Statistical bounds for distributed Gaussian regression algorithms

Gianluigi Pillonetto * **Luca Schenato** * **Damiano Varagnolo** **

* *Department of Information Engineering, University of Padova, Padova, Italy*
** *Department of Computer Science, Electrical and Space Engineering, Lulea University of Technology, Lulea, Sweden*

Abstract We consider distributed nonparametric function estimation in the framework of Gaussian regression, i.e., with the estimand being modelled as a Gaussian random field whose covariance (kernel) encodes expected properties like smoothness. We assume that some agents with limited computational and communication capabilities collect M direct and noisy measurements of the unknown map on input locations drawn from a common probability density. Moreover we assume that their goal is to collaborate for obtaining a common and shared estimate of the estimand. For large number of measurements M, distributedly computing the minimum variance estimate is difficult: to do so one has first to exchange all the measurements and the corresponding input locations, plus invert an $M \times M$ matrix. To overcome this limitation a possibility suggested in the existing literature is to perform an opportune Karhunen-Loève (KL) expansion of the kernel and then approximate the estimand as belonging to the space spanned by a finite number of kernel eigenfunctions. In this paper we characterize statistically this strategy by providing a rigorous probabilistic bound which returns crucial information on M and the number of eigenfunctions which the network needs to exchange to obtain a certain level of accuracy in the estimate.

Keywords: Gaussian processes, distributed estimation, reproducing kernel Hilbert spaces, regularization, nonparametric estimation, average consensus

1. INTRODUCTION

Many modern engineering problems involve networks containing a large number of agents which have to cooperate to obtain a common goal. Examples include estimation of the wind speed and direction field in a wind farm from local measurements of the turbines, or the reconstruction of the temperature field in a datacenter from local measurements at each server. Even if suitable when the size of these networks is small, centralized estimation approaches are non-scalable, and it may be preferred to implement distributed cooperation approaches Xu et al. (2015).

In this paper we consider the problem of distributed nonparametric function estimation via Gaussian regression. We model the estimand as a Gaussian random field whose covariance (also called kernel in the machine learning literature) has to embed expected properties like smoothness. In this framework it is typically assumed that agents first collect M direct and noisy measurements of the unknown map on input locations drawn from a common (and known) probability density, then aim at obtaining a common function estimate. Another typical assumption is that the agents are equipped with limited computational and data storage capabilities and can communicate only with a restricted number of neighbors.

⋆ The research leading to these results has received funding from the Swedish research council Norrbottens Forskningsrad. Corresponding author: Damiano Varagnolo, `damiano.varagnolo@ltu.se`

Achieving the minimum variance estimate of the function in distributed settings is complicated, since its computation requires the agents first to exchange a great amount of information (in practice all the measurements and the input locations where they have been collected), and then to invert an $M \times M$ matrix (see Section 2).

To face this problem it has been proposed to use opportune KL expansions (computed w.r.t. the probability density governing the extraction of the input locations). A strategy is indeed to approximate the Gaussian random field by the E kernel eigenfunctions associated to the largest eigenvalues. This is the best approximation of the process before seeing the data. A posteriori (i.e., after seeing the data) the situation is instead more subtle. In fact, there exist basis functions that depend on the input locations where data are collected and approximate better the minimum variance estimator Trecate et al. (1999).

However, the a-priori basis has still some advantages: first, as proved in Zhu et al. (1998), the first E kernel eigenfunctions describing the Gaussian field are however asymptotically optimal, i.e. when the number of measurements M grows to infinity; second, differently from the a-posteriori basis described in Trecate et al. (1999), the a-priori basis can be computed off-line; third, the a-priori basis leads to estimators that are amenable to distributed computations. All these motivations suggest the use of these KL-based E-dimensional approximations. The so-derived estimators can then provide an accurate approximation of the minimum variance estimate by performing consensus over an $E \times E$ matrix, with possibly $E \ll M$.

Literature review: this paper complements our previous efforts Varagnolo et al. (2010), where we identified some sufficient conditions on M that guarantee the statistical meaningfulness of performing distributed average-consensus based estimates, and Varagnolo et al. (2012), and also derived other statistical error bounds connected with the same estimator analyzed in this paper. These bounds however apply to a different framework, since they state how much the uncertainty in the prior information is going to affect the estimate, while the bounds analyzed here deal with the problem of deciding a suitable dimension E of the hypothesis space (this is further remarked in the statement of contributions below).

Our stream of research pairs the ones of other authors, also focusing on distributed kernel regression. An example is Predd et al. (2009), that proposes a distributed regularized kernel Least Squares (LS) regression algorithm that exploits successive orthogonal projections, or Perez-Cruz and Kulkarni (2010), that extends Predd et al. (2009) by proposing some mechanisms for reducing the assumptions on the communication burden and synchronization needs.

