build Figure 1). The difference however is that our estimators now depend on unknown hyperparameters that need to be inferred from data. More specifically, when $\widehat{f}_A$ is adopted we fix $E = 400$ and the regularization parameter is searched over a grid Γ containing 50 logarithmically spaced values between 10^{-3} and 10^3. When using $\widehat{f}_B$ the grid Γ contains only the three values $\{10^{-3}, 0, 10^3\}$ while E is estimated from data over $\Omega = \{1, 5, 10, 20, 50, 100, 200, 300, 400\}$. We still consider a Monte Carlo study of 1000 runs where at any run independent realizations of f, of the M input locations and of the measurement noises are generated. Hyperparameters tuning is then performed by:

- "$\widehat{f}_A+$ oracle" and "$\widehat{f}_B+$ oracle", where "oracle" indicates that these approaches know at any run the realization of f (which is the object to estimate) and select exactly those hyperparameters that minimize the MSE achievable by those two estimators. For instance, assume that $f = \sum_{e=1}^{\infty} a_e \phi_e$ is the realization of the function at a certain run. Let also $\widehat{a}(\gamma, E')$ denote the vector with the estimates of the first E' coefficients a_e returned by $\widehat{f}_B$. Then $\widehat{f}_B+$ oracle determines the hyperparameters as
$$\left(\widehat{\gamma}, \widehat{E}\right) := \arg \min_{\gamma \in \Gamma, E' \in \Omega} \sum_{e=1}^{\infty} \left(a_e - \widehat{a}_e(\gamma, E')\right)^2$$
where $\widehat{a}_e(\gamma, E') := 0$ for $e > E'$. Thus, this estimator is not implementable in practice and provides the lower bound on the MSEs (25) and (26) achievable by the two estimators;
- "$\widehat{f}_A+$ SURE" and "$\widehat{f}_B+$ SURE", where the hyperparameters tuning step is performed following the SURE approaches described in the previous subsection. Recall that "$\widehat{f}_A+$ SURE" requires only a single consensus on a vector of size $O(E^2)$ to obtain simultaneously both the hyperparameters and function estimates, while "$\widehat{f}_B+$ SURE" requires two consensus operations of size $O(E)$.

Data set size	M = 100	M = 1000	M = 10000
$\mathcal{S}_p$	0.93	0.987	0.99

Table I
SURE's performance index $\mathcal{S}_p$ summarizing four Monte Carlo studies as a function of the number M of available measurements. A value of $\mathcal{S}_p$ close to 1 indicates that SURE's performance is close to that of the oracle. For $M = 10000$, $\mathcal{S}_p$ represents a resume of the entire Figure 2.

Figure 2 compares with a scatter-plot the various (normalized) MSEs obtained by the oracle- and SURE-based approaches as a function of the Monte Carlo run. Remarkably, SURE's performance (dashed lines) is very close to that of the oracle (solid lines). When using "$\widehat{f}_B+$ SURE" (right panels) the curves are in practice indistinguishable.

The set of four Monte Carlo experiments have been also repeated adopting different data set sizes M. To synthesize SURE's performance with an index function only of M, let $\mathcal{S}_p \in [0, 1]$ denote the ratio between the mean of the 400 errors obtained by the oracle and the SURE strategies respectively for a certain value of M. Note that a value of $\mathcal{S}_p = 1$ indicates that SURE is performing as well as the oracle and that, for $M = 10000$, $\mathcal{S}_p$ becomes the distillate of Figure 2. Table I reports $\mathcal{S}_p$ for $M = 100, 1000, 10000$: one can see that the proposed hyperparameter estimation procedure behaves very nicely.

E. Numerical study on field data – Colorado rain

Let us now consider the reconstruction of monthly precipitations using data collected in Colorado in the years 1995-1997 [58]. Many alternative solutions are available in the context of weather forecasts, but they are limited to centralized solutions, such as [5], [6], for example, thus not suitable in our distributed framework. Measurements y_m are a series of monthly average precipitations measured at 367 stations within the rectangular longitude/latitude region

[2] T

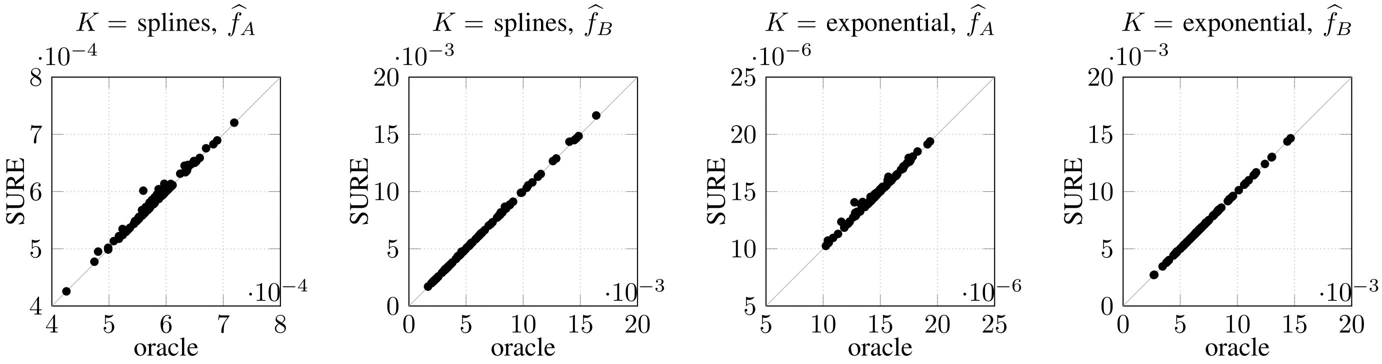

Figure 2. Comparison of the MSE indexes obtained by the SURE- and oracle-based strategies. Each circle corresponds to the result of a certain Monte Carlo run (the x-axis being associated to oracle-based estimators, and the y-axis to SURE-based ones). The fact that the circles groups are close to the bisector of the first quadrant indicates that the performance of SURE is almost equivalent to that of the oracle.

$[-109.5, -101] \times [36.5, 41.5]$ remapped for convenience into the unitary square so that $x_m \in [0, 1]^2$ for every m.

We test the SURE strategies (40) and (41). When using $\widehat{f}_A$ we set $E = 20$ and Γ to the grid containing 50 values logarithmically spaced between 10^{-5} and 10^5. When adopting $\widehat{f}_B$ we use $\Gamma = 0$ and $\Omega = \{2, 4, \ldots, 20\}$, i.e., consider only E' as a regularization parameter. In both cases, we consider the Gaussian kernel

$$K(x, x') = \exp\left(-10\|x - x'\|^2\right).$$

The eigenfunctions are computed by assuming that the input locations are not know a-priori and extracted uniformly from the monitored region, i.e., $\mu(x)$ is a uniform distribution. We design a Monte Carlo study of 1000 runs where, at any run, we select randomly two months within the 1995-1997 dataset obtaining three different sets. The first one is a training set $\mathcal{D}_{\text{train}}$ of average precipitations obtained by selecting randomly and uniformly 2/3 of the measurements from the first selected month. The second is a test set $\mathcal{D}_{\text{test}}$ corresponding to the remaining 1/3 measurements from the first selected month. The last one is $\mathcal{D}_{\sigma^2_\nu}$ and contains measurements in the second selected month which are used to estimate the noise variance via least squares based on E eigenfunctions. This corresponds to using $\widehat{f}_A$ with $\gamma = 0$ obtaining as estimate of the noise variance

$$\widehat{\sigma}^2_\nu = \frac{1}{\dim(\mathcal{D}_{\sigma^2_\nu}) - E} \sum_{m=1}^{\dim(\mathcal{D}_{\sigma^2_\nu})} \left(\widehat{f}_A\left(x_m; 0\right) - y_m\right)^2$$

where $\dim(\mathcal{D}_{\sigma^2_\nu})$ is the cardinality of $\mathcal{D}_{\sigma^2_\nu}$. Overall, this represents a situation where noise levels are determined by a centralized approach before running the estimators $\widehat{f}_A$ and $\widehat{f}_B$.

