Distributed multi-agent Gaussian regression via Karhunen-Loève expansions

Gianluigi Pillonetto, Luca Schenato, Damiano Varagnolo

Abstract

We consider the problem of distributedly estimating Gaussian random fields in multi-agent frameworks. Each sensor collects few measurements and aims to collaboratively reconstruct a common estimate based on all data. Agents are assumed to have 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 benefit is twofold: first, the computation and communication complexities scale with $E$ and not with $M$. Second, computing the required sufficient statistics can be performed via standard average consensus algorithms. We obtain probabilistic non-asymptotic bounds for both approaches, useful to determine a priori the desired level of estimation accuracy. Furthermore, we also derive new distributed strategies to tune the regularization parameter which rely on the Stein’s unbiased risk estimate (SURE) paradigm and can again be implemented via standard average consensus algorithms. 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 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], or the reconstruction of the temperature field in a datacenter from local measurements at each server [4]. 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 sought [5].

This paper considers a distributed nonparametric Gaussian regression approach. In this context, the unknown map is modeled as a zero-mean Gaussian random field whose covariance (also called kernel in the machine learning literature) has to embed expected properties like smoothness [6], [7]. In our distributed framework 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.

Achieving the minimum variance estimate requires knowing all the $M$ measurements and related input locations, plus invert an $M \times M$ matrix which requires $O(M^3)$ 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 [8], [9], [10], the use of matrix factorizations [11] and approximations of the kernel function [12], [13] through the Nyström method or greedy techniques [14], [15], [16]. Along this way, KL expansions [17] 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 random field via the $E$ kernel eigenfunctions associated...
to the largest eigenvalues, an approximation that corresponds to perform the best process approximation before seeing the data [17] (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 [18]. However, the a priori basis given by the KL expansion has important advantages. In fact, as proved in [19], 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 [18], the a priori basis can be computed off-line. Also, as detailed in Section III-B the sufficient statistics necessary to compute the corresponding estimates involve the computation of the average of \( M \) local matrices and vectors of dimension \( E \times E \) and \( E \), respectively, with possibly \( E \ll M \). Such averages can be readily computed by each agent via the popular average consensus algorithms available in the literature [20] as long as the communication graph is connected. This explains why KL-based estimators will be the focus of this paper.

Our stream of research pairs the ones of other authors focusing on distributed kernel regression. An example is [21], that proposes a distributed regularized kernel Least Squares (LS) regression algorithm that exploits successive orthogonal projections, or [22] that extends [21] by designing strategies to reduce the communication and synchronization needs. Estimators with reduced order model complexity have been proposed in [23], while nonparametric schemes using Nearest-Neighbors interpolation strategies have been studied also in [24]. Another Gaussian estimation approach is considered in [25], 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 [26] with the aim of estimating a dynamic Gaussian random field and its gradient, while in [27] 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 \( \hat{f}_A \) and \( \hat{f}_B \) which have computational and communication complexities of order \( O(E^2) \) and \( O(E) \), respectively. These two strategies were also proposed in [28] designing a Monte Carlo approach to measure their performance. Here, instead, assigned the kernel and the input locations statistics, their a priori prediction capability on future data is characterized by deriving non-asymptotic error bounds function of \( E, M \) and \( \alpha \). 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 [29]. 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 \( \hat{f}_A \) and \( \hat{f}_B \) 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 \( \hat{f}_A \) is always consistent, i.e., convergent in probability to the true function as \( E, M \to \infty \), consistency of \( \hat{f}_B \) 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 the terminal part of the paper, we address this problem by proposing a novel distributed tuning strategy based on the SURE criterion [30]. Standard approaches proposed in the literature in the context of a centralized framework, like cross-validation and maximum likelihood [31], MAP estimation [32], expected improvement [33] and Bayesian Monte Carlo [34], 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 \( \hat{f}_A \) is employed, and via only two averages of size \( O(E) \) when \( \hat{f}_B \) is employed.

The paper is organized as follows. Section II formulates the Bayesian estimation problem while Section III describes the KL expansion of the Gaussian random field 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. 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

\[ y_m = f(x_m) + \nu_m, \quad m = 1, \ldots, M \]  

(1)

with the input locations \( x_m \) following the stochastic generation scheme

\[ x_m \sim \mu(X) \text{ i.i.d.,} \quad m = 1, \ldots, M, \]  

(2)

with \( \mu \) a non-degenerate probability measure on the compact \( X \). The unknown function \( f : X \to \mathbb{R} \) is a zero-mean Gaussian random field with continuous covariance \( K : X \times X \to \mathbb{R} \), i.e.,

\[ f \sim \mathcal{N}(0, K). \]  

(3)

The measurement noise is also Gaussian of known variance \( \sigma^2 \nu \):

\[ \nu_m \sim \mathcal{N}(0, \sigma^2 \nu). \]  

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

\[ \hat{f}_{\text{MAP}}(x) = \left[ K(x, x_1) \ldots K(x, x_M) \right] H_{\text{MAP}} \begin{bmatrix} y_1 \\ \vdots \\ y_M \end{bmatrix} \]

with

\[ H_{\text{MAP}} := \left( \begin{array}{ccc} K(x_1, x_1) & \cdots & K(x_1, x_M) \\ \vdots & \ddots & \vdots \\ K(x_M, x_1) & \cdots & K(x_M, x_M) \end{array} \right) + \sigma^2 \nu I \right)^{-1}. \]

The storage and computational requirements needed to compute \( \hat{f}_{\text{MAP}} \) are thus \( O(M^2) \) and \( O(M^3) \), respectively. The communication complexity is either \( O(\dim(X)M) \) if agents share the input locations \( x_m \) or \( O(M^2) \) 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 \( \hat{f}_{\text{MAP}} \) 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 \( \phi_e \) orthonormal w.r.t. the measure \( \mu \) in (2) and related eigenfunctions \( \lambda_e \) [17]. They are defined by

\[ \lambda_e \phi_e(x) = \int_X K(x, x') \phi_e(x') d\mu(x'), \]  

(4)

\[ K(x, x') = \sum_{e=1}^{+\infty} \lambda_e \phi_e(x) \phi_e(x') \quad \lambda_1 \geq \lambda_2 \ldots > 0, \]  

(5)

and, using \( \delta_{ij} \) for the Kronecker delta,

\[ \int_X \phi_i(x) \phi_j(x) d\mu(x) = \delta_{ij}. \]  

(6)

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

\[ f(x) = \sum_{e=1}^{E} a_e \phi_e(x) + \sum_{e=1}^{+\infty} b_e \phi_E(x) \]

\[ =: f_a(x) + : f_b(x) : \]  

(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
\[
a_e \sim \mathcal{N}(0, \lambda_e), \quad e = 1, \ldots, E \tag{8a}
\]
\[
b_e \sim \mathcal{N}(0, \lambda_{E+e}), \quad e = 1, 2, \ldots \tag{8b}
\]
It is well known that
\[
S := \text{span} \{ \phi_1(\cdot), \ldots, \phi_E(\cdot) \} \tag{9}
\]
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 \([19]\).