Estimators with reduced order model complexity are proposed also in Honeine et al. (2008): here the agents construct an estimate considering only a subset of the representing functions that would be used to form the optimal solution. Nonparametric schemes are applied also in Martinez (2010), where the mobile network distributely estimates a noisily sampled scalar random field through opportune Nearest-Neighbors interpolation schemes. Another Gaussian estimation approach is considered in Xu et al. (2013), with focus on the problem of sequentially predicting the most informative locations of future measurements to minimize simultaneously the prediction error and the uncertainty in the hyperparameters of the prior. Other distributed regression algorithms are proposed in Cortés (2009) which introduces introduces an algorithm used to estimate a dynamic Gaussian random field and its gradient (in this particular case agents estimate their own neighborhood and not to the global scenario, differently from the approach considered here). In the same framework, in Choi et al. (2009) authors develop a distributed learning and cooperative control algorithm where agents estimate a static field modeled as a network of radial basis functions whose number and centers location are known in advance by agents.

Statement of contributions: to the best of our knowledge, the fundamental question that has not been answered up to now is how to choose the dimension E of the distributed estimator, i.e. the number of basis functions necessary to well approximate the minimum variance estimator. Remarkably, one cannot find in the literature a bound on the distance between the unknown function and the E-dimensional approximation of the minimum variance estimator as a function of the number of data points M, the kernel nature and the input locations statistics. The main contribution of this paper is to provide such missing bound, measuring the error by the L_2-norm weighted by the input locations probability density. Our results thus provide crucial information on the prediction capability of the distributed estimator.

The analysis reported in this paper can be also seen as the extension to the Bayesian context of the concept of effective dimension developed in deterministic frameworks, e.g., in Zhang (2005). 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 and reveals which subspace of the unknown function can be really influenced by the measurements.

Structure of the manuscript Section 2 defines the Bayesian estimation problem. Section 3 describes the KL expansion of the Gaussian random field used to define the distributed estimator. Section 4 presents the distributed estimator. Section 5 characterizes it and provides the main contribution of this manuscript. Section 6 shows numerically the accuracy of the obtained bound. Section 7 collects some conclusions and future research directions. Proofs and a summary of the notation (to ease the readability of the documents) are collected in the appendix.

2. THE BAYESIAN ESTIMATION PROBLEM

2.1 The measurements model

We assume the scalar measurements model

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

with the input locations x_m following the stochastic generation scheme

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

with μ a non-degenerate measure (w.r.t. Lebesgue) on the compact $\mathcal{X}$. The unknown function is modeled as

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

i.e., $f : \mathcal{X} \to \mathbb{R}$ is a zero-mean Gaussian random field with continuous covariance $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Finally, the measurement noise is

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

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

2.2 The Bayesian estimator

The Gaussian assumptions of Section 2.1 imply that the posterior of f given the dataset $\{x_m, y_m\}_{m=1}^M$ is again Gaussian, and that the Maximum A Posteriori (MAP) (or minimum variance) estimator is

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

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 associated to the computation of $\widehat{f}_{\text{MAP}}(\cdot)$ are thus, respectively, $O\left(M^2\right)$ and $O\left(M^3\right)$. The communication complexity, moreover, is either $d \cdot O\left(M\right)$ (with d the dimensionality of $\mathcal{X}$, in case agents share the input locations x_m) or $O\left(M^2\right)$ (in case agents share the covariances $K(x_m, x_{m'})$).

Storage, computational and communication complexities of the MAP estimators thus scale non favorably with the dataset size M, and this pushes for finding approximate estimators with more favorable scalability properties.

Our aim is thus to find approximators of $\widehat{f}_{MAP}(\cdot)$ that are both suitable for distributed implementations and have favorable statistical properties in a sense specified in Section 5.

3. REFORMULATING THE MEASUREMENTS MODEL (1) THROUGH A KL EXPANSION

Our first step is to approximate the MAP estimator $\widehat{f}_{\text{MAP}}(\cdot)$ through an opportune KL expansion.

Thanks to the assumptions in Section 2.1, we expand the kernel K in (3) in terms of that eigenfunctions of K that are orthonormal w.r.t. the measure μ in (2). In other words, we let these eigenfunctions be defined by

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

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

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 that has been fixed a priori. Then (5), (4) and (6) allow us to reformulate f via a KL expansion of the form

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

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

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

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

Importantly, for every fixed and finite E the subspace

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

is, thanks to the KL interpretation of K, that E-dimensional subspace that captures the biggest part of the statistical energy of the estimand. In other words, $f_a(x)$ is that E-dimensional part of $f(x)$ that captures the biggest statistical energy of $f(x)$, while $f_b(x)$ can be considered a remainder.

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.

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

Assumption 1 is necessary for the subsequent formal derivations. It is naturally satisfied for all the finite-dimensional kernels and also several of the classical infinite-dimensional ones (e.g., the splines kernels). More in general, approximation of the KL expansion of kernels like the Gaussian and Laplacian can be numerically obtained with arbitrary accuracy, obtaining also the value of the constant k.