The following tuning strategies are used:

1) "$\widehat{f}_A$+ oracle" and "$\widehat{f}_B$+ oracle", where "oracle" now indicates that these approaches can select those hyperparameters minimizing the following prediction errors on the test set

$$\text{RSS}_A\left(\gamma\right) := \frac{1}{\dim(\mathcal{D}_{\text{test}})} \sum_{m=1}^{\dim(\mathcal{D}_{\text{test}})} \left(\widehat{f}_A\left(x_m; \gamma\right) - y_m\right)^2 \tag{43}$$

$$\text{RSS}_B\left(\gamma, E'\right) := \frac{1}{\dim(\mathcal{D}_{\text{test}})} \sum_{m=1}^{\dim(\mathcal{D}_{\text{test}})} \left(\widehat{f}_B\left(x_m; \gamma, E'\right) - y_m\right)^2 \tag{44}$$

where (x_m, y_m) are all elements of $\mathcal{D}_{\text{test}}$. Note that RSS_A and RSS_B can be seen as approximations of the MSEs (25) and (26) and that the oracle provides a lower bound on their values;

2) "$\widehat{f}_A$+ SURE" and "$\widehat{f}_B$+ SURE", where the hyperparameters tuning step is performed minimizing the estimated risks $J_A\left(\gamma\right)$ and $J_B\left(\gamma, E'\right)$ defined in (40) and (41).

Figure 3 compares with a scatter-plot the prediction errors (43) and (44) achieved by the estimators after the 1000 Monte Carlo runs. The situation is not dissimilar from the case of synthetic data: as for the estimators "$\widehat{f}_A$+ SURE" and "$\widehat{f}_B$+ SURE", the performance of the SURE strategies is close to that of the oracles.

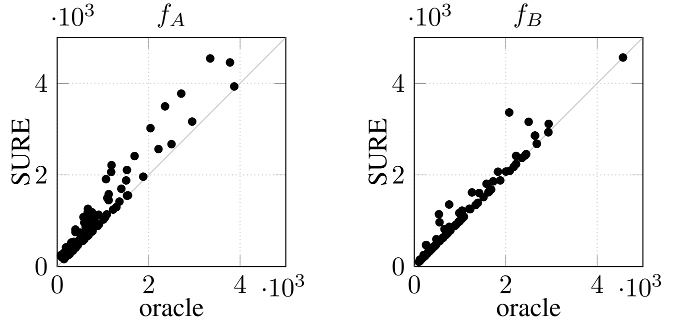

Figure 3. Comparison of the RSS prediction error indexes obtained by the oracle- and SURE-based strategies. Each opaque circle corresponds to the result of one of the 1000 Monte Carlo runs. The closer the circles are to the bisector of the first quadrant performance means that the closer the performance of that SURE-based or Nyström-SURE estimator is to the ones of the oracle-based estimator.

Specifically considering the SURE-based strategies, Figure 4 also compares the estimated risks $J_A\left(\gamma\right)$ and $J_B\left(\gamma, E'\right)$ against the indexes $\text{RSS}_A\left(\gamma\right)$ and $\text{RSS}_B\left(\gamma, E'\right)$ in (43) and (44) in the first Monte Carlo run. The curves show that hyperparameters values have a major effect on the estimation performance and that our SURE approach leads to a good

	CCPP		CPU	
	$\widehat{f}_A$	$\widehat{f}_B$	$\widehat{f}_A$	$\widehat{f}_B$
fit oracle	99.3	73.3	75.7	76.1
fit SURE	99.3	73.3	71.2	66.5
γ oracle	10^{-3}	-	10^{-4}	-
γ SURE	10^{-3}	-	10^{-5}	-
E' oracle	-	2.00	-	10.00
E' SURE	-	2.00	-	27.00

Table II

Summary of the performance of the proposed parameters calibration strategies against oracles for different publicly available datasets. "CCPP" indicates the UCI Combined Cycle Power Plant regression dataset, with 9568 instances and a 4-dimensional input domain $\mathcal{X}$. "CPU" indicates the UCI Computer hardware regression dataset, with 209 instances and a 6-dimensional input domain $\mathcal{X}$ (corresponding to the quantitative features available in the dataset). For each dataset 1/3 of the data has been used for test purposes.

regularization tuning. The related function estimates are visible in Figure 5.

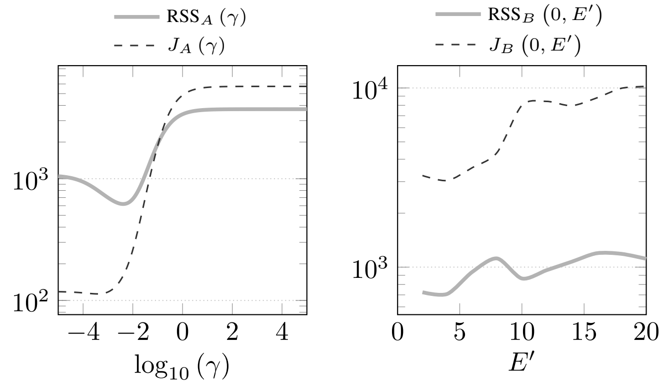

Figure 4. Comparison of the predictive performance of the estimators $\widehat{f}_A$ (left panel) and $\widehat{f}_B$ (right panel) over the test set in Figure 5 for $\gamma \in \Gamma$ and $E' \in \Omega$ against the SURE scores $J_A(\gamma)$ and $J_B(\gamma)$ in the first Monte Carlo run.

F. Numerical study on field data – UCI datasets

The second study is performed on two datasets from the public UCI repository, and is executed using the Nyström-based strategy described in Section V-C3 to compute the basis functions for the estimators using all the input locations that define the training set. Our purpose is here to compare the proposed SURE-based strategy for tuning the regularization parameters against an oracle that selects as the best regularization parameters that ones that give the best fit performance in the test set. As for the kernel, we consider a Gaussian kernel with an unitary variance (not accurately tuned, since the purpose of this section is more checking that the proposed SURE strategy chooses the regularization parameters accurately rather than actually maximizing the generalization capabilities of the regression algorithms). As for the grid for tuning γ, we then consider the set $\Gamma = \{0, 10^{-5}, 10^{-4}, \dots, 10^{2}\}$; as for E', we consider $\Omega = \{1, 2, \dots, 30\}$. Table II summarizes then the obtained results, and numerically confirms the efficacy of the proposed parameters tuning strategies.