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.
\[
\sup_{x \in \mathcal{X}} |\phi_e(x)| \leq \sqrt{k} < +\infty \quad e = 1, 2, \ldots \tag{10}
\]

Assumption [10] is satisfied by all the finite-dimensional kernels and also by classical covariances like the spline kernels, e.g., see \([35]\) for the case of uniform \( \mu \). 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 \([36]\), also permitting to achieve 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 \( \hat{f}_A \) and \( \hat{f}_B \), which assume values in the finite-dimensional subspace \( S \) defined in (9). First, it is useful to rewrite model (1) in a more compact form.

Let
\[
x := [x_1, \ldots, x_M]^T
\]
\[
y := [y_1, \ldots, y_M]^T \quad \nu := [\nu_1, \ldots, \nu_M]^T
\]
\[
a := [a_1, \ldots, a_E]^T \quad b := [b_1, b_2, \ldots]^T
\]
\[
G := \begin{bmatrix} G_{11} & \cdots & G_{1E} \\ \vdots & \ddots & \vdots \\ G_{M1} & \cdots & G_{ME} \end{bmatrix} \quad Z := \begin{bmatrix} Z_{11} & Z_{12} & \cdots \\ \vdots & \vdots & \vdots \\ Z_{M1} & Z_{M2} & \cdots \end{bmatrix}
\]
\[
G_{me} := \phi_e(x_m), \quad m = 1, \ldots, M, \quad e = 1, \ldots, E,
\]
\[
Z_{me} := \phi_{E+e}(x_m), \quad m = 1, \ldots, M, \quad e = 1, 2, \ldots
\]

Considering decomposition (7), definitions (11)-(13) and using classical algebraic notation to handle infinite-dimensional objects, the measurements model (1) becomes
\[
y = Ga + Zb + \nu. \tag{14}
\]

With this novel notation \( Ga \) accounts for the contribution from \( f_a \) while \( Zb \) accounts for the contribution from \( f_b \).

**C. The \( E \)-dimensional estimator \( \hat{f}_A \)**

Let
\[
\hat{f}_A(x) := [\phi_1(x) \cdots \phi_E(x)] H_A y
\]
where
\[
H_A := \left( \frac{G^T G}{M} + \frac{\sigma^2}{M} \Lambda_E^{-1} \right)^{-1} \frac{G^T}{M}
\]
and \( \Lambda_E := \text{diag}(\lambda_1, \ldots, \lambda_E) \). The estimator \( \hat{f}_A \) is suitable for distributed computations. In fact, defining
\[
G_m := [\phi_1(x_m), \ldots, \phi_E(x_m)]
\]
one has

\[ \frac{G^T G}{M} = \frac{1}{M} \sum_{m=1}^{M} G_m^T G_m, \quad \frac{G^T y}{M} = \frac{1}{M} \sum_{m=1}^{M} G_m^T y_m. \]  

(17)

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, (17) points out that $\hat{f}_A$ 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 $\hat{f}_B$

As stated in (6), one has

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

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

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

Thus, it is tempting to use the approximation

\[ \frac{G^T G}{M} \approx I \]  

(18)

and use, in place of $H_A$ in (15), the matrix

\[ H_B := \left( I + \frac{\sigma^2}{M} \Lambda_E^{-1} \right)^{-1} \frac{G^T}{M}. \]  

(19)

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

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

The estimator $\hat{f}_B$ is more advantageous than $\hat{f}_A$ 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 $\hat{f}_A$), and does not require any expensive matrix inversion since $I + \frac{\sigma^2}{M} \Lambda_E^{-1}$ is diagonal.

IV. STATISTICAL ANALYSIS OF $\hat{f}_A$ AND $\hat{f}_B$

Ideally one would like to compute $\mathbb{E} \left[ \| f - \hat{f}_A \|^2 \right]$ and $\mathbb{E} \left[ \| f - \hat{f}_B \|^2 \right]$, 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., $M, E, \sigma^2$. To circumvent this challenge, we will exploit the assumption that the input locations are randomly drawn from a known distribution $\mu$ and the orthonormality of the eigenfunctions to find bounds on $\mathbb{E} \left[ \| f - \hat{f}_A \|^2 \right]$ 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 $\mathbb{E} \left[ \| f - \hat{f}_A \|^2 \mid \mathcal{E} \right]$ can be computed. This is formally described in the next sections.

A. Performance indexes and lower bound

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

\[ \text{Err}_A(x) := \mathbb{E} \left[ \| f - \hat{f}_A \|^2 \mid x \right] \]

(20)

\[ \text{Err}_B(x) := \mathbb{E} \left[ \| f - \hat{f}_B \|^2 \mid x \right] \]

where

\[ \| g \|^2 := \int_X g^2(x) \, d\mu(x). \]
The variables $\text{Err}_A(x)$ and $\text{Err}_B(x)$ are stochastic, since they are functions of the random input locations $x$ 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 $x$. Note also that $\| \cdot \|$ depends on $\mu$ so that $\text{Err}_A$ and $\text{Err}_B$ 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 $\text{Err}_A(x)$ and $\text{Err}_B(x)$ 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 $\hat{f}_*\star$ be any generic estimator of $f$, function of $x$ and $y$ and assuming values in any generic $E$-dimensional space fixed a priori. Then

$$\min_{\hat{f}_*} \mathbb{E} \left[ \|f - \hat{f}_*\|^2 \mid x \right] \geq \sum_{e=E+1}^{+\infty} \lambda_e. \quad (21)$$

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

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

$$\mathbb{E} \left[ \text{Err}_A(x) \mid \mathcal{E} \right] \leq q \quad (22)$$

or

$$\mathbb{E} \left[ \text{Err}_B(x) \mid \mathcal{E} \right] \leq q. \quad (23)$$

Thus, if $\alpha$ is close to zero saying that $\overline{\text{Err}}_A \leq q$ with probability $1 - \alpha$ is equivalent to saying that the average error associated to $\hat{f}_A$ is smaller than $q$ with high probability. Finally, note that setting $\mathcal{E}$ to the entire sample space, the expectations in the lhs of (22) and (23) become the Mean Square Errors (MSEs) of $\hat{f}_A$ and $\hat{f}_B$, i.e.

$$\text{MSE}_{\hat{f}_A} = \int_X \text{Err}_A(x) d\mu(x), \quad (24)$$

$$\text{MSE}_{\hat{f}_B} = \int_X \text{Err}_B(x) d\mu(x). \quad (25)$$
B. Non asymptotic error bounds

The key issue is to bound the performance indexes $\text{Err}_A$ and $\text{Err}_B$ for any finite number of measurements $M$ and eigenfunctions $E$. The following theorem provides the desired bounds. It depends on the input locations distribution $\mu$, the kernel eigenvalues $\lambda_e$ 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 $\frac{\sigma^2 F}{M}$, as detailed in the proof contained in the Appendix.