4. A FINITE-DIMENSIONAL APPROXIMATION OF THE MAP ESTIMATOR

Our next step is to search for a finite-dimensional estimator of f that is suitable for distributed implementations.

Given the KL interpretation in Section 3, we force our estimator $\widehat{f}$ to assume values in the finite-dimensional subspace $\mathcal{S}$ defined in (9). The explanation of why $\widehat{f}$ is an approximation of $\widehat{f}_{\text{MAP}}$, along with its statistical characterization, is delegated to Section 5.

4.1 Rewriting model (1) in a compact form

Let

$$\boldsymbol{x} := \left[x_1, \ldots, x_M\right]^T$$

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

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

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

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

$$Z_{me} := \phi_{E+e}(x_m), \quad m = 1, \ldots, M, \; e = 1, 2, \ldots \tag{13}$$

Consider decomposition (7) and the definitions (11)-(13). Using classical algebraic notation to handle also infinite-dimensional objects we can then compact the measurements model (1) into

$$\boldsymbol{y} = G\boldsymbol{a} + Z\boldsymbol{b} + \boldsymbol{\nu}. \tag{14}$$

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

4.2 The finite-dimensional estimator $\widehat{f}$

Let

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

where

$$H := \left(\frac{G^T G}{M} + \frac{\sigma_\nu^2}{M} \Lambda_E^{-1} \right)^{-1} \frac{G^T}{M} \tag{15}$$

and where $\Lambda_E := \operatorname{diag}(\lambda_1, \ldots, \lambda_E)$.

Estimator $\widehat{f}$ is suitable for distributed computations in the following sense: defining

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

one has

$$\frac{G^T G}{M} = \sum_{m=1}^{M} \frac{G_m^T G_m}{M}, \qquad \frac{G^T \boldsymbol{y}}{M} = \sum_{m=1}^{M} \frac{G_m^T y_m}{M}. \tag{16}$$

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, (16) indicates that $\widehat{f}$ 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. Incidentally, we notice that average consensus can be implemented in networks with delays and dynamically changing directed graph topologies with failing communication links Hadjicostis and Charalambous (2012).

5. STATISTICAL CHARACTERIZATION

As performance indexes for $\widehat{f}$ we consider the conditional expectation

$$\mathrm{Err} := \mathbb{E}\left[\left\| f - \widehat{f} \right\|^2 \mid \boldsymbol{x} \right]$$

where $\| \cdot \|$ is the norm induced by μ, i.e.,

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

In our settings Err is stochastic, since it is function of the random input locations $\{x_m\}_{m=1}^M$.

The following Theorem 2 states a lower bound on the performance achievable by a generic estimator of f that comes directly from the KL expansion introduced in Section 3.

Theorem 2. Let $\widehat{f}_\star$ be any generic estimator of f that is function of $\boldsymbol{y}$ and that takes values in any generic E-dimensional space that has been fixed a-priori. Then

$$\arg\min_{\widehat{f}_\star} \mathbb{E}\left[\left\| f - \widehat{f}_\star \right\|^2 \mid \boldsymbol{x} \right] \geq \sum_{e=E+1}^{+\infty} \lambda_e. \tag{17}$$

The next important result is that $\widehat{f}$ asymptotically reaches the bound (17). In combination with the above theorem, this implies that $\mathcal{S}$ is asymptotically the optimal range of finite-dimensional approximations of the minimum variance estimators.

Theorem 3. Given the assumptions in Section 2.1 and Assumption 1,

$$\lim_{M \to +\infty} \mathrm{Err} = \sum_{e=E+1}^{+\infty} \lambda_e \qquad \text{in probability} \tag{18}$$

To obtain further insight on $\widehat{f}$, we now bound its performance index Err for any finite number of measurements M and opportune number of eigenfunctions E. To this aim we need to introduce some probabilistic results on the eigenvalues of the matrix $\frac{G^T G}{M}$. First of all we notice that, as stated in (6),

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

and that, given the assumptions in Section 2.1 and Assumption 1,

$$\frac{G^T G}{M} = \frac{1}{M} \sum_{m=1}^{M} G_m^T G_m \xrightarrow{M \to +\infty} \mathbb{E}\left[\frac{G^T G}{M} \right] = I. \tag{19}$$

(19) obviously implies

$$\lambda_{\min}\left(\mathbb{E}\left[\frac{G^T G}{M} \right] \right) = \lambda_{\max}\left(\mathbb{E}\left[\frac{G^T G}{M} \right] \right) = 1.$$

Moreover, since we also assume both $M \geq E$ and the measure $\mu(\mathcal{X})$ to be non-trivial, the rank of the matrix $\frac{G^T G}{M}$ will be full almost surely; thus, also, $\mathbb{P}\left[\lambda_{\min}\left(\frac{G^T G}{M} \right) > 0 \right] = 1$. Finally,