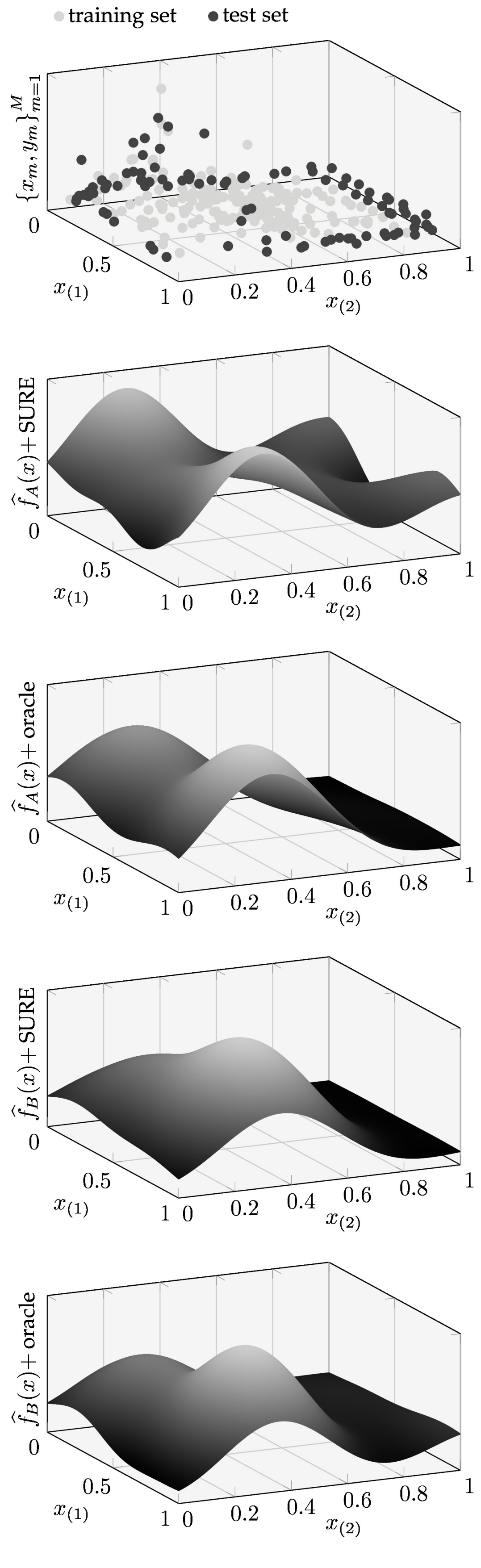

Figure 5. Visualization of the training and test sets (top panel, respectively 173 and 87 samples), and of the estimates returned by "$\widehat{f}_A+$ SURE", "$\widehat{f}_A+$ oracle", "$\widehat{f}_B+$ SURE", and "$\widehat{f}_B+$ oracle" in the first Monte Carlo run.

VI. Conclusions

Distributed function estimation is an important problem where agents with limited computational, data storage and communication capabilities collect noisy measurements and have to reconstruct an unknown map in a collaborative way. In this context, we have studied Gaussian regression providing rigorous statistical bounds on the performance of two distributed estimators, also characterizing their asymptotic behavior. On the practical side, our study indicates how the dimension E of the adopted estimator has to depend on the number of measurements M collected by the agents to guarantee the desired statistical performance. The analysis clarifies merits and limitations of the two approaches also in function of the different amount of information exchange required to the network (linear or quadratic in E). We have also introduced novel distributed strategies which learn from data possibly unknown hyperparameters entering the estimators, and that do not necessarily require solving potentially numerically intensive eigenfunctions-eigenvalues decompositions of kernel functions. For the first time, to our knowledge, this paper has shown how it is possible to estimate the regularization parameter and the unknown function via a single average consensus operation.

Overall, the theoretical achievements and the numerical strategies here described provide sound tools to reconstruct static functions in distributed networks. An important future direction is to extend all the analysis to an even more challenging situation where the unknown map may change in time and has to be tracked in an on-line manner.

Appendix

A. Preliminary results

The following result will be especially useful in what follows. In fact, it will be often used to obtain bounds on intricate conditional expectations just calculating unconditional means.

Lemma 12 *Let Ω denote a sample space, $\omega \in \Omega$ its generic element. Let $\mathcal{E}$ be an event such that*

$$\mathbb{P}\left[\omega \in \mathcal{E}\right] \geq 1 - \alpha. \tag{45}$$

If $g(\omega)$ is positive and (45) *holds then*

$$\mathbb{E}\left[g(\omega) \ \mid \omega \in \mathcal{E}\right] \leq \frac{1}{1-\alpha}\mathbb{E}\left[g(\omega)\right].$$

Proof of Lemma 12: Let η be the probability measure on the σ-algebra Ω is equipped with. In general, for every $\mathcal{E}'$,

$$\begin{aligned}\mathbb{P}\left[\omega \in \mathcal{E}' \left| \omega \in \mathcal{E}\right.\right] &= \frac{\mathbb{P}\left[\omega \in \mathcal{E}' \cap \mathcal{E}\right]}{\mathbb{P}\left[\omega \in \mathcal{E}\right]} \\ &\leq \frac{\mathbb{P}\left[\omega \in \mathcal{E}'\right]}{\mathbb{P}\left[\omega \in \mathcal{E}\right]} \leq \frac{1}{1-\alpha}\mathbb{P}\left[\omega \in \mathcal{E}'\right].\end{aligned}$$

If $\eta_{\mathcal{E}}$ denotes the probability measure η conditional on $\mathcal{E}$, one then has

$$\begin{aligned}\int_{\mathcal{E}} g(\omega) d\eta_{\mathcal{E}}(\omega) &\leq \frac{1}{1-\alpha}\int_{\mathcal{E}} g(\omega) d\eta(\omega) \\ &\leq \frac{1}{1-\alpha}\int_{\Omega} g(\omega) d\eta(\omega).\end{aligned}$$

■

The following result exploits the Chernoff bound and will be important to obtain Bnd_A and Bnd_B. It will also clarify the role played by the ε entering the bounds.

Lemma 13 *Let $\alpha \in (0,1)$ be a desired confidence level (e.g., 0.01 or 0.05), and $\varepsilon \in (0,1]$ represent a given distance index for λ_{min} and λ_{max} as specified in* (46) *and* (47)*. If E, M and k in* (10) *satisfy* (27) *then*

$$\mathbb{P}\left[\lambda_{min}\left(\frac{G^TG}{M}\right) \geq \varepsilon\right] \geq 1-\alpha. \tag{46}$$

If instead E, M and k satisfy (30) *then*

$$\mathbb{P}\left[\lambda_{min}\left(\frac{G^TG}{M}\right) \geq \varepsilon \ \cap \ \lambda_{max}\left(\frac{G^TG}{M}\right) \leq 2-\varepsilon\right] \geq 1-\alpha. \tag{47}$$

Proof of Lemma 13: Since the assumptions in [59, Thm. 1.1] are satisfied, for any $\varepsilon \in (0,1]$ one has

$$\mathbb{P}\left[\lambda_{\min}\left(\frac{G^TG}{M}\right) \leq \varepsilon\right] \leq E\left(\frac{e^{-(1-\varepsilon)}}{\varepsilon^{\varepsilon}}\right)^{\overline{Ek}^{\,M}}. \tag{48}$$