**Theorem 4** Let the assumptions in Section II-A and Assumption 7 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

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

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

$$\text{Err}_A \leq \text{Bnd}_A$$

with

$$\text{Bnd}_A := \frac{k M}{1 - \alpha} \left( \sum_{e=1}^{E} \frac{\lambda_e^2}{(\varepsilon M \lambda_e + \sigma^2_\nu)^2} \right) \left( \sum_{e=E+1}^{\infty} \lambda_e \right)$$

$$+ \frac{\sigma^2_\nu}{1 - \alpha} \left( \sum_{e=1}^{E} \frac{\lambda_e}{\varepsilon M \lambda_e + \sigma^2_\nu} \right) + \left( \sum_{e=E+1}^{\infty} \lambda_e \right),$$

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

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

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

$$\text{Err}_B \leq \text{Bnd}_B$$

with

$$\text{Bnd}_B := \frac{k M}{1 - \alpha} \left( \sum_{e=1}^{E} \frac{\lambda_e^2}{(M \lambda_e + \sigma^2_\nu)^2} \right) \left( \sum_{e=E+1}^{\infty} \lambda_e \right)$$

$$+ \frac{\sigma^2_\nu}{1 - \alpha} \left( \sum_{e=1}^{E} \frac{\lambda_e}{\varepsilon M \lambda_e + \sigma^2_\nu} \right) + \left( \sum_{e=E+1}^{\infty} \lambda_e \right)$$

$$+ \kappa \left( \frac{E}{M} \sigma^2_\nu + \sum_{e=1}^{E} \lambda_e \right),$$

where

$$\kappa = \frac{1}{1 - \alpha} \left( \varepsilon + \frac{\lambda^{-1}_{\epsilon} \sigma^2_\nu}{M} \right)^{-4} (1 - \varepsilon)^2 (2 - \varepsilon)^2$$

The obtained bounds are now tested via a numerical example.

C. Numerical study

Consider the first-order spline kernel [37] 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 $\mu$ in (2) set to the uniform distribution on $[0,1]$. With these settings

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

and $k = 2$. To make the bounds only depend on $E$ we set $M = 10000$, $1 - \alpha = 0.95$, the noise variance $\sigma^2_\nu = 0.1^2$, and $\varepsilon \in (0,1]$ that minimizes the bound while satisfying (26) or (29) accordingly.

The thick lines in the two top panels of Figure 1 show how $\text{Bnd}_A$ (left) and $\text{Bnd}_B$ (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 (24) and (25), 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 $\text{Bnd}_A$, 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 $\hat{f}_A$ 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 $\text{Bnd}_B$ is instead different and exhibit a clear minimum at $E = 7$. The reason is that $\hat{f}_B$ relies on the asymptotic matrix approximation [18]. The bound $\text{Bnd}_B$ 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 $\hat{f}_B$ is not guaranteed to converge to the minimum variance estimator). Note that, in this case, $\text{Bnd}_B$ 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 $\hat{f}_B$ 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 $\hat{f}_A$ and $\hat{f}_B$ asymptotically reach the lower bound (21).

**Theorem 5** Given the assumptions in Section II-A and Assumption 7

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

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

We now discuss the statistical consistency of our estimators. In this case, the conditions under which $\hat{f}_A$ and $\hat{f}_B$ 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 7

$$
\lim_{M \to +\infty} \lim_{E \to +\infty} \text{Err}_A = 0 \quad \text{in probability}
$$

**Theorem 7** Let $E = E(M)$ such that $E(M) \log E^8(M) \leq M^8$, $\lim_{M \to +\infty} E(M) = +\infty$ for some $\delta \in (0, 1)$. Given the assumptions in Section II-A and Assumption 7 then

$$
\lim_{M \to +\infty} \lim_{E \to +\infty} \text{Err}_B = 0 \quad \text{in probability}
$$

**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

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

must be satisfied. Asymptotically, the lhs tends to $0^+$ while the second term becomes $k - \frac{E}{M} \log(\alpha/2)$ and is larger than $k$ when $\alpha$ 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 $E(M) = M^\alpha$, for any $\alpha \in (0, 1)$, satisfies the consistency condition, which can be readily verified by choosing $\delta = \frac{1+\alpha}{2}$.

The consistency properties of $\hat{f}_A$ and $\hat{f}_B$ are thus remarkably different. For what regards $\hat{f}_A$, 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 $\hat{f}_{\text{MAP}}$. This agrees with what already discussed in the previous subsection: when using $\hat{f}_A$ it is convenient for the network to use a dimension $E$ as large as possible, just compatible with its communication constraints. Differently, the estimator $\hat{f}_B$ 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 $\hat{f}_A$ and $\hat{f}_B$ 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 $\gamma$. In addition, the prior is never perfect and the tuning of $\gamma$ could also hinder possible undermodeling. So, in place of \( [3] \), in practical applications it is beneficial to consider

\[
\gamma \sim \mathcal{N}(0, \gamma^{-1} K)
\]

with $\gamma$ to be estimated from the observed noisy outputs and related input locations. Furthermore, when $\hat{f}_B$ 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 [31][Section 7.4].

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

\[
z = \eta + e
\]

and consider also future measurements

\[
z^* = \eta + e^*
\]

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

\[
\|z - \hat{z}\|^2 + 2 \text{tr} (S\Sigma).
\]

(35)

The quantity $\text{tr} (S\Sigma)$ entering the second part of the objective \( (35) \) is connected to the concept of equivalent degrees of freedom [38], [39].

In what follows, we assume that $\gamma$ is unknown but belongs to the finite set $\Gamma$ 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 $\hat{f}_A$ has been used, each agent has stored $G^T G_M$ and $G^T y_M$ so that, letting

\[
H_A(\gamma) := \left( \frac{G^T G}{M} + \frac{\gamma \sigma^2}{M} \Lambda_E \right)^{-1} G^T M,
\]

it can compute $H_{Ay}$ for any $\gamma \in \Gamma$.

If $\hat{f}_B$ 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 $G^T y_M$ and, letting

\[
H_B(\gamma, E') := \mathcal{I}_{E'} \left( I + \frac{\gamma \sigma^2}{M} \Lambda_E \right)^{-1} G^T M,
\]

where

\[
\mathcal{I}_{E'} := \begin{bmatrix} I_{E'} & 0_{E-E'} \end{bmatrix},
\]

it can compute $H_{By}$ for any $\gamma \in \Gamma$ and integer $E' \in \Omega$.