Theorem 4. 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 deviance index for $\lambda_{\min}$ and $\lambda_{\max}$ as specified in (21). If E, M and k in (10) satisfy

$$1 - \varepsilon + \varepsilon \log(\varepsilon) \geq \frac{Ek}{M} \log\left(\frac{E}{\alpha} \right) \tag{20}$$

then

$$\mathbb{P}\left[\lambda_{\min}\left(\frac{G^T G}{M} \right) \geq \varepsilon \right] \geq 1 - \alpha. \tag{21}$$

Thanks to Theorem 4 we can claim the following:

Theorem 5. Let the assumptions in Section 2.1 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 a given deviance index for $\lambda_{\min}$ and $\lambda_{\max}$ as in (21). If moreover E, M and k in (10) satisfy (20) then with probability at least $1 - \alpha$ it holds that

$$\mathrm{Err} \leq \mathrm{Bnd}(E)$$

with

$$\begin{aligned} \mathrm{Bnd}(E) := {} & \frac{kM}{1-\alpha} \left(\sum_{e=1}^{E} \frac{\lambda_e^2}{(\varepsilon M \lambda_e + \sigma_\nu^2)^2} \right) \left(\sum_{e=E+1}^{+\infty} \lambda_e \right) \\ & + \frac{\sigma_\nu^2}{1-\alpha} \left(\sum_{e=1}^{E} \frac{\lambda_e}{\varepsilon M \lambda_e + \sigma_\nu^2} \right) + \left(\sum_{e=E+1}^{+\infty} \lambda_e \right). \end{aligned} \tag{22}$$

It is worth stressing that $\mathrm{Bnd}(E)$ holds for any possible stochastic machinery that generate the x_m. In particular, it depens on the input locations distribution and the adopted kernel only through the eigenvalues λ_e.

The dependency of $\mathrm{Bnd}(E)$ on α, ε, M and k is assumed tacit. The rationale for this notation is that, assuming α, ε, M and k to be given, then there is an interval [1] $1, \ldots, E_{\max}$ for which any E in this interval satisfies (20).

Given this interpretation, Figure 1 shows the dependency of $\mathrm{Bnd}(E)$ on E for some fixed M, ε and α. Bound $\mathrm{Bnd}(E)$

[1] Implicitly we assume M sufficiently big to guarantee $E_{\max} \geq 1$.

is monotonically decreasing with E, that means that, as for $\widehat{f}$, the bigger the E the better the bound.

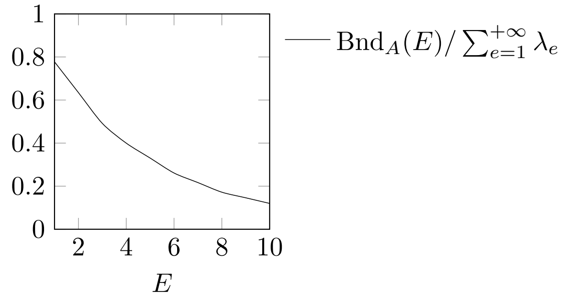

Figure 1. Dependency of $\mathrm{Bnd}(E)$ (normalized by the a-priori variance of the estimand) on E for $M = 10000$, $\varepsilon = 0.75$, $\alpha = 0.05$ and σ_ν^2 with a Gaussian kernel as in (23) (implying $k \approx 5.3$).

6. NUMERICAL ASSESSMENTS

As the prior K we use the Gaussian kernel over the uniform input locations measure $\mathcal{X} = [0,1] \times [0,1]$ defined by

$$K\left(x_1, x_2; x_1', x_2'\right) = \exp\left(-\frac{\left(x_1 - x_1'\right)^2 + \left(x_2 - x_2'\right)^2}{0.02}\right) \tag{23}$$

and a measurement noise level $\sigma_\nu^2 = 0.05$.

To describe graphically the effectiveness of our bound we then plot in Figures 2 and 3 respectively the following two situations: as for 2, consider M fixed and E variable, then for every (M, E) couple perform 100 Monte-Carlo iterations of the kind "extract f and $\boldsymbol{x}$, then compute $\boldsymbol{y}$, $\widehat{f}$ and Err", then plot for that (M, E) the boxplot of Err and the value of $\mathrm{Bnd}(E)$, so to assess how much the bound is conservative. As for 3, perform the same kind of Monte-Carlo simulations but keeping E fixed and M variable.

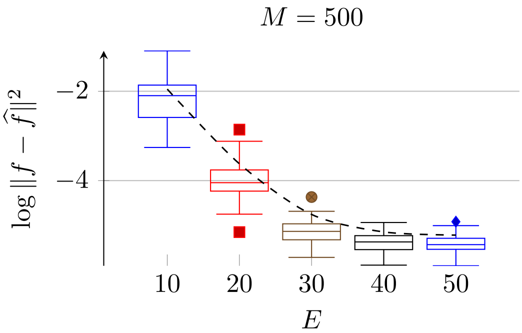

Figure 2. Monte-Carlo analysis of the conservativeness of the bound $\mathrm{Bnd}(E)$ keeping M fixed and E variable. Every boxplot summarizes the statistics for 100 independent realizations of Err, while the dashed black line represents $\mathrm{Bnd}(E)$ for the case $\alpha = 0.05$, $\varepsilon = 0.75$.