Condition (27) is obtained by picking α larger than the Right Hand Side (RHS) of (48) and manipulating this inequality. Then, (46) follows from (48) just considering that if $\overline{\star}$ is the complementary of $\star$ then $\mathbb{P}\left[\star\right] \leq \alpha \Leftrightarrow \mathbb{P}\left[\overline{\star}\right] \geq 1-\alpha$. Now, we can use again [59, Thm. 1.1] to claim that, for every $\varepsilon \in [0,1]$,

$$\mathbb{P}\left[\lambda_{\max}\left(\frac{G^TG}{M}\right) \geq 2-\varepsilon\right] \leq E\left(\frac{e^{(1-\varepsilon)}}{(2-\varepsilon)^{(2-\varepsilon)}}\right)^{\overline{Ek}^{\,M}}. \tag{49}$$

Let the arguments in the $\mathbb{P}\left[\cdot\right]$ in the LHS of (48) and (49) be respectively $\star_{\lambda_{\min}}$ and $\star_{\lambda_{\max}}$. Let also the RHSs of (48) and (49) be upper bounded respectively by $\alpha_{\lambda_{\min}}$ and $\alpha_{\lambda_{\max}}$. Then, it follows that

$$\begin{aligned}\mathbb{P}\left[\star_{\lambda_{\min}} \cup \star_{\lambda_{\max}}\right] &\leq \mathbb{P}\left[\star_{\lambda_{\min}}\right] + \mathbb{P}\left[\star_{\lambda_{\max}}\right] \\ &\leq \alpha_{\lambda_{\min}} + \alpha_{\lambda_{\max}} \leq 2\alpha_{\lambda_{\min}}\end{aligned}$$

with the last inequality following from the fact that $\varepsilon \in [0,1] \implies \alpha_{\lambda_{\min}} \geq \alpha_{\lambda_{\max}}$ since

$$\frac{e^{-(1-\varepsilon)}}{\varepsilon^{\varepsilon}} \geq \frac{e^{(1-\varepsilon)}}{(2-\varepsilon)^{(2-\varepsilon)}}.$$

Thus, letting the novel α be $2\alpha_{\lambda_{\min}}$ (i.e., assuming (30) to be satisfied), we obtain

$$\mathbb{P}\left[\overline{\star}_{\min} \cap \overline{\star}_{\max}\right] \geq 1-\alpha,$$

and this proves (47).

■

The following lemma is just a generalization of the fact that convergence in mean (L_1-norm) of random variables implies convergence in probability. The proof is simple and therefore omitted.

Lemma 14 *Let $g(\boldsymbol{x})$ denote a stochastic variable whose randomness derives from the input locations $\boldsymbol{x} := [x_1, \dots, x_M]^T$. Assume that $g(\boldsymbol{x}) \geq q$ almost surely with q independent of M. In addition, assume also that for any $1-\alpha$ and $\varepsilon > 0$ there exists M_0 such that $\forall M \geq M_0$ one has*

$$\overline{g(\boldsymbol{x})} \leq q + \varepsilon \quad \text{with probability} \quad 1-\alpha,$$

in accordance with Definition 3. Then

$$\lim_{M \to +\infty} g(\boldsymbol{x}) = q \qquad \text{in probability.}$$

■

B. Proof of Theorem 4

We start by computing the general expression for $\mathrm{Err}_A(\boldsymbol{x})$ in (21), then evaluate its expectation like in (23), and finally transport the results to the case of $\mathrm{Err}_B(\boldsymbol{x})$.

As for finding the general expression for $\mathrm{Err}_A(\boldsymbol{x})$, we recall the decomposition of the estimand as $f = f_a + f_b$ in (7), the definition of $\mathcal{S}$ in (9) and the design requirement $\widehat{f}_A, \widehat{f}_B \in \mathcal{S}$, that imply $f_a, \widehat{f}_A \in \mathcal{S}$ and $f_b \in \mathcal{S}^{\perp}$. By construction, then, $\|f\|^2 = \|f_a\|^2 + \|f_b\|^2$ and

$$\mathbb{E}\left[\left\|f - \widehat{f}_A\right\|^2 \mid \boldsymbol{x}\right] = \mathbb{E}\left[\left\|f_a - \widehat{f}_A\right\|^2 \mid \boldsymbol{x}\right] + \|f_b\|^2 \tag{50}$$

where the expectations are w.r.t. the noises $\boldsymbol{\nu}$, so that $\mathbb{E}\left[\|f_b\|^2 \mid \boldsymbol{x}\right] = \|f_b\|^2$ since $\boldsymbol{\nu}, f_b$ and $\boldsymbol{x}$ are all mutually independent. Notice that a similar decomposition holds also for $\widehat{f}_B$.

As for $\mathbb{E}\left[\left\|f_a - \widehat{f}_A\right\|^2 \mid \boldsymbol{x}\right]$ in (50), we notice that $\left\|\widehat{f}_A\right\|^2 = \|\widehat{\boldsymbol{a}}\|_2^2 = \|H_A \boldsymbol{y}\|_2^2$. Since (15) implies

$$\widehat{\boldsymbol{a}} = H_A\left(G\boldsymbol{a} + Z\boldsymbol{b} + \boldsymbol{\nu}\right),$$

and since both $\boldsymbol{a} \perp \boldsymbol{b}$ and $\boldsymbol{\nu} \perp \boldsymbol{b}$, it eventually follows that

$$\begin{aligned} \mathrm{Err}_A(\boldsymbol{x}) = \quad & \mathbb{E}\left[\|\boldsymbol{a} - H_A(G\boldsymbol{a} + \boldsymbol{\nu})\|_2^2 \mid \boldsymbol{x}\right] \\ & + \mathbb{E}\left[\|H_A Z\boldsymbol{b}\|_2^2 \mid \boldsymbol{x}\right] \\ & + \|f_b\|^2 . \end{aligned} \tag{51}$$

Proof of equations (28) **and** (29)**:**

Let $\overline{\mathcal{E}}$ be the event

$$\overline{\mathcal{E}} := \left\{\lambda_{\min}\left(\frac{G^T G}{M}\right) \geq \varepsilon\right\}, \tag{52}$$

and assume ε, α, M and E satisfy (27). Since in this case we can apply Theorem 13, it holds that $\mathbb{P}\left[\overline{\mathcal{E}}\right] \geq 1 - \alpha$.

We can now write the LHS of (23) as

$$\mathbb{E}\left[\mathrm{Err}_A(\boldsymbol{x}) \mid \overline{\mathcal{E}}\right] = \mathbb{E}\left[\mathbb{E}\left[\left\|f - \widehat{f}_A\right\|^2 \mid \boldsymbol{x}\right] \mid \overline{\mathcal{E}}\right].$$

Since $f_b \perp \overline{\mathcal{E}}$, (51) implies

$$\begin{aligned} \mathbb{E}\left[\mathrm{Err}_A(\boldsymbol{x}) \mid \overline{\mathcal{E}}\right] = & \\ & \mathbb{E}\left[\mathbb{E}\left[\|\boldsymbol{a} - H_A(G\boldsymbol{a} + \boldsymbol{\nu})\|^2 \mid \boldsymbol{x}\right] \mid \overline{\mathcal{E}}\right] \\ & + \mathbb{E}\left[\mathbb{E}\left[\|H_A Z\boldsymbol{b}\|^2 \mid \boldsymbol{x}\right] \mid \overline{\mathcal{E}}\right] \\ & + \mathbb{E}\left[\|f_b\|^2\right]. \end{aligned} \tag{53}$$

As for $\mathbb{E}\left[\|f_b\|^2\right]$, we know from (7), (8b) and the mutual independence of the b_e's, that

$$\mathbb{E}\left[\|f_b\|^2\right] = \sum_{e=E+1}^{+\infty} \lambda_e . \tag{54}$$

This term is thus an approximation error influenced only by the dimension E of our search space $\mathcal{S}$.