A. Distributed SURE for $\hat{f}_A$: tuning of $\gamma$

The first strategy is suited for $\hat{f}_A$. Surprisingly, we will see that the tuning of $\gamma$ 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

\[
y = Ga + Zb + \nu
\]
where $a$ is $E$-dimensional. Hereby, we break away from the assumptions on prior correctness by thinking of $Ga + Zb$ 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 pre-multiplication by $G^T/M$, i.e.,

$$
\frac{G^Ty}{M} = \frac{G^TG}{M} a + \frac{G^TZ}{M} b + \frac{GT\nu}{M}
$$

(36)

where the correspondences with the key quantities defining the risk estimator \([55]\) have been pointed out. From such definitions, we also obtain $\hat{z} = \frac{G^TG}{M} H A \frac{G^Ty}{M} = S \hat{z}$ where

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

and

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

Recall that the matrix $V = \frac{G^TG}{M} = \frac{1}{M} \sum_{m=1}^{M} G^T_m G_m$ and the vector $z = \frac{1}{M} \sum_{m=1}^{M} G^T_m y_m$ have been already computed by each agent via a distributed consensus algorithm \([20]\) to implement $\hat{f}_A$. Then, since the network cardinality $M$ is known, each agent can tune $\gamma$ by optimizing the SURE score \([55]\) connected with the prediction risk on the future data $z^* = \frac{G^Ty^*}{M}$, i.e.,

$$
\hat{\gamma}_A = \arg \min_{\gamma \in \Gamma} J_A(\gamma)
$$

(37)

with

$$
J_A(\gamma) := \| (I - S) z \|^2 + 2 \text{tr} (S \Sigma)
$$

$$
= \left\| \frac{\gamma}{M} \sigma^2 \left( V \Lambda_E + \frac{\gamma}{M} \sigma^2 I \right)^{-1} z \right\|^2 + \frac{2 \sigma^2}{M} \text{tr} \left( V^2 \left( V + \frac{\gamma}{M} \sigma^2 \Lambda_E^{-1} \right)^{-1} \right).
$$

To understand the rationale underlying this strategy we have just to consider that the novel process \([36]\) 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 $G^T Z b$ vanishes so that $\eta \approx a$. 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 $\hat{f}_B$. In fact, it requires each agent to know $\frac{G^TG}{M}$. But if this quantity were known, each agent could implement $\hat{f}_A$, an estimator that has more favorable features than $\hat{f}_B$.

**B. Distributed SURE for \(\hat{f}_B\): tuning of $E$ and $\gamma$**

The second strategy is designed for $\hat{f}_B$. 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 \([36]\), i.e., the $E$-dimensional projected measurement space, where $z = \frac{G^Ty}{M} = \frac{1}{M} \sum_{m=1}^{M} G^T_m y_m \in \mathbb{R}^E$ has been computed to implement $\hat{f}_B$ via a standard distributed consensus algorithm and is therefore known to each agent. Let us define $\tilde{a}(\gamma, E') = H_B(\gamma, E') z \in \mathbb{R}^E$. Clearly $\tilde{a}(\gamma, E')$ for $E' < E$ is simply the truncated version of $\tilde{a}(\gamma, E)$ where the last $E - E'$ components are set to zero. Moreover, the vectors $\tilde{a}(\gamma, E')$ can be independently computed by each agent for each value of $\gamma \in \Gamma$, $E' \in \Omega$ once $z$ is available.

The output prediction can be written as

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

Hence, each agent can compute the vectors $\tilde{z}(\gamma, E')$ for each $\gamma \in \Gamma$ and $E' \in \Omega$ by running an additional consensus of size $O(\dim(\Omega) \dim(\Gamma))$. As so, the first part of the SURE score $\| z - \tilde{z}(\gamma, E') \|^2$ 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 $\text{tr} (S(\gamma, E') \Sigma)$ where

$$
S(\gamma, E') = \frac{G^T G}{M} \mathbb{I}_{E'} \left( I + \frac{\gamma}{M} \sigma^2 \Lambda_E^{-1} \right)^{-1}, \quad \Sigma = \frac{\sigma^2}{M} G^T G.
$$

Obviously, this would not make too much sense in the context of $\hat{f}_B$ since the computation of $V = \frac{G^TG}{M}$ would allow us to compute $\hat{f}_A$ which has better performance anyways. Therefore we will approximate such matrix $V$ (similarly to what we did
to obtain \( \hat{f}_B \) by replacing it with an identity matrix. This corresponds to use a sort of expected equivalent degrees of freedom:
\[
\text{tr} \left( S(\gamma, E') \Sigma \right) \approx \frac{\sigma^2}{M} \text{tr} \left( I + \frac{\gamma \sigma^2}{M} \Lambda_E^{-1} \right) \]
\[
= \frac{\sigma^2}{M} \sum_{e=1}^{E'} \lambda_e + \gamma \sigma^2/\tilde{m}.
\]
The optimal tuning of the parameter is then obtained as
\[
\left( \hat{\gamma}_B, \hat{E}_B \right) = \arg\min_{\gamma \in \Gamma, E' \in \Omega} J_B(\gamma, E')
\]
with
\[
J_B(\gamma, E') := \| z - \hat{z}(\gamma, E') \|^2 + 2 \frac{\sigma^2}{M} \sum_{e=1}^{E'} \lambda_e + \gamma \sigma^2/\tilde{m}.
\]

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

C. 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 \( \hat{f}_A \) and \( \hat{f}_B \). The errors are still the MSES defined in (24) and (25) 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 \( \hat{f}_A \) is adopted we fix \( E = 400 \) and the regularization parameter is searched over a grid \( \Gamma \) containing 50 logarithmically spaced values between \( 10^{-3} \) and \( 10^3 \). When using \( \hat{f}_B \), the grid \( \Gamma \) 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:

• “\( \hat{f}_A + \text{oracle} \)” and “\( \hat{f}_B + \text{oracle} \)” 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 \( \hat{a}(\gamma, E') \) denote the vector with the estimates of the first \( E' \) coefficients \( a_e \) returned by \( \hat{f}_B \). Then \( \hat{f}_B + \text{oracle} \) determines the hyperparameters as

\[
\left( \hat{\gamma}, \hat{E} \right) := \arg\min_{\gamma \in \Gamma, E' \in \Omega} \sum_{e=1}^{\infty} (a_e - \hat{a}_\gamma(\gamma, E'))^2
\]

where \( \hat{a}_\gamma(\gamma, E') := 0 \) for \( e > E' \). Thus, this estimator is not implementable in practice and provides the lower bound on the MSES (24) and (25) achievable by the two estimators;

• “\( \hat{f}_A + \text{SURE} \)” and “\( \hat{f}_B + \text{SURE} \)” where the hyperparameters tuning step is performed following the SURE approaches described in the previous subsection. Recall that “\( \hat{f}_A + \text{SURE} \)” requires only a single consensus on a vector of size \( O(E^2) \) to obtain simultaneously both the hyperparameters and function estimates, while “\( \hat{f}_B + \text{SURE} \)” requires two consensus operations of size \( O(E) \).

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 “\( \hat{f}_B + \text{SURE} \)” (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 \( S_p \in [0, 1] \) 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 \( S_p = 1 \) indicates that SURE is performing as well as the oracle and that, for \( M = 10000 \), \( S_p \) becomes the distillate of Figure 2. Table I reports \( S_p \) for \( M = 100, 1000, 10000 \): one can see that the proposed hyperparameter estimation procedure behaves very nicely.

