Università degli Studi di Padova

Dipartimento di Ingegneria dell'Informazione

Distributed Parametric-Nonparametric Estimation in Networked Control Systems

Ph.D. candidate
Damiano Varagnolo

Advisor
Prof. Luca Schenato

Ph.D. School in
Information Engineering
2011

Contents

Abstract

In the framework of parametric and nonparametric distributed estimation, we introduce and mathematically analyze some consensus-based regression strategies characterized by a guess of the number of agents in the network as a parameter. The parametric estimators assume a-priori information about the finite set of parameters to be estimated, while the the nonparametric use a reproducing kernel Hilbert space as the hypothesis space. The analysis of the proposed distributed regressors offers some sufficient conditions assuring the estimators to perform better, under the variance of the estimation error metric, than local optimal ones. Moreover it characterizes, under euclidean distance metrics, the performance losses of the distributed estimators with respect to centralized optimal ones. We also offer a novel on-line algorithm that distributedly computes *certificates of quality* attesting the goodness of the estimation results, and show that the nonparametric distributed regressor is an approximate distributed Regularization Network requiring small computational, communication and data storage efforts. We then analyze the problem of estimating a function from different noisy data sets collected by spatially distributed sensors and subject to unknown temporal shifts, and perform time delay estimation through the minimization of functions of inner products in reproducing kernel Hilbert spaces.

Due to the importance of the knowledge of the number of agents in the previously analyzed algorithms, we also propose a design methodology for its distributed estimation. This algorithm is based on the following paradigm: some locally randomly generated values are exchanged among the various sensors, and are then modified by known consensus-based strategies. Statistical analysis of the a-consensus values allows the estimation of the number of sensors participating in the process. The first main feature of this approach is that algorithms are completely distributed, since they do not require leader election steps. Moreover sensors are not requested to transmit authenticating information like identification numbers or similar data, and thus the strategy can be implemented even if privacy problems arise. After a rigorous formulation of the paradigm we analyze some practical examples, fully characterize them from a statistical point of view, and finally provide some general theoretical results among with asymptotic analyses.

Sommario

In questa tesi vengono introdotti e analizzati alcuni algoritmi di regressione distribuita parametrica e nonparametrica, basati su tecniche di consenso e parametrizzati da un parametro il cui significato è una stima del numero di sensori presenti nella rete. Gli algoritmi parametrici assumono la conoscenza di informazione a-priori sulle quantità da stimare, mentre quelli nonparametrici utilizzano come spazio delle ipotesi uno spazio di Hilbert a nucleo riproducente. Dall'analisi degli stimatori distribuiti proposti si ricavano alcune condizioni sufficienti che, se assicurate, garantiscono che le prestazioni degli stimatori distribuiti sono migliori di quelli locali (usando come metrica la varianza dell'errore di stima). Inoltre dalla stessa analisi si caratterizzano le perdite di prestazioni che si hanno usando gli stimatori distribuiti invece che quelli centralizzati e ottimi (usando come metrica la distanza euclidea tra le due diverse stime ottenute). Inoltre viene offerto un nuovo algoritmo che calcola in maniera distribuita dei *certificati di qualità* che garantiscono la bontà dei risultati ottenuti con gli stimatori distribuiti. Si mostra inoltre come lo stimatore nonparametrico distribuito proposto sia in realtà una versione approssimata delle cosiddette "Reti di Regolarizzazione", e come esso richieda poche risorse computazionali, di memoria e di comunicazione tra sensori. Si analizza quindi il caso di sensori spazialmente distribuiti e soggetti a ritardi temporali sconosciuti. Si mostra dunque come si possano stimare, minimizzando opportune funzioni di prodotti interni negli spazi di Hilbert precedentemente considerati, sia la funzione vista dai sensori che i relativi ritardi visti da questi.

A causa dell'importanza della conoscenza del numero di agenti negli algoritmi proposti precedentemente, viene proposta una nuova metodologia per sviluppare algoritmi di stima distribuita di tale numero, basata sulla seguente idea: come primo passo gli agenti generano localmente alcuni numeri, in maniera casuale e da una densità di probabilità nota a tutti. Quindi i sensori si scambiano e modificano questi dati usando algoritmi di consenso quali la media o il massimo; infine, tramite analisi statistiche sulla distribuzione finale dei dati modificati, si può ottenere dell'informazione su quanti agenti hanno partecipato al processo di consenso e modifica. Una caratteristica di questo approccio è che gli algoritmi sono completamente distribuiti, in

quanto non richiedono passi di elezione di leaders. Un'altra è che ai sensori non è richiesto di trasmettere informazioni sensibili quali codici identificativi o altro, quindi la strategia è implementabile anche se in presenza di problemi di riservatezza. Dopo una formulazione rigorosa del paradigma, analizziamo alcuni esempi pratici, li caratterizziamo completamente dal punto di vista statistico, e infine offriamo alcuni risultati teorici generali e analisi asintotiche.