Given (54), what we actually need to bound is the first two terms in the RHS of (53). We perform this task in the next two subsections.

Bounding $\mathbb{E}\left[\mathbb{E}\left[\|H_A Z\boldsymbol{b}\|^2 \mid \boldsymbol{x}\right] \mid \overline{\mathcal{E}}\right]$ in (53)

Exploiting the nature of the event $\overline{\mathcal{E}}$ to bound H_A in (17), it is not difficult to prove that

$$\mathbb{E}\left[\mathbb{E}\left[\|H_A Z\boldsymbol{b}\|^2 \mid \boldsymbol{x}\right] \mid \overline{\mathcal{E}}\right] \leq \mathbb{E}\left[\left\|\left(\varepsilon I_E + \frac{\sigma_\nu^2}{M}\Lambda_E^{-1}\right)^{-1} \frac{G^T Z}{M}\boldsymbol{b}\right\|^2 \mid \overline{\mathcal{E}}\right]. \tag{55}$$

Defining

$$d_e := \frac{\varepsilon M \lambda_e + \sigma_\nu^2}{M \lambda_e}, \qquad e = 1, \dots, E, \tag{56}$$

it follows that

$$\left(\varepsilon I_E + \frac{\sigma_\nu^2}{M}\Lambda_E^{-1}\right)^{-1} = \operatorname{diag}\left(d_1^{-1}, \dots, d_E^{-1}\right). \tag{57}$$

Consider moreover that from the definition of f_b in (7), of $\boldsymbol{b}$ in (12) and of Z in (14) it follows that $[Z\boldsymbol{b}]_m = f_b\left(x_m\right)$. Let then

$$c_e := \left[G^T Z\boldsymbol{b}\right]_e = \sum_{m=1}^{M} \phi_e\left(x_m\right) f_b\left(x_m\right) \qquad e = 1, \dots, E \tag{58}$$

so that

$$\boldsymbol{b}^T Z^T G \left(\varepsilon I_E + \frac{\sigma_\nu^2}{M}\Lambda_E^{-1}\right)^{-2} G^T Z\boldsymbol{b} = \sum_{e=1}^{E} \frac{c_e^2}{d_e^2} . \tag{59}$$

Combining (57) and (59), and considering that the d_e's are deterministic, we can thus rewrite (55) as

$$\begin{aligned} \mathbb{E}\left[\mathbb{E}\left[\|H_A Z\boldsymbol{b}\|^2 \mid \boldsymbol{x}\right] \mid \overline{\mathcal{E}}\right] & \leq \frac{1}{M^2} \sum_{e=1}^{E} \frac{\mathbb{E}\left[c_e^2 \mid \overline{\mathcal{E}}\right]}{d_e^2} \\ & \leq \frac{1}{(1-\alpha)M^2} \sum_{e=1}^{E} \frac{\mathbb{E}\left[c_e^2\right]}{d_e^2}, \end{aligned} \tag{60}$$

where in the last inequality we applied Lemma 12. In view of the definition of the c_e's in (58) and the linearity of $\mathbb{E}\left[\cdot\right]$, one has

$$\begin{aligned} \mathbb{E}\left[c_e^2\right] = \quad & \sum_{m=1}^{M} \mathbb{E}\left[\phi_e^2\left(x_m\right) f_b^2\left(x_m\right)\right] \\ & + \sum_{m \neq m'} \mathbb{E}\left[\phi_e\left(x_m\right) \phi_e\left(x_{m'}\right) f_b\left(x_m\right) f_b\left(x_{m'}\right)\right]. \end{aligned} \tag{61}$$

As for the first term in the RHS of (61), combining (8b) with bound (10), one immediately has

$$\mathbb{E}\left[\phi_e^2\left(x_m\right) f_b^2\left(x_m\right)\right] \leq k \sum_{e=E+1}^{+\infty} \lambda_e.$$

As for the second term in the RHS of (61), due to the independence of the $\{x_m\}_{m=1}^M$ we know that

$$\mathbb{E}\left[\phi_e\left(x_m\right) \phi_e\left(x_{m'}\right) f_b\left(x_m\right) f_b\left(x_{m'}\right) \mid f_b\right] = \mathbb{E}\left[\phi_e\left(x_m\right) f_b\left(x_m\right) \mid f_b\right] \mathbb{E}\left[\phi_e\left(x_{m'}\right) f_b\left(x_{m'}\right) \mid f_b\right].$$

Moreover, since $e = 1, \dots, E$, from the definition of f_b in (7) it comes that

$$\mathbb{E}\left[\phi_e\left(x_m\right) f_b\left(x_m\right) \mid f_b\right] = 0.$$

Combining the two results, one has

$$\mathbb{E}\left[c_e^2\right] \leq kM \sum_{e=E+1}^{+\infty} \lambda_e \quad e = 1, \dots, E. \tag{62}$$

Finally, combining (56), (60), (62) and Lemma 12 one obtains

$$\mathbb{E}\left[\mathbb{E}\left[\|H_A Z \boldsymbol{b}\|^2 \mid \boldsymbol{x}\right] \mid \overline{\mathcal{E}}\right] \leq \\ \leq \frac{kM}{1-\alpha}\left(\sum_{e=1}^{E} \frac{\lambda_e^2}{(\varepsilon M \lambda_e + \sigma_\nu^2)^2}\right)\left(\sum_{e=E+1}^{+\infty} \lambda_e\right). \tag{63}$$

Bounding $\mathbb{E}\left[\mathbb{E}\left[\|\boldsymbol{a} - H_A(G\boldsymbol{a} + \boldsymbol{\nu})\|^2 \mid \boldsymbol{x}\right] \mid \overline{\mathcal{E}}\right]$ *in* (53)

To characterize $\|\boldsymbol{a} - H_A(G\boldsymbol{a} + \boldsymbol{\nu})\|^2$, note that this term corresponds to the MSE of a classical MAP estimator for a standard linear and finite-dimensional Gaussian model (where the term $\boldsymbol{b}$ is not involved). More precisely, if the measurements models conditional on $\boldsymbol{x}$ were

$$\boldsymbol{y} = G\boldsymbol{a} + \nu, \quad \boldsymbol{a} \sim \mathcal{N}(0, \Lambda), \quad v \sim \mathcal{N}(0, \sigma_\nu^2) \tag{64}$$

with $\boldsymbol{a}$ independent of ν, the optimal estimator would indeed be

$$\widehat{a}_A = H_A \boldsymbol{y} = H_A(G\boldsymbol{a} + \boldsymbol{\nu}). \tag{65}$$

Exploiting standard results on Gaussian estimation, the covariance matrix of the error is

$$\operatorname{var}\left(\boldsymbol{a} - H_A(G\boldsymbol{a} + \boldsymbol{\nu}) \mid \boldsymbol{x}\right) = \frac{\sigma_\nu^2}{M}\left(\frac{G^T G}{M} + \frac{\sigma_\nu^2}{M}\Lambda_E^{-1}\right)^{-1}. \tag{66}$$