D. Numerical study on field data

Let us now consider the reconstruction of monthly precipitations using data collected in Colorado in the years 1995-1997 [40]. Measurements \( y_m \) are a series of monthly average precipitations measured at 367 stations within the rectangular longitude/latitude region \([-109.5, -101] \times [36.5, 41.5] \) remapped for convenience into the unitary square so that \( x_m \in [0, 1]^2 \) for every \( m \).
Figure 2. Comparison of the errors obtained by the SURE- and oracle-based strategies. Each opaque circle corresponds to the result of a certain Monte Carlo run. 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.

We test the SURE strategies (37) and (38). When using $\hat{f}_A$ we set $E = 20$ and $\Gamma$ to the grid containing 50 values logarithmically spaced between $10^{-3}$ and $10^5$. When adopting $\hat{f}_B$ 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

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

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 $D_{\text{train}}$ 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 $D_{\text{test}}$ corresponding to the remaining $1/3$ measurements from the first selected month. The last one is $D_{\sigma_2^2}$ 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 $\hat{f}_A$ with $\gamma = 0$ obtaining as estimate of the noise variance

$$\hat{\sigma}_e^2 = \frac{1}{\text{dim}(D_{\sigma_2^2}) - E} \sum_{m=1}^{\text{dim}(D_{\sigma_2^2})} \left(\hat{f}_A(x_m; 0) - y_m\right)^2$$

where $\text{dim}(D_{\sigma_2^2})$ is the cardinality of $D_{\sigma_2^2}$. Overall, this represents a situation where noise levels are determined by a centralized approach before running the estimators $\hat{f}_A$ and $\hat{f}_B$.

The following tuning strategies are used:

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

$$\text{RSS}_A(\gamma) := \frac{1}{\text{dim}(D_{\text{test}})} \sum_{m=1}^{\text{dim}(D_{\text{test}})} \left(\hat{f}_A(x_m; \gamma) - y_m\right)^2$$

$$\text{RSS}_B(\gamma, E') := \frac{1}{\text{dim}(D_{\text{test}})} \sum_{m=1}^{\text{dim}(D_{\text{test}})} \left(\hat{f}_B(x_m; \gamma, E') - y_m\right)^2$$

where $(x_m, y_m)$ are all elements of $D_{\text{test}}$. Note that RSS$_A$ and RSS$_B$ can be seen as approximations of the MSEs (24) and (25) and that the oracle provides a lower bound on their values;

2) “$\hat{f}_A+$ SURE” and “$\hat{f}_B+$ SURE”, where the hyperparameters tuning step is performed minimizing the estimated risks $J_A(\gamma)$ and $J_B(\gamma, E')$ defined in (37) and (38).

Table I

| Data set size | $S_p$ |
|--------------|-------|
| $M = 100$    | 0.93  |
| $M = 1000$   | 0.987 |
| $M = 10000$  | 0.99  |

SURE’S PERFORMANCE INDEX $S_p$ SUMMARIZING FOUR MONTE CARLO STUDIES AS A FUNCTION OF THE NUMBER $M$ OF AVAILABLE MEASUREMENTS. A VALUE OF $S_p$ CLOSE TO 1 INDICATES THAT SURE’S PERFORMANCE IS CLOSE TO THAT OF THE ORACLE. FOR $M = 10000$, $S_p$ REPRESENTS A RESUME OF THE ENTIRE FIGURE.

Figure 3 compares with a scatter-plot the prediction errors (39) and (40) achieved by the estimators after the 1000 Monte Carlo runs. The situation is not dissimilar from the case of synthetic data: the performance of the SURE strategies is close to that of the oracles.
Figure 3. Comparison of the prediction errors obtained by the SURE- and oracle-based strategies. Each opaque circle corresponds to the result of one of the 1000 Monte Carlo runs. The performance of SURE and of the oracle is quite similar.

Figure 4 also compares the estimated risks $J_A(\gamma)$ and $J_B(\gamma, E')$ against the indexes $RSS_A(\gamma)$ and $RSS_B(\gamma, E')$ in (39) and (40) in the first Monte Carlo run. The related function estimates are visible in Figure 5.

Figure 4. Comparison of the predictive performance of the estimators $\hat{f}_A$ (left panel) and $\hat{f}_B$ (right panel) over the test set in Figure 3 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.

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 behaviour. 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. 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.

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. Shiyun, 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.
Figure 5. Visualization of the training and test sets (top panel, respectively 173 and 87 samples), and of the estimates returned by “$\hat{f}_A + \text{SURE}$”, “$\hat{f}_A + \text{oracle}$”, “$\hat{f}_B + \text{SURE}$”, and “$\hat{f}_B + \text{oracle}$” in the first Monte Carlo run.
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 11** Let $\Omega$ denote a sample space, $\omega \in \Omega$ its generic element. Let $\mathcal{E}$ be an event such that
\[ P[\omega \in \mathcal{E}] \geq 1 - \alpha. \] (41)

If $g(\omega)$ is positive and (41) holds then
\[ E[g(\omega) \mid \omega \in \mathcal{E}] \leq \frac{1}{1 - \alpha} E[g(\omega)]. \]

**Proof of Lemma 11**: Let $\eta$ be the probability measure on the $\sigma$-algebra $\Omega$ equipped with. In general, for every $\mathcal{E}'$,
\[ P[\omega \in \mathcal{E}' \cap \mathcal{E}] = \frac{P[\omega \in \mathcal{E}]}{P[\omega \in \mathcal{E}]} \leq \frac{P[\omega \in \mathcal{E}]}{1 - \alpha} \leq \frac{1}{1 - \alpha} P[\omega \in \mathcal{E}]. \]

If $\eta_\mathcal{E}$ denotes the probability measure $\eta$ conditional on $\mathcal{E}$, one then has
\[ \int_\mathcal{E} g(\omega)d\eta_\mathcal{E}(\omega) \leq \frac{1}{1 - \alpha} \int_\Omega g(\omega)d\eta(\omega) \leq \frac{1}{1 - \alpha} \int_\Omega g(\omega)d\eta(\omega). \]

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 $\varepsilon$ entering the bounds.