Introduction

New low-cost technologies and wireless communication are promoting the deployment of networks, commonly referred as Networked Control Systems (NCSs), composed by a large number of devices with the capacity to sense, interact with the environment, communicate and collaborate to achieve a common objective. These networks, which popularity and diffusion is increasing, are enabling a whole new wide range of applications such as remote surveillance / environmental monitoring, indoor target tracking, multi-robot exploration and others, as listed in the surveys of Akyildiz et al. (2002) and Puccinelli and Haenggi (2005). The key assumption is that agents form a connected network: this implies that even if they might not be able to communicate directly, there exists a path that allows information to travel from any node to any other node - even if in the presence of lossy communications, bandwidth limitations and energy constraints. A more detailed example of such a system is given by the next generation power grids (Glanzmann et al., 2007) where each energy producer or user will be connected through a communication network to exchange information and estimate some unknown parameters of the network, like its efficiency, capacity, current utilization, etc.

Even if there has been a wide interest on the subject and even if methodological strategies are recently appearing (Papachristodoulou et al., 2004), the design of large scale networks of cooperating systems is still a difficult task. For example, since these networks are likely to be dynamic (i.e. new nodes can appear, disappear, or change their characteristics without warning the other agents) it is necessary to do not rely on a-priori knowledge on network topology and parameters, and be robust to node failure and dynamic changes. Moreover, these networks inherit a multitude of possible strong peculiarities from the variety of the suitable applications. As a natural consequence, these systems are posing challenging novel questions both from theoretical and practical perspectives. Examples of those are in terms of information compression (Xiao et al., 2006; Nakamura et al., 2007, and references therein), distributed learning (Predd et al., 2006c), event detection (Viswanathan and Varshney, 1997; Blum et al., 1997), and many others, just to name a few.

Among all the networked systems-related problematics expressed up to now, in

this thesis we develop tools aiding the scenario where a swarm of agents, deployed in an unknown environment, have to perform monitoring and research operations. Some practical and important cases are:

- underwater unmanned vehicles looking for illegally deployed radioactive waste;
- unmanned aerial vehicles searching for survivals in a stricken area;
- mobile robots monitoring borderlines.

In such scenarios agents have to face some additional peculiarities given by the lack of a-priori knowledge about the environment. For this reason, it is natural to assume that:

(a) it is unknown when some decision has to be taken, and which kind of decision it will be: examples are "should agents take some measurements" or "should agents communicate";

(b) it is unlikely that agents will be all in the same situation and condition: some may be moving, some may be offline. Moreover they will not have a precise knowledge on the situation of the other agents;

(c) the gathered information will not be uniform. For example, data will be neither spatially nor temporally uniformly distributed, and agents may do not know if they are obtaining information that has already been obtained -even partially- by somebody else.

A key factor for the success of these networks in these scenarios is then their ability to accommodate themselves to the unknown environment and to proactively face the difficulties and the variabilities without requiring the human intervention. In this thesis we then seek to *augment the level of autonomy of these networks*, aiming for a completely distributed and self-governing system that is unaffected by the uncertainties.

Our first effort is on the *characterization of the performance of distributed estimators*. In more details, we start seeking answers to the questions:

are distributed estimation algorithms performing better than local ones? And are they performing worse than centralized schemes?

Paraphrasing, we ask if we can compare the performance of distributed estimation algorithms with respect to local strategies, where every agent considers only its dataset and do not share information. And we ask also if we can do the same with respect to centralized strategies, where all the information is collected in an unique place and then processed. In the most general framework, the just-posed questions are extremely complex and difficult to answer. We thus restrict our focus posing a list of assumptions:

- over the multitude of different sensor networks typologies and interpretations (Poor, 2009), we consider collaborative Wireless Sensor Networks (WSNs), i.e. networks in which sensors are randomly distributed over a region of interest and collaborate to achieve a common goal. We assume that agents have limited computational and communication capabilities, that there are no central coordinating units or fusion centers, and that each sensor aims at obtaining a shared knowledge close to the one computable through a centralized strategy. Since we assume that the topology can be dynamic, allowing agents to randomly appear, disappear or move, we will let the nodes to have only a limited topological knowledge: in particular we assume that they only know some statistical properties about the probability density of their physical location. Examples of such networks are WSNs for forest-monitoring where identical sensors are dropped from an helicopter, or a network of sensing robots exploring an unknown but limited region;

- we consider distributed *regression* algorithms -and not classification ones- which data-fitting properties are regulated by cost functions that quadratically weight the estimation errors, both in parametric and nonparametric frameworks;

- we assume that all agents want to obtain global and identical knowledge about the quantity to be estimated. This means that the approximation capability of a given sensor is not focused on its neighborhood, but rather on all the domain where the estimation is performed.