Applying Lemma 12 and using (52) to bound H_A in (17), from definitions (56) and (57) it follows that

$$\mathbb{E}\left[\mathbb{E}\left[\|\boldsymbol{a} - H_A(G\boldsymbol{a} + \boldsymbol{\nu})\|^2 \mid \boldsymbol{x}\right] \mid \overline{\mathcal{E}}\right] \leq \frac{\sigma_\nu^2}{1-\alpha}\left(\sum_{e=1}^{E} \frac{\lambda_e}{\varepsilon M \lambda_e + \sigma_\nu^2}\right). \tag{67}$$

Hence, the bound on $\mathrm{Err}_A(\boldsymbol{x})$ is obtained by combining (54), (63) and (67). ■

Proof of equations (31) **and** (32)**:**

Let $\overline{\mathcal{E}}$ be now the event

$$\overline{\mathcal{E}} := \left\{\lambda_{\min}\left(\frac{G^T G}{M}\right) \geq \varepsilon \cap \lambda_{\max}\left(\frac{G^T G}{M}\right) \leq 2 - \varepsilon\right\},$$

and assume that ε, α, M and E satisfy (30). Substituting H_A with H_B in the derivation of (53) one obtains

$$\begin{aligned} \mathbb{E}\left[\mathrm{Err}_B(\boldsymbol{x}) \mid \overline{\mathcal{E}}\right] & \\ &= \mathbb{E}\left[\mathbb{E}\left[\|\boldsymbol{a} - H_B(G\boldsymbol{a} + \boldsymbol{\nu})\|^2 \mid \boldsymbol{x}\right] \mid \overline{\mathcal{E}}\right] \\ &+\mathbb{E}\left[\mathbb{E}\left[\|H_B Z \boldsymbol{b}\|^2 \mid \boldsymbol{x}\right] \mid \overline{\mathcal{E}}\right] \\ &+\mathbb{E}\left[\|f_b\|^2\right] \end{aligned} \tag{68}$$

We already know that $\mathbb{E}\left[\|f_b\|^2\right] = \sum_{e=E+1}^{+\infty} \lambda_e$. Hence, we have to bound the first two terms in the RHS of (68).

Bounding $\mathbb{E}\left[\mathbb{E}\left[\|H_B Z \boldsymbol{b}\|^2 \mid \boldsymbol{x}\right] \mid \overline{\mathcal{E}}\right]$ *in* (68)

From the definition of H_B in (20), one has

$$\mathbb{E}\left[\mathbb{E}\left[\|H_B Z \boldsymbol{b}\|^2 \mid \boldsymbol{x}\right] \mid \overline{\mathcal{E}}\right] \leq \mathbb{E}\left[\left\|\left(I_E + \frac{\sigma_\nu^2}{M}\Lambda_E^{-1}\right)^{-1} \frac{G^T Z}{M}\boldsymbol{b}\right\|^2 \mid \overline{\mathcal{E}}\right],$$

which corresponds to (55) with $\varepsilon = 1$. Thus, we just need to plug $\varepsilon = 1$ in (63) to obtain the desired result, i.e.,

$$\mathbb{E}\left[\mathbb{E}\left[\|H_B Z \boldsymbol{b}\|^2 \mid \boldsymbol{x}\right] \mid \overline{\mathcal{E}}\right] \leq \\ \leq \frac{kM}{1-\alpha}\left(\sum_{e=1}^{E} \frac{\lambda_e^2}{(M \lambda_e + \sigma_\nu^2)^2}\right)\left(\sum_{e=E+1}^{+\infty} \lambda_e\right).$$

Bounding $\mathbb{E}\left[\mathbb{E}\left[\|\boldsymbol{a} - H_B(G\boldsymbol{a} + \boldsymbol{\nu})\|^2 \mid \boldsymbol{x}\right] \mid \overline{\mathcal{E}}\right]$ *in* (68)

It is useful again to reason as if the measurements were generated according to (64) so that $\boldsymbol{y} = G\boldsymbol{a} + \boldsymbol{\nu}$. Then, let

$$\widehat{a}_B = H_B \boldsymbol{y}$$

and

$$\Phi_B = \operatorname{var}\left(\boldsymbol{a} - \widehat{a}_B | \boldsymbol{x}\right).$$

Recalling (65) and (66), let also

$$\widehat{a}_A = H_A \boldsymbol{y}$$

and

$$\Phi_A = \operatorname{var}\left(\boldsymbol{a} - \widehat{a}_A | \boldsymbol{x}\right) = \frac{\sigma_\nu^2}{M}\left(\frac{G^T G}{M} + \frac{\sigma_\nu^2}{M}\Lambda_E^{-1}\right)^{-1}.$$

After some simple calculations one obtains

$$\Phi_B = \Phi_A + \underbrace{(H_A - H_B)(G\Lambda G^T + \sigma_\nu^2 I)(H_A - H_B)^T}_{\widetilde{\Phi}}.$$

The bound for $\mathbb{E}\left[\operatorname{tr}\left(\Phi_A\right) \mid \overline{\mathcal{E}}\right]$ was already obtained in (67) so that now we can just focus on bounding $\mathbb{E}\left[\operatorname{tr}\left(\widetilde{\Phi}\right) \mid \overline{\mathcal{E}}\right]$. Define

$$A = \frac{G^T G}{M} + \frac{\sigma_\nu^2 \Lambda^{-1}}{M}, \quad B = I + \frac{\sigma_\nu^2 \Lambda^{-1}}{M}.$$

Then, it follows that

$$\begin{aligned} \widetilde{\Phi} &= (A^{-1} - B^{-1})\left(\tfrac{G^T G}{M}\Lambda\tfrac{G^T G}{M} + \tfrac{G^T G}{M^2}\sigma_\nu^2\right)(A^{-1} - B^{-1}) \\ &= A^{-1}(\underbrace{B - A}_{=:C})B^{-1}\left(\tfrac{G^T G}{M}\Lambda\tfrac{G^T G}{M} + \tfrac{G^T G}{M^2}\sigma_\nu^2\right) \\ &\times B^{-1}(B - A)A^{-1} \end{aligned}$$

so that

$$\begin{aligned} \operatorname{tr}\left(\widetilde{\Phi}\right) &= \operatorname{tr}\left(B^{-1}CA^{-2}CB^{-1}\left(\tfrac{G^T G}{M}\Lambda\tfrac{G^T G}{M} + \tfrac{G^T G}{M^2}\sigma_\nu^2\right)\right) \\ &= \operatorname{tr}\left(\Lambda^{1/2}\tfrac{G^T G}{M}B^{-1}CA^{-2}CB^{-1}\tfrac{G^T G}{M}\Lambda^{1/2}\right) \\ &\quad + \operatorname{tr}\left(\sigma_\nu^2 A^{-1}CB^{-1}\tfrac{G^T G}{M^2}B^{-1}CA^{-1}\right). \end{aligned}$$

Since

$$\varepsilon I \leq \frac{G^T G}{M} \leq (2-\varepsilon) I, \quad \Lambda \leq \lambda_1 I$$

we obtain

$$\begin{aligned} A &\geq \left(\varepsilon + \tfrac{\lambda_1^{-1}\sigma_\nu^2}{M}\right) I, \\ B &\geq \left(1 + \tfrac{\lambda_1^{-1}\sigma_\nu^2}{M}\right) I \geq \left(\varepsilon + \tfrac{\lambda_1^{-1}\sigma_\nu^2}{M}\right) I, \\ C^2 &= \left(I - \tfrac{G^T G}{M}\right)^2 \leq (1-\varepsilon)^2 I. \end{aligned}$$

Exploiting such inequalities and Lemma 12 we obtain

$$\mathbb{E}\left[\operatorname{tr}\left(\widetilde{\Phi}\right) \mid \overline{\mathcal{E}}\right] \leq$$

$$\left(\varepsilon + \frac{\lambda_1^{-1}\sigma_\nu^2}{M}\right)^{-4} (1-\varepsilon)^2 \frac{(2-\varepsilon)^2}{1-\alpha} \left(\sigma_\nu^2 \frac{E}{M} + \sum_{e=1}^{E} \lambda_e\right)$$

which, combined with (67), provides the desired result, also leading to the overall bound for $\mathrm{Err}_B(\boldsymbol{x})$.