**Lemma 12** 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 $\lambda_{\min}$ and $\lambda_{\max}$ as specified in [42] and [43]. If $E$, $M$, and $k$ in (10) satisfy (26) then
\[ P[\lambda_{\min} \left( \frac{G^T G}{M} \right) \geq \varepsilon] \geq 1 - \alpha. \] (42)

If instead $E$, $M$, and $k$ satisfy (26) then
\[ P[\lambda_{\min} \left( \frac{G^T G}{M} \right) \geq \varepsilon] \cap \lambda_{\max} \left( \frac{G^T G}{M} \right) \leq 2 - \varepsilon \] (43)

**Proof of Lemma 12**: Since the assumptions in [41] Thm. 1.1 are satisfied, for any $\varepsilon \in (0, 1]$ one has
\[ P \left[ \frac{\lambda_{\min} \left( \frac{G^T G}{M} \right)}{E} \leq \varepsilon \right] \leq E \left( \frac{e^{-(1-\varepsilon) \frac{1}{\varepsilon}}}{e \varepsilon} \right) \geq (2 - \varepsilon)(2 - \varepsilon). \] (44)

Condition (26) is obtained by picking $\alpha$ larger than the Right Hand Side (RHS) of (44) and manipulating this inequality. Then, (42) follows from (44) just considering that if $\overline{\mathcal{T}}$ is the complementary of $\mathcal{T}$ then $P[\mathcal{T}] \leq \alpha \Leftrightarrow P[\overline{\mathcal{T}}] \geq 1 - \alpha$. Now, we can use again (41) Thm. 1.1 to claim that, for every $\varepsilon \in [0, 1]$,
\[ P \left[ \frac{\lambda_{\max} \left( \frac{G^T G}{M} \right)}{E} \geq 2 - \varepsilon \right] \leq E \left( \frac{e^{-(1-\varepsilon) \frac{1}{\varepsilon}}}{e \varepsilon} \right) \geq (2 - \varepsilon)(2 - \varepsilon). \] (45)

Let the arguments in the $P[\cdot]$ in the Left Hand Side (LHS) of (44) and (45) be respectively $\star_{\lambda_{\min}}$ and $\star_{\lambda_{\max}}$. Let also the RHSs of (44) and (45) be upper bounded respectively by $\alpha_{\lambda_{\min}}$ and $\alpha_{\lambda_{\max}}$. Then, it follows that
\[ P \left[ \star_{\lambda_{\min}} \cup \star_{\lambda_{\max}} \right] \leq P[\star_{\lambda_{\min}}] + P[\star_{\lambda_{\max}}] \leq \alpha_{\lambda_{\min}} + \alpha_{\lambda_{\max}} \leq 2\alpha_{\lambda_{\min}} \]

with the last inequality following from the fact that $\varepsilon \in [0, 1] \Rightarrow \alpha_{\lambda_{\min}} \geq \alpha_{\lambda_{\max}}$ since
\[ \frac{e^{-(1-\varepsilon) \frac{1}{\varepsilon}}}{e \varepsilon} \geq \alpha_{\lambda_{\min}} \geq \frac{e^{-(1-\varepsilon) \frac{1}{\varepsilon}}}{e \varepsilon} \geq (2 - \varepsilon)(2 - \varepsilon). \]
Thus, letting the novel \( \alpha \) be \( 2\alpha_{\lambda_{\text{min}}} \) (i.e., assuming (29) to be satisfied), we obtain
\[
P[\bar{\tau}_{\text{min}} \cap \bar{\tau}_{\text{max}}] \geq 1 - \alpha,
\]
and this proves (33).

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 13** Let \( g(x) \) denote a stochastic variable whose randomness derives from the input locations \( x := [x_1, \ldots, x_M]^T \).
Assume that \( g(x) \geq q \) almost surely with \( q \) independent of \( M \). In addition, assume also that for any \( 1 - \alpha \) and \( \epsilon > 0 \) there exists \( M_0 \) such that \( \forall M \geq M_0 \) one has
\[
\overline{g(x)} \leq q + \epsilon \quad \text{with probability} \quad 1 - \alpha,
\]
in accordance with Definition 3. Then
\[
\lim_{M \rightarrow +\infty} g(x) = q \quad \text{in probability}.
\]

**B. Proof of Theorem 4**

We start by computing the general expression for \( \text{Err}_A(x) \) in (20), then evaluate its expectation like in (22), and finally transport the results to the case of \( \text{Err}_B(x) \).

As for finding the general expression for \( \text{Err}_A(x) \), we recall the decomposition of the estimand as \( f = f_a + f_b \) in (7), the definition of \( S \) in (9) and the design requirement \( \hat{f}_A, \hat{f}_B \in S \), that imply \( f_a, \hat{f}_A \in S \) and \( f_b \in S^\perp \). By construction, then, \( \|f\|^2 = \|f_a\|^2 + \|f_b\|^2 \) and
\[
\mathbb{E}\left[\|f - \hat{f}_A\|^2 \mid x\right] = \mathbb{E}\left[\|f_a - \hat{f}_A\|^2 \mid x\right] + \|f_b\|^2
\]
where the expectations are w.r.t. the noises \( \nu \), so that \( \mathbb{E}\left[\|f_b\|^2 \mid x\right] = \|f_b\|^2 \) since \( \nu, f_b \) and \( x \) are all mutually independent. Notice that a similar decomposition holds also for \( \hat{f}_B \).

As for \( \mathbb{E}\left[\|f_a - \hat{f}_A\|^2 \mid x\right] \) in (46), we notice that \( \|\hat{f}_A\|^2 = \|\hat{a}\|^2 = \|H_A y\|^2_2 \). Since (14) implies
\[
\hat{a} = H_A (Ga + Zb + \nu),
\]
and since both \( a \perp b \) and \( \nu \perp b \), it eventually follows that
\[
\text{Err}_A(x) = \mathbb{E}\left[\|a - H_A(Ga + \nu)\|^2_2 \mid x\right] + \mathbb{E}\left[\|H_A Zb\|^2_2 \mid x\right] + \|f_b\|^2.
\]

**Proof of equations (27) and (28):**

Let \( \mathcal{E} \) be the event
\[
\mathcal{E} := \left\{ \lambda_{\text{min}} \left( \frac{G^T G}{M} \right)^{1/2} \geq \varepsilon \right\},
\]
and assume \( \varepsilon, \alpha, M \) and \( E \) satisfy (26). Since in this case we can apply Theorem 12 it holds that \( P\left[\mathcal{E}\right] \geq 1 - \alpha \).

We can now write the LHS of (22) as
\[
\mathbb{E}\left[\text{Err}_A(x) \mid \mathcal{E}\right] = \mathbb{E}\left[\|f - \hat{f}_A\|^2 \mid x\right] \mid \mathcal{E}.
\]
Since \( f_b \perp \mathcal{E} \), (47) implies
\[
E[\text{Err}_A(x) \mid \mathcal{E}] =
E\left[ \mathbb{E}\left[ \left\| a - H_A(Ga + \nu) \right\|^2 \mid x \right] \mid \mathcal{E} \right] + \mathbb{E}\left[ \left\| H_AZb \right\|^2 \mid x \right] + \mathbb{E}\left[ \left\| f_b \right\|^2 \right].
\] (49)

As for \( \mathbb{E}\left[ \left\| f_b \right\|^2 \right] \), we know from (7), (8b) and the mutual independence of the \( b_e \)'s, that
\[
\mathbb{E}\left[ \left\| f_b \right\|^2 \right] = \sum_{e=E+1}^{\infty} \lambda_e.
\] (50)
This term is thus an approximation error influenced only by the dimension \( E \) of our search space \( S \).