As one can notice, the bound is capturing the error in a statistical sense and provides a good indication of its size.

7. CONCLUSIONS

We statistically characterized a well-known distributed Gaussian regression algorithm that tackles estimation problems where agents have with limited computational, data storage and communication capabilities, collect M direct and noisy measurements, and aim at estimating the underlying map from these measurements.

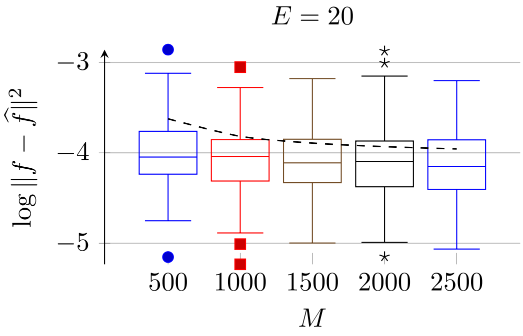

Figure 3. Monte-Carlo analysis of the conservativeness of the bound $\mathrm{Bnd}(E)$ keeping E fixed and M variable. Every boxplot summarizes the statistics for 100 independent realizations of Err, while the dashed black line represents $\mathrm{Bnd}(E)$ for the case $\alpha = 0.05$, $\varepsilon = 0.75$.

More precisely we provide a rigorous probabilistic a-priori bound on the statistical performance of the estimator, and thus find a tool that indicates how the dimension E of the hypothesis space where the estimator lives (a hyperparameter that influences the communications needs) should depend on the number of measurements M collected by the agents in order to guarantee a certain statistical performance.

In other words we contribute to the existing literature by answering the fundamental question of how to choose a-priori the hyperparameter E of the estimator based on the expected number of samples that the agents will take so to guarantee a certain given statistical performance index.

Finding this specific answer nonetheless does not conclude the topic: the bound that we found relies on the crucial hypothesis that the prior is correct. An important direction is thus now to understand what changes in the structure and the answers of the bound in case that the knowledge of the prior is imperfect.

REFERENCES

Jongeun Choi, Songhwai Oh, and Roberto Horowitz. Distributed learning and cooperative control for multi-agent systems. *Automatica*, 45(12):2802–2814, 2009.

Jorge Cortés. Distributed kriged kalman filter for spatial estimation. *IEEE Transactions on Automatic Control*, 54(12):2816–2827, 2009.

Christoforos N Hadjicostis and Themistoklis Charalambous. Average consensus in the presence of delays and dynamically changing directed graph topologies. *arXiv preprint arXiv:1210.4778*, 2012.

Paul Honeine, Mehdi Essoloh, Cédric Richard, and Hichem Snoussi. Distributed regression in sensor networks with a reduced-order kernel model. In *IEEE GLOBECOM 2008-2008 IEEE Global Telecommunications Conference*, pages 1–5. IEEE, 2008.

Sonia Martinez. Distributed interpolation schemes for field estimation by mobile sensor networks. *IEEE*

Transactions on Control Systems Technology, 18(2): 491–500, 2010.

Fernando Perez-Cruz and Sanjeev R Kulkarni. Robust and low complexity distributed kernel least squares learning in sensor networks. *IEEE Signal Processing Letters*, 17 (4):355–358, 2010.

Joel B Predd, Sanjeev R Kulkarni, and H Vincent Poor. A collaborative training algorithm for distributed learning. *IEEE Transactions on Information Theory*, 55(4): 1856–1871, 2009.

Giancarlo Ferrari Trecate, Christopher KI Williams, and Manfred Opper. Finite-dimensional approximation of gaussian processes. In *Proceedings of the 1998 conference on Advances in neural information processing systems II*, pages 218–224. MIT Press, 1999.

Joel A. Tropp. User-Friendly Tail Bounds for Sums of Random Matrices. *Foundations of Computational Mathematics*, 12(4):389–434, aug 2011.

Damiano Varagnolo, Gianluigi Pillonetto, and Luca Schenato. Distributed consensus-based bayesian estimation: sufficient conditions for performance characterization. In *Proceedings of the 2010 American Control Conference*, pages 3986–3991. IEEE, 2010.

Damiano Varagnolo, Gianluigi Pillonetto, and Luca Schenato. Distributed parametric and nonparametric regression with on-line performance bounds computation. *Automatica*, 48(10):2468–2481, 2012.

Yunfei Xu, Jongeun Choi, Sarat Dass, and Tapabrata Maiti. Efficient bayesian spatial prediction with mobile sensor networks using gaussian markov random fields. *Automatica*, 49(12):3520–3530, 2013.