C. Proof of Theorem 5

For both the cases the proof is obtained using Lemma 14. In particular, for both the estimators the lower bound is

$$q = \sum_{e=E+1}^{+\infty} \lambda_e$$

according to Theorem 2. Then, from the expressions of the bounds (29) and (32), it is immediate to verify that, for fixed E and α, they are monotonically decreasing in M. Now, for any $0 < \delta < 1$, let $\alpha = \delta$ and ε such that $\kappa < \frac{\delta}{2} \frac{4\delta}{1-\alpha} \sum_{e=1}^{\infty} \lambda_e$. Then, there exists M_0 such that for $M > M_0$ conditions (27) and (30) are satisfied and

$$\overline{\mathrm{Err}}_A \leq q + \delta \quad \text{with probability} \quad 1 - \delta$$

and

$$\overline{\mathrm{Err}}_B \leq q + \delta \quad \text{with probability} \quad 1 - \delta.$$

As anticipated, the use of Lemma 14 then concludes the proof.

D. Proof of Theorem 6

For what regards (34), it is sufficient to recall that, for finite M, as $E \to +\infty$ the estimator $\widehat{f}_A$ coincides with the MAP estimator which is consistent, see [60][Appendix] for details.

E. Proof of Theorem 7

The convergence (35) related to $\widehat{f}_B$ is more delicate and we will exploit bound (31). The rationale is to establish conditions on the convergence of E, M to infinity to make both the confidence level α and the bound (32) tend to zero. As for α, we can make it vanish to zero by setting $\alpha = \frac{1}{E^{1-\delta}}$ with $\delta \in (0,1)$. As for the bound (32), its first components naturally vanish with $E \to +\infty$, while the last two do not. In particular, one needs κ in (33) to go to zero and this requires $\varepsilon \to 1$. Hence, we set ε such that $\frac{(1-\varepsilon)^2}{2k} = \frac{1}{M^{1-\delta}}$. However, the convergence of α and (32) is not enough, since condition (30) must be always satisfied. This condition is indeed verified since

$$\frac{Ek}{M} \log \frac{E}{\alpha} = \frac{Ek}{M} \log E^{\delta} \leq \frac{kM^{\delta}}{M} = \frac{(1-\varepsilon)^2}{2} \leq 1 - \varepsilon + \varepsilon \log \varepsilon.$$

References

[1] T. Poggio and F. Girosi, "Networks for approximation and learning," *Proceedings of the IEEE*, vol. 78, no. 9, pp. 1481–1497, 1990.

[2] F. Cucker and S. Smale, "On the Mathematical Foundations of Learning," *Bulletin of the American Mathematical Society*, vol. 39, no. 101, pp. 1–49, 2001.

[3] M. Lei, L. Shiyan, J. Chuanwen, L. Hongling, and Z. Yan, "A review on the forecasting of wind speed and generated power," *Renewable and Sustainable Energy Reviews*, vol. 13, no. 4, pp. 915–920, 2009.

[4] L. Parolini, B. Sinopoli, B. H. Krogh, and Z. Wang, "A cyber–physical systems approach to data center modeling and control for energy efficiency," *Proceedings of the IEEE*, vol. 100, no. 1, pp. 254–268, 2012.

[5] A. Gelfand and S. a. G. S. Banerjee, "Spatial process modelling for univariate and multivariate dynamic spatial data," *Environmetrics*, vol. 16, p. 465?479, 2005.

[6] A. Datta, S. Banerjee, A. O. Finley, and A. E. Gelfand, "Hierarchical nearest-neighbor gaussian process models for large geostatistical datasets," *Journal of the American Statistical Association*, vol. 111, no. 514, pp. 800–812, 2016.

[7] Yunfei Xu, Jongeun Choi, S. Dass, and T. Maiti, "Sequential Bayesian Prediction and Adaptive Sampling Algorithms for Mobile Sensor Networks," *IEEE Transactions on Automatic Control*, vol. 57, no. 8, pp. 2078–2084, aug 2012.

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

[9] C. E. Rasmussen and C. K. I. Williams, *Gaussian processes for machine learning*. The MIT Press, apr 2006, vol. 14, no. 2.

[10] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society, Series B*, vol. 58, no. 1, pp. 267–288, 1996.

[11] N. Meinshausen and B. Yu, "Lasso-type recovery of sparse representations for high-dimensional data," *The Annals of Statistics*, vol. 37, no. 1, pp. 246–270, 2009.

[12] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, 2005.

[13] D. Wipf and S. Nagarajan, "A new view of automatic relevance determination," in *Proc. of NIPS*, 2007.

[14] A. Aravkin, J. Burke, A. Chiuso, and G. Pillonetto, "Convex vs. nonconvex approaches for sparse estimation: GLASSO, multiple kernel learning, and HGLASSO," *Journal of Machine Learning Research*, vol. 15, pp. 217–252, 2014.

[15] S. Boyd, P. Neal, E. Chu, B. Peleato, and E. Jonathan, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, 2011.

[16] J. Quiñonero-Candela and C. E. Rasmussen, "A Unifying View of Sparse Approximate Gaussian Process Regression," *Journal of Machine Learning Research*, vol. 6, no. Dec, pp. 1939–1959, 2005.

[17] E. Snelson and Z. Ghahramani, "Sparse Gaussian processes using pseudo-inputs," *Advances in Neural Information Processing Systems*, vol. 18, pp. 1257—-1264, 2006.

[18] M. Lázaro-Gredilla, J. Quiñonero-Candela, C. E. Rasmussen, and A. R. Figueiras-Vidal, "Sparse Spectrum Gaussian Process Regression," *The Journal of Machine Learning Research*, vol. 11, pp. 1865–1881, 2010.

[19] S. Ambikasaran, D. Foreman-Mackey, L. Greengard, D. W. Hogg, and M. O'Neil, "Fast Direct Methods for Gaussian Processes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 2, pp. 252–265, feb 2016.

[20] F. R. Bach and M. I. Jordan, "Predictive low-rank decomposition for kernel methods," in *Proceedings of the 22nd international conference on Machine learning - ICML '05*. New York, New York, USA: ACM Press, 2005, pp. 33–40.

[21] B. Kulis, M. Sustik, and I. Dhillon, "Learning low-rank kernel matrices," in *Proceedings of the 23rd international conference on Machine learning - ICML '06*. New York, New York, USA: ACM Press, 2006, pp. 505–512.

[22] C. Williams and M. Seeger, "Using the Nyström Method to Speed Up Kernel Machines," *Advances in Neural Information Processing Systems*, vol. 13, pp. 682—-688, 2001.

[23] K. Zhang and J. T. Kwok, "Clustered Nyström Method for Large Scale Manifold Learning and Dimension Reduction," *IEEE Transactions on Neural Networks*, vol. 21, no. 10, pp. 1576–1587, oct 2010.