Given (50), what we actually need to bound is the first two terms in the RHS of (49). We perform this task in the next two subsections.

**Bounding \( \mathbb{E}\left[ \left\| H_AZb \right\|^2 \mid x \right] \mid \mathcal{E} \) in (49)**

Exploiting the nature of the event \( \mathcal{E} \) to bound \( H_A \) in (16), it is not difficult to prove that
\[
\mathbb{E}\left[ \left\| H_AZb \right\|^2 \mid x \right] \leq E\left[ \left( \varepsilon I_E + \frac{\sigma^2}{M} A^{-1} \right)^{-1} G^T Zb \right] \left\| b \right\|^2.
\] (51)
Defining
\[
d_e := \frac{\varepsilon M \lambda_e + \sigma^2}{M \lambda_e}, \quad e = 1, \ldots, E,
\] (52)
it follows that
\[
\left( \varepsilon I_E + \frac{\sigma^2}{M} A^{-1} \right)^{-1} = \text{diag} \left( d_1^{-1}, \ldots, d_E^{-1} \right).
\] (53)
Consider moreover that from the definition of \( f_b \) in (7), of \( b \) in (12) and of \( Z \) in (13) it follows that \( [Zb]_m = f_b(x_m) \). Let then
\[
c_e := \left[ G^T Zb \right]_e = \sum_{m=1}^M \phi_e(x_m) f_b(x_m), \quad e = 1, \ldots, E
\] (54)
so that
\[
b^T Z^T G \left( \varepsilon I_E + \frac{\sigma^2}{M} A^{-1} \right)^{-2} G^T Zb = \sum_{e=1}^E c_e^2 \left( \frac{d_e^2}{d_e} \right).
\] (55)
Combining (53) and (55), and considering that the \( d_e \)'s are deterministic, we can thus rewrite (51) as
\[
\mathbb{E}\left[ \left\| H_AZb \right\|^2 \mid x \right] \mid \mathcal{E} \leq \frac{1}{M^2} \sum_{e=1}^E \mathbb{E}\left[ c_e^2 \mid \mathcal{E} \right] \leq \frac{1}{(1-\alpha)M^2} \sum_{e=1}^E \mathbb{E}\left[ c_e^2 \right],
\] (56)
where in the last inequality we applied Lemma 11. In view of the definition of the \( c_e \)'s in (54) and the linearity of \( \mathbb{E}[\cdot] \), one has
\[
\mathbb{E}\left[ c_e^2 \right] = \sum_{m=1}^M \mathbb{E}\left[ \phi_e^2(x_m) f_b^2(x_m) \right] + \sum_{m \neq m'} \mathbb{E}\left[ \phi_e(x_m) \phi_e(x_{m'}) f_b(x_m) f_b(x_{m'}) \right].
\] (57)
As for the first term in the RHS of (57), combining (8b) with bound (10), one immediately has
\[
\mathbb{E}\left[ \phi_e^2(x_m) f_b^2(x_m) \right] \leq k \sum_{e=E+1}^{\infty} \lambda_e.
\]
As for the second term in the RHS of (57), due to the independence of the \( \{x_m\}_{m=1}^M \) we know that
\[
\mathbb{E}\left[ \phi_e(x_m) \phi_e(x_{m'}) f_b(x_m) \phi_f(x_{m'}) \mid f_b = f_b \right] = \mathbb{E}\left[ \phi_e(x_m) f_b(x_m) \mid f_b \right] \mathbb{E}\left[ \phi_e(x_{m'}) f_b(x_{m'}) \mid f_b \right].
\]
Moreover, since }e = 1, \ldots, E\text{, from the definition of } f_b \text{ in } (7) \text{ it comes that}
\[
\mathbb{E} [\phi_e (x_m) f_b (x_m) \mid f_b] = 0.
\]
Combining the two results, one has
\[
\mathbb{E} [\varepsilon_e^2] \leq k M \sum_{e = E + 1}^{+\infty} \lambda_e \quad e = 1, \ldots, E.
\] (58)

Finally, combining (52), (56), (58) and Lemma 11 one obtains
\[
\mathbb{E} \left[ \sum_{e = 1}^{E} \lambda_e^2 \left( \frac{\varepsilon M \lambda_e + \sigma^2}{2} \right)^2 \right] \leq \frac{k M}{1 - \alpha} \left( \sum_{e = 1}^{E} \lambda_e \right)^2 \left( \sum_{e = E + 1}^{+\infty} \lambda_e \right).
\] (59)

Bounding \( \mathbb{E} \left[ \|a - H_A (Ga + \nu)\|^2 \mid x \right] \mid \mathcal{E} \) in (49)

To characterize \( \|a - H_A (Ga + \nu)\|^2 \), 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 \( b \) is not involved). More precisely, if the measurements models conditional on \( x \) were
\[
y = Ga + \nu, \quad a \sim \mathcal{N}(0, \Lambda), \quad \nu \sim \mathcal{N}(0, \sigma^2_{\nu})
\] (60)
with \( a \) independent of \( \nu \), the optimal estimator would indeed be
\[
\hat{a}_A = H_A y = H_A (Ga + \nu).
\] (61)

Exploiting standard results on Gaussian estimation, the covariance matrix of the error is
\[
\text{var} (a - H_A (Ga + \nu) \mid x) = \frac{\sigma^2_{\nu} M}{M} \left( \frac{G^T G}{M} + \frac{\sigma^2_{\nu}}{M} \Lambda^{-1} \right)^{-1}.
\] (62)

Applying Lemma 11 and using (48) to bound \( H_A \) in (16), from definitions (52) and (53) it follows that
\[
\mathbb{E} \left[ \sum_{e = 1}^{E} \lambda_e \left( \sum_{e = E + 1}^{+\infty} \lambda_e \right)^2 \right] \leq \frac{k M}{1 - \alpha} \left( \sum_{e = 1}^{E} \lambda_e \right)^2 \left( \sum_{e = E + 1}^{+\infty} \lambda_e \right).
\] (63)

Hence, the bound on \( \text{Err}_A (x) \) is obtained by combining (50), (59) and (63).

Proof of equations (30) and (31):

Let \( \mathcal{E} \) be now the event
\[
\mathcal{E} := \left\{ \lambda_{\min} \left( \frac{G^T G}{M} \right) \geq \varepsilon \cap \lambda_{\max} \left( \frac{G^T G}{M} \right) \leq 2 - \varepsilon \right\},
\]
and assume that \( \varepsilon, \alpha, M \) and \( E \) satisfy (29). Substituting \( H_A \) with \( H_B \) in the derivation of (49) one obtains
\[
\mathbb{E} \left[ \text{Err}_B (x) \mid \mathcal{E} \right] = \mathbb{E} \left[ \mathbb{E} \left[ \|a - H_B (Ga + \nu)\|^2 \mid x \right] \mid \mathcal{E} \right] + \mathbb{E} \left[ \|H_B Z_b\|^2 \mid x \right] \left( \mathcal{E} \right) + \mathbb{E} \left[ \|f_b\|^2 \right] \left( \mathcal{E} \right)
\] (64)

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 (64).