Yunfei Xu, Jongeun Choi, Sarat Dass, and Tapabrata Maiti. *Bayesian Prediction and Adaptive Sampling Algorithms for Mobile Sensor Networks: Online Environmental Field Reconstruction in Space and Time*. Springer, 2015.

Tong Zhang. Learning bounds for kernel regression using effective data dimensionality. *Neural Computation*, 17 (9):2077–2098, 2005.

H Zhu, C K I Williams, R J Rohwer, and M Morciniec. Gaussian regression and optimal finite dimensional linear models. In C M Bishop, editor, *Neural Networks and Machine Learning*. Springer-Verlag, Berlin, 1998.

APPENDIX

Remark 6. For ease of notation in this appendix we will shorten $\mathbb{E}\left[\left\|f-\widehat{f}_\star\right\|^2 \mid \boldsymbol{x}\right]$ with $\mathbb{E}\left[\left\|f-\widehat{f}_\star\right\|^2\right]$. I.e., in all the expectations the conditioning on the input locations $\boldsymbol{x} := [x_1, \ldots, x_M]^T$ is tacit.

Moreover we indicate with $\star|_\mathcal{E}$ a generic r.v. $\star$ that is conditioned on a generic event $\mathcal{E}$.

Appendix A. PRELIMINARY RESULT ON CONDITIONAL EXPECTATIONS COMPUTATION

The following theorem will be used to characterize the Mean Square Error (MSE) of the proposed estimator.

Let Ω denote a sample space, $\omega \in \Omega$ its generic element, η a probability measure on a suitable σ-algebra on Ω, $\mathcal{E}$ an element of the σ-algebra s.t.

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

If $g(\omega)$ is a positive random variable on Ω (i.e., s.t. $g(\omega) \geq 0\ \forall \omega \in \Omega$) then the expectation of $g(\omega)$ conditioned on $\omega \in \mathcal{E}$ can be bounded with a scaled version of the unconditioned expectation of $g(\omega)$:

Theorem 7. If $g(\omega)$ is positive and (A.1) 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 Theorem 7: In general, for every $\mathcal{E}'$,

$$\begin{aligned}\mathbb{P}\left[\omega \in \mathcal{E}' \mid \omega \in \mathcal{E}\right] &= \frac{\mathbb{P}\left[\omega \in \mathcal{E}' \cap \mathcal{E}\right]}{\mathbb{P}\left[\omega \in \mathcal{E}\right]} \\ &\leq \frac{\mathbb{P}\left[\omega \in \mathcal{E}'\right]}{\mathbb{P}\left[\omega \in \mathcal{E}\right]} \leq \frac{1}{1-\alpha}\mathbb{P}\left[\omega \in \mathcal{E}'\right].\end{aligned} \tag{A.2}$$

If $\eta_\mathcal{E}$ denotes the probability measure η conditional on $\mathcal{E}$, thus, (A.2) implies that $\eta_\mathcal{E}(\omega) \leq \frac{1}{1-\alpha}\eta(\omega)$. Thus

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

Appendix B. PROBABILISTIC BOUNDS ON THE EIGENVALUES OF $\frac{G^T G}{M}$

Proof of (21) **in Theorem 4:** since we satisfy the assumptions in (Tropp, 2011, Thm. 1.1), we can thus claim that, for every $\varepsilon \in (0,1]$,

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

To claim (21) we then consider that: *i)* equivalence (20) follows from majorizing the Right Hand Side (RHS) of (B.1) with α (i.e., letting $\alpha \geq$ RHS) and then opportunely manipulating this inequality; *ii)* (21) follows from (B.1) by considering that if $\overline{\star}$ is the complementary of $\star$ then $\mathbb{P}\left[\star\right] \leq \alpha \Leftrightarrow \mathbb{P}\left[\overline{\star}\right] \geq 1-\alpha$. □

We eventually remark that the event "$\lambda_{\min}\left(\frac{G^T G}{M}\right) \geq \varepsilon$" in Theorem 4 can be interpreted as particular instances of "$\{x_m\}_{m=1}^M =: \omega \in \mathcal{E}$" in (A.1) where $\mathcal{E}$ is an opportune function of the threshold ε.

Appendix C. PROOF OF THEOREMS 3 AND 5

If $\overline{\mathcal{E}}$ is an event s.t. $\mathbb{P}\left[\overline{\mathcal{E}}\right] \geq 1-\alpha$ then

$$\mathbb{P}\left[\mathbb{E}\left[\left\|f-\widehat{f}\right\|^2\right] = \mathbb{E}\left[\left\|f-\widehat{f}\right\|^2 \mid \overline{\mathcal{E}}\right]\right] \geq 1-\alpha.$$

Suppose moreover that

$$\mathbb{E}\left[\left\|f-\widehat{f}\right\|^2 \mid \overline{\mathcal{E}}\right] \leq \mathrm{Bnd}(E) \tag{C.1}$$

with $\mathrm{Bnd}(E)$ defined in (22) (this will be proved from the next paragraph). This means that if $\overline{\mathcal{E}}$ is s.t. $\mathbb{P}\left[\overline{\mathcal{E}}\right] \geq 1-\alpha$ and if (C.1) holds then

$$\mathbb{P}\left[\mathbb{E}\left[\left\|f-\widehat{f}\right\|^2\right] \leq \mathrm{Bnd}(E)\right] \geq 1-\alpha. \tag{C.2}$$

Since (C.2) is the claim, the previous discussion reduces the problem to find an $\overline{\mathcal{E}}$ and a $\mathrm{Bnd}(E)$ s.t. $\mathbb{P}\left[\overline{\mathcal{E}}\right] \geq 1-\alpha$ and (C.1) hold simultaneously.