After the analysis of the performances of basic estimator, we focus on an ad-hoc scheme for the *estimation of unknown random fields noisily sampled with unknown delays and in non-uniform locations*. The aim is to provide the sensor network with the capability of distributedly estimate quantities like elevations or intensity of wind speeds despite possible unknown time delays or non-uniformities on the spatial distribution of the measurements. More precisely, we consider the problem of fusing different streams of measurements of a single function observed by various sensors, and subject to unknown temporal shifts. Examples of important applications captured by this framework include estimation of the average force of the wind blowing through a set of wind turbines from noisy samples, or of the time-course of the average concentration of a medicine from plasma samples coming from a set of different patients. In both cases, one needs to adopt a cooperative approach where all the measurements coming from the different sources are exploited to determine the different translations to which signals are subject and to improve function estimation.

We then consider a problem which importance is highlighted by the results shown in the points analyzed before. In fact we will discover, through the way, that the knowledge of the actual number of sensors in the network is an important parameter affecting the performance of the proposed estimators. For this reason, we thus offer a distributed algorithm increasing this knowledge requiring neither leader election steps nor additional topological knowledge like, for example, to be in the neighborhood of a given agent.

It is important to notice that all the techniques we propose in this thesis rely on distributed computation of averages, operation that can be performed through the well-known consensus algorithms (Olfati-Saber and Murray, 2004; Boyd et al., 2006; Olfati-Saber et al., 2007; Fagnani and Zampieri, 2008b; Garin and Schenato, 2011). These algorithms are attractive because of their simplicity, their completely asynchronous and distributed communication schemes, their robustness to nodes and links failures, and their scalability with the network size. In the following we will assume the communication graph to be sufficiently connected in order to allow the computation of consensus algorithms (Cortés, 2008) and that a sufficient number of consensus steps are performed to guarantee convergence to the true average. We notice that, despite their simple structure, they have been proven to be able to compute a wide class of functions (Cortés, 2008), to estimate important physical parameters (Bolognani et al., 2010), or even to synchronize clocks (Bolognani et al., 2009).

Literature review: before stating the novelties introduced in this thesis, we briefly review a list of works related to our framework.

In general, all the research areas involved in this thesis are well estabilished: distributed estimation and distributed computation (Varshney, 1996; Bertsekas and Tsitsiklis, 1997), parametric estimation (Kay, 1993; Anderson and Moore, 1979), nonparametric estimation (Hastie et al., 2001; Schölkopf and Smola, 2001; Wahba, 1990).

In the framework of Bayesian estimation, several authors focused on distributed or decentralized computations. For example, in Kearns and Seung (1995) authors analyze how to combine multiple independent results of learning algorithms performed by identical agents, providing bounds on the number of agents necessary to obtain a desired level of accuracy. In Yamanishi (1997) the author proposes estimation strategies using a hierarchical structure: the sensor nodes perform measurements of the process and preprocess this data, then a supervisor node fuses these local outputs and compute a global estimate. It considers also the expected losses for predicted data, giving upper bounds as functions of the number of samples of each agent. There is also a wide literature on distributed estimation subject to communication constraints: in Predd et al. (2005) authors propose a message-passing scheme for a nonparametric distributed regression algorithm, while in Predd et al. (2006c) they survey the problems related to the distribution of the learning process in wireless sensor networks, analyzing both parametric and nonparametric scenarios. In Predd et al. (2006a) the same authors analyze the existence of decision and fusion rules assuring consistency for a binary classification problem, where the measurements are performed by a set of agents with limited communication capabilities and transmitting information to a central unit. In this framework also some authors propose some asymptotic results on the performance of decision transmission strategies, seeking for optimality in terms of decision error probability for the central unit (Chamberland and Veeravalli, 2004). In Schizas et al. (2008) the authors focus on consensus-based decentralized estimation of deterministic parameter vectors, considering both Maxi-

mum Likelihood (ML) and Best Linear Unbiased Estimator (BLUE) schemes, solved through a set of convex minimization subproblems. Distributed convex optimization has also been used in Schizas