Bounding \( \mathbb{E} \left[ \|H_B Z_b\|^2 \mid x \right] \mid \mathcal{E} \) in (64)

From the definition of \( H_B \) in (19), one has
\[
\mathbb{E} \left[ \|H_B Z_b\|^2 \mid x \right] \mid \mathcal{E} \leq \mathbb{E} \left[ \left( I_E + \frac{\sigma^2_{\nu}}{M} \Lambda_{E}^{-1} \right)^{-1} \frac{G^T Z}{M} b \right]^2 \mid \mathcal{E} \]
\]
which corresponds to (51) with \( \varepsilon = 1 \). Thus, we just need to plug \( \varepsilon = 1 \) in (59) to obtain the desired result, i.e.
\[
\mathbb{E} \left[ \|H_B Z_b\|^2 \mid x \right] \mid \mathcal{E} \leq \frac{k M}{1 - \alpha} \left( \sum_{e = 1}^{E} \lambda_e \right)^2 \left( \sum_{e = E + 1}^{+\infty} \lambda_e \right).
\]
Then, there exists and according to Theorem 2. Then, from the expressions of the bounds (28) and (31), it is immediate to verify that, for fixed $E$

$$
\tilde{\alpha}_B = H_B y
$$

and

$$
\Phi_B = \text{var} (a - \tilde{\alpha}_B | x).
$$

Recalling (61) and (62), let also

$$
\tilde{\alpha}_A = H_A y
$$

and

$$
\Phi_A = \text{var} (a - \tilde{\alpha}_A | x) = \frac{\sigma_B^2}{M} \left( \frac{G^T G}{M} + \frac{\sigma_B^2}{M} \Lambda \right)^{-1}.
$$

After some simple calculations one obtains

$$
\Phi_B = \Phi_A + (H_A - H_B)(G\Lambda G^T + \sigma_B^2 I)(H_A - H_B)^T.
$$

The bound for $\mathbb{E} [\text{tr} (\Phi_A) | \overline{x}]$ was already obtained in (63) so that now we can just focus on bounding $\mathbb{E} [\text{tr} (\tilde{\Phi}) | \overline{x}]$.

Define

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

Then, it follows that

$$
\tilde{\Phi} = (A^{-1} - B^{-1}) \left( \frac{G^T G}{M} \Lambda \frac{G^T G}{M} + \frac{\sigma_B^2}{M} \Lambda \right) (A^{-1} - B^{-1})
$$

and

$$
= A^{-1}(B - A)B^{-1} \left( \frac{G^T G}{M} \Lambda \frac{G^T G}{M} + \frac{\sigma_B^2}{M} \Lambda \right)
$$

so that

$$
\text{tr} (\tilde{\Phi}) = \text{tr} \left( B^{-1} C A^{-2} C B^{-1} \left( \frac{G^T G}{M} \Lambda \frac{G^T G}{M} + \frac{\sigma_B^2}{M} \Lambda \right) \right)
$$

$$
= \text{tr} \left( A^{-1/2} \frac{G^T G}{M} B^{-1} C A^{-2} C B^{-1} \frac{G^T G}{M} A^{-1/2} \right)
$$

$$
+ \text{tr} \left( \sigma_B^2 A^{-1} C B^{-1} \frac{G^T G}{M} B^{-1} C A^{-1} \right).
$$

Since

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

we obtain

$$
A \geq \left( \epsilon + \frac{\lambda_1^{-1} \sigma_B^2}{M} \right) I,
$$

$$
B \geq \left( 1 + \frac{\lambda_1 \sigma_B^2}{M} \right) I \geq \left( \epsilon + \frac{\lambda_1 \sigma_B^2}{M} \right) I,
$$

$$
C^2 = \left( I - \frac{G^T G}{M} \right)^2 \leq (1 - \epsilon)^2 I.
$$

Exploiting such inequalities and Lemma 11 we obtain

$$
\mathbb{E} [\text{tr} (\tilde{\Phi}) | \overline{x}] \leq \left( \epsilon + \frac{\lambda_1^{-1} \sigma_B^2}{M} \right)^{-4} (1 - \epsilon)^2 \left( 2 - \epsilon \right)^2 \frac{E \sigma_B^2 I}{M} + \frac{E}{\sum_{e=1}^{+\infty} \lambda_e}
$$

which, combined with (63), provides the desired result, also leading to the overall bound for $\text{Err}_B(x)$.

C. Proof of Theorem 5

For both the cases the proof is obtained using Lemma 13. In particular, for both the estimators the lower bound is

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

according to Theorem 2. Then, from the expressions of the bounds (28) and (31), it is immediate to verify that, for fixed $E$ and $\alpha$, they are monotonically decreasing in $M$. Now, for any $0 < \delta < 1$, let $\alpha = \delta$ and $\epsilon$ such that $\kappa < \delta \frac{4 \delta}{1 - \delta} \sum_{e=1}^{+\infty} \lambda_e$. Then, there exists $M_0$ such that for $M > M_0$ conditions (26) and (29) are satisfied and

$$
\text{Err}_A \leq q + \delta
$$

with probability $1 - \delta$.
and 
\[ \text{Err}_B \leq q + \delta \quad \text{with probability} \quad 1 - \delta. \]

As anticipated, the use of Lemma 13 then concludes the proof.

**D. Proof of Theorem 6**

For what regards (33), it is sufficient to recall that, for finite \( M \), as \( E \rightarrow +\infty \) the estimator \( \hat{f}_A \) coincides with the MAP estimator which is consistent, see [42] [Appendix] for details.

**E. Proof of Theorem 7**

The convergence (34) related to \( \hat{f}_B \) is more delicate and we will exploit bound (30). The rationale is to establish conditions on the convergence of \( E, M \) to infinity to make both the confidence level \( \alpha \) and the bound (31) tend to zero. As for \( \alpha \), we can make it vanish to zero by setting \( \alpha = \frac{1}{E^{\delta}} \) with \( \delta \in (0,1) \). As for the bound (31), its first components naturally vanish with \( E \rightarrow +\infty \), while the last two do not. In particular, one needs \( \kappa \) in (32) to go to zero and this requires \( \varepsilon \rightarrow 1 \). Hence, we set \( \varepsilon \) such that \( \frac{(1-\varepsilon)^2}{2E} = \frac{1}{M^{1-\varepsilon}} \). However, the convergence of \( \alpha \) and (31) is not enough, since condition (29) must be always satisfied. This condition is indeed verified since
\[
\frac{Ek}{M} \log \frac{E}{\alpha} = \frac{Ek}{M} \log E^\delta \leq \frac{kM^\delta}{M} = \frac{(1-\varepsilon)^2}{2} \leq 1 - \varepsilon + \varepsilon \log \varepsilon.
\]