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

$$\overline{\mathcal{E}} := \left\{\lambda_{\min}\left(\frac{G^T G}{M}\right) \geq \varepsilon\right\}, \tag{C.3}$$

and assume ε, α, M and E satisfy (20). Since in this case we can apply Theorem 4, we are ensured that $\mathbb{P}\left[\overline{\mathcal{E}}\right] \geq 1-\alpha$. The next step is thus to verify that bound (22) satisfies (C.1).

To this aim, 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 $\widehat{f} \in \mathcal{S}$, that imply $f_a, \widehat{f}, \widehat{f} \in \mathcal{S}$ and $f_b \in \mathcal{S}^{\perp}$). By construction, then, $\|f\|^2 = \|f_a\|^2 + \|f_b\|^2$ and

$$\mathbb{E}\left[\left\|f-\widehat{f}\right\|^2\right] = \mathbb{E}\left[\left\|f_a-\widehat{f}\right\|^2\right] + \mathbb{E}\left[\|f_b\|^2\right]. \tag{C.4}$$

We thus proceed explicating the terms in the right hand side of (C.4).

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

$$\mathbb{E}\left[\|f_b\|^2\right] = \sum_{e=E+1}^{+\infty} \lambda_e. \tag{C.5}$$

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

As for $\mathbb{E}\left[\left\|f_a-\widehat{f}\right\|^2\right]$ in (C.4), we notice that $\left\|\widehat{f}\right\|^2 = \|\widehat{\boldsymbol{a}}\|_2^2 = \|H\boldsymbol{y}\|_2^2$. Since (14) implies

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

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

$$\mathbb{E}\left[\left\|f_a-\widehat{f}\right\|^2\right] = \mathbb{E}\left[\|\boldsymbol{a} - H(G\boldsymbol{a}+\boldsymbol{\nu})\|^2\right] + \mathbb{E}\left[\|HZ\boldsymbol{b}\|^2\right]$$

so that, summarizing, Err is in general

$$\begin{aligned}\mathbb{E}\left[\left\|f-\widehat{f}\right\|^2\right] = \; & \mathbb{E}\left[\|\boldsymbol{a} - H(G\boldsymbol{a}+\boldsymbol{\nu})\|^2\right] \\ & + \mathbb{E}\left[\|HZ\boldsymbol{b}\|^2\right] \\ & + \mathbb{E}\left[\|f_b\|^2\right].\end{aligned} \tag{C.6}$$

Since $f_b \perp \overline{\mathcal{E}}$, (C.6) in its turn implies

$$\begin{aligned}\mathbb{E}\left[\left\|f-\widehat{f}\right\|^2 \mid \overline{\mathcal{E}}\right] = \; & \mathbb{E}\left[\|\boldsymbol{a} - H(G\boldsymbol{a}+\boldsymbol{\nu})\|^2 \mid \overline{\mathcal{E}}\right] \\ & + \mathbb{E}\left[\|HZ\boldsymbol{b}\|^2 \mid \overline{\mathcal{E}}\right] \\ & + \mathbb{E}\left[\|f_b\|^2\right].\end{aligned} \tag{C.7}$$

Given (C.5), what we actually need to bound is the first two terms in the RHS of (C.7). We perform this task in the two dedicated sections Section C.1 and Section C.2.

C.1 Bounding $\mathbb{E}\left[\|HZ\boldsymbol{b}\|^2 \mid \overline{\mathcal{E}}\right]$ *in* (C.7)

Conditioning on (C.3) it is possible to minorize H in (15), so that

$$\mathbb{E}\left[\|HZ\boldsymbol{b}\|^2 \mid \overline{\mathcal{E}}\right] \leq \mathbb{E}\left[\left\|\left(\varepsilon I_E + \frac{\sigma_\nu^2}{M}\Lambda_E^{-1}\right)^{-1}\frac{G^T Z}{M}\boldsymbol{b}\right\|^2 \mid \overline{\mathcal{E}}\right]. \tag{C.8}$$

Defining the deterministic quantities

$$d_e := \frac{\varepsilon M \lambda_e + \sigma_\nu^2}{M\lambda_e}, \qquad e = 1, \ldots, E, \tag{C.9}$$

it follows that

$$\left(\varepsilon I_E + \frac{\sigma_\nu^2}{M}\Lambda_E^{-1}\right)^{-1} = \mathrm{diag}\left(d_1^{-1}, \ldots, d_E^{-1}\right). \tag{C.10}$$

Consider moreover that from the definition of f_b in (7), of $\boldsymbol{b}$ in (12) and of Z in (13) it follows that $[Z\boldsymbol{b}]_m = f_b(x_m)$. Let then

$$c_e := \left[G^T Z\boldsymbol{b}\right]_e = \sum_{m=1}^{M} \phi_e(x_m) f_b(x_m) \qquad e = 1, \ldots, E \tag{C.11}$$

so that

$$\boldsymbol{b}^T Z^T G G^T Z\boldsymbol{b} = \sum_{e=1}^{E} c_e^2\,. \tag{C.12}$$

Combining (C.10) and (C.12), and considering that the d_e's are deterministic, we can thus rewrite (C.8) as

$$\begin{aligned}\mathbb{E}\left[\|HZ\boldsymbol{b}\|^2 \mid \overline{\mathcal{E}}\right] & \leq \frac{1}{M^2}\sum_{e=1}^{E}\frac{\mathbb{E}\left[c_e^2 \mid \overline{\mathcal{E}}\right]}{d_e^2} \\ & \leq \frac{1}{(1-\alpha)M^2}\sum_{e=1}^{E}\frac{\mathbb{E}\left[c_e^2\right]}{d_e^2},\end{aligned} \tag{C.13}$$

where in the last inequality we applied Theorem 7. Considering the definition of the c_e's in (C.11) and the linearity of $\mathbb{E}[\cdot]$, then, implies

$$\begin{aligned}\mathbb{E}\left[c_e^2\right] = \; & \sum_{m=1}^{M}\mathbb{E}\left[\phi_e^2(x_m) f_b^2(x_m)\right] \\ & + \sum_{m \neq m'}\mathbb{E}\left[\phi_e(x_m)\phi_e(x_{m'}) f_b(x_m) f_b(x_{m'})\right].\end{aligned} \tag{C.14}$$

As for the first term in the RHS of (C.14), combining (8b) with bound (10) we can then state that

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

As for the second term in the RHS of (C.14), due to the independence of the $\{x_m\}_{m=1}^M$ we know that

$$\begin{aligned}&\mathbb{E}\left[\phi_e(x_m)\phi_e(x_{m'}) f_b(x_m) f_b(x_{m'})\right] = \\ &\quad \mathbb{E}\left[\phi_e(x_m) f_b(x_m)\right]\mathbb{E}\left[\phi_e(x_{m'}) f_b(x_{m'})\right].\end{aligned}$$

Moreover, due to the definition of f_b in (7), the fact that $b_e \perp x_m$, and the fact that $\mathbb{E}[b_e] = 0$ for every e, we can state that

$$\mathbb{E}\left[\phi_e\left(x_m\right)f_b\left(x_m\right)\right] = \sum_{e'=1}^{+\infty}\mathbb{E}\left[b_{e'}\right]\mathbb{E}\left[\phi_e\left(x_m\right)\phi_{E+e'}\left(x_m\right)\right] = 0.$$

Combining the two results, thus, implies

$$\mathbb{E}\left[c_e^2\right] = kM\sum_{e=E+1}^{+\infty}\lambda_e \quad e = 1,\ldots,E. \tag{C.15}$$

Combining (C.9), (C.13) and (C.15) leads thus to

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

C.2 Bounding $\mathbb{E}\left[\|\boldsymbol{a}-H(G\boldsymbol{a}+\boldsymbol{\nu})\|^2 \mid \overline{\mathcal{E}}\right]$ *in* (C.7)

To characterize $\|\boldsymbol{a}-H(G\boldsymbol{a}+\boldsymbol{\nu})\|^2$ we notice that, independently of the conditioning on $\boldsymbol{x}$ or $\overline{\mathcal{E}}$, by construction this term corresponds to the MSE of a classical MAP estimator for a standard linear Gaussian model for which the term $\boldsymbol{b}$ does not exist. Using the Woodbury matrix identity form of the variance of the error, thus, we obtain

$$\mathrm{var}\left(\boldsymbol{a}-H(G\boldsymbol{a}+\boldsymbol{\nu})\right) = \frac{\sigma_\nu^2}{M}\left(\frac{G^TG}{M}+\frac{\sigma_\nu^2}{M}\Lambda_E^{-1}\right)^{-1}.$$

Applying simultaneously Theorem 7 and the fact that, conditioning on (C.3), it is possible to minorize H in (15) and exploit definitions (C.9) and (C.10), it then follows that

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

and this concludes the proof for Theorem 5.

C.3 Proof of (18)

Given Theorem 5, we can now compute what happens to Err when $M \to +\infty$ starting from considering bound (5). The claim then follows directly from the fact that first and second terms of the bound vanish for $M \to +\infty$.