[24] A. J. Smola, A. J. Smola, and B. Schölkopf, "Sparse Greedy Matrix Approximation for Machine Learning," in *Proceedings of the Seventeenth International Conference on Machine Learning*, 2000, pp. 911—-918.

[25] B. C. Levy, "Karhunen-loève expansion of gaussian processes," *Principles of Signal Detection and Parameter Estimation*, pp. 1–47, 2008.

[26] G. Ferrari-Trecate, C. K. I. Williams, and M. Opper, "Finite-dimensional approximation of Gaussian processes," *Proceedings of the 1998 conference on Advances in neural information processing systems*, no. n 3, pp. 218–224, 1999.

[27] H. Zhu, C. K. I. Williams, R. J. Rohwer, and M. Morciniec, "Gaussian regression and optimal finite dimensional linear models," in *Neural Networks and Machine Learning*, C. M. Bishop, Ed. Berlin: Springer-Verlag, 1998.

[28] L. Xiao and S. Boyd, "Fast linear iterations for distributed averaging," *Systems and Control Letters*, no. 53, pp. 65–78, 1998.

[29] F. Garin and L. Schenato, "A survey on distributed estimation and control applications using linear consensus algorithms," in *Networked Control Systems*. Springer, 2010, pp. 75–107.

[30] N. Bof, R. Carli, and L. Schenato, "Average consensus with asynchronous updates and unreliable communication," in *Proceedings of IFAC Word Congress*, ser. IFAC'17, vol. 1, 2017, pp. 601–606.

[31] J. B. Predd, S. R. Kulkarni, and H. V. Poor, "A Collaborative Training Algorithm for Distributed Learning," *IEEE Transactions on Information Theory*, vol. 55, no. 4, pp. 1856–1871, apr 2009.

[32] F. Perez-Cruz and S. R. Kulkarni, "Robust and Low Complexity Distributed Kernel Least Squares Learning in Sensor Networks," *IEEE Signal Processing Letters*, vol. 17, no. 4, pp. 355–358, apr 2010.

[33] P. Honeine, M. Essoloh, C. Richard, and H. Snoussi, "Distributed Regression in Sensor Networks with a Reduced-Order Kernel Model," in *IEEE GLOBECOM 2008 - 2008 IEEE Global Telecommunications Conference*. IEEE, 2008, pp. 1–5.

[34] S. Martinez, "Distributed Interpolation Schemes for Field Estimation by Mobile Sensor Networks," *IEEE Transactions on Control Systems Technology*, vol. 18, no. 2, pp. 491–500, mar 2010.

[35] Y. Xu, J. Choi, S. Dass, and T. Maiti, "Efficient Bayesian spatial prediction with mobile sensor networks using Gaussian Markov random fields," *Automatica*, vol. 49, no. 12, pp. 3520–3530, 2013.

[36] J. Cortés, "Distributed Kriged Kalman filter for spatial estimation," *IEEE Transactions on Automatic Control*, vol. 54, no. 12, pp. 2816–2827, 2009.

[37] J. Choi, S. Oh, and R. Horowitz, "Distributed learning and cooperative control for multi-agent systems," *Automatica*, vol. 45, no. 12, pp. 2802–2814, dec 2009.

[38] D. Varagnolo, G. Pillonetto, and L. Schenato, "Distributed parametric and nonparametric regression with on-line performance bounds computation," *Automatica*, vol. 48, no. 10, pp. 2468–2481, 2012.

[39] T. Zhang, "Learning Bounds for Kernel Regression Using Effective Data Dimensionality," *Neural Computation*, vol. 17, no. 9, pp. 2077–2098, sep 2005.

[40] C. M. Stein, "Estimation of the Mean of a Multivariate Normal Distribution," *The Annals of Statistics*, vol. 9, no. 6, pp. 1135–1151, 1981.

[41] J. Rice, "Choice of smoothing parameter in deconvolution problems," *Contemporary Math.*, vol. 59, pp. 137–151, 1986.

[42] G. Golub, M. Heath, and G. Wahba, "Generalized cross-validation as a method for choosing a good ridge parameter," *Technometrics*, vol. 21, no. 2, pp. 215–223, 1979.

[43] J. S. Maritz and T. Lwin, *Empirical Bayes Method*. Chapman and Hall, 1989.

[44] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York: Springer, 2009.

[45] Y. Xu and J. Choi, "Adaptive Sampling for Learning Gaussian Processes Using Mobile Sensor Networks," *Sensors*, vol. 11, no. 3, pp. 3051–3066, 2011.

[46] J. Snoek, H. Larochelle, and R. P. Adams, "Practical bayesian optimization of machine learning algorithms," in *Proceedings of the 25th International Conference on Neural Information Processing Systems*, ser. NIPS'12, 2012, pp. 2951–2959.

[47] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, *{Markov} chain {Monte Carlo} in Practice*. London: Chapman and Hall, 1996.

[48] P. Magni, R. Bellazzi, and G. D. Nicolao, "Bayesian function learning using mcmc methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 12, pp. 1319–1331, 1998.

[49] B. M. Bell and G. Pillonetto, "Estimating parameters and stochastic functions of one variable using nonlinear measurement models," *Inverse Problems*, vol. 20, no. 3, pp. 627–646, jun 2004.

[50] G. De Nicolao and G. F. Trecate, "Consistent identification of narx models via regularization networks," *IEEE Transactions on Automatic Control*, vol. 44, no. 11, pp. 2045–2049, 1999.

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

[52] D. J. C. MacKay, "Bayesian interpolation," *Neural Computation*, vol. 4, no. 3, pp. 415–447, 1992.

[53] G. De Nicolao, G. Sparacino, and C. Cobelli, "Nonparametric input estimation in physiological systems: problems, methods and case studies," *Automatica*, vol. 33, pp. 851–870, 1997.

[54] G. Kimeldorf and G. Wahba, "A correspondence between bayesian estimation on stochastic processes and smoothing by splines," *The Annals of Mathematical Statistics*, vol. 41, no. 2, pp. 495–502, 1970.

[55] B. Schölkopf, R. Herbrich, and A. J. Smola, "A generalized representer theorem," *Neural Networks and Computational Learning Theory*, vol. 81, pp. 416–426, 2001.

[56] P. Drineas and M. Mahoney, "On the Nyström Method for Approximating a Gram Matrix for Improved Kernel-Based Learning," *J. Mach. Learn. Res.*, vol. 6, pp. 2153–2175, Dec. 2005.

[57] T. Yang, Y. Li, M. Mahdavi, R. Jin, and Z. Zhou, "Nyström method vs random Fourier features: A theoretical and empirical comparison," in *Advances in neural information processing systems*, ser. NIPS'12, 2012, pp. 476–484.

[58] "Colorado monthly meteorology dataset 1995-1997," https://www.image.ucar.edu/Data/US.monthly.met/CO.shtml, accessed: 2017-06-20.

[59] J. A. Tropp, "User-Friendly Tail Bounds for Sums of Random Matrices," *Foundations of Computational Mathematics*, vol. 12, no. 4, pp. 389–434, aug 2011.

[60] G. Pillonetto and B. M. Bell, "Bayes and empirical Bayes semi-blind deconvolution using eigenfunctions of a prior covariance," *Automatica*, vol. 43, no. 10, pp. 1698–1712, oct 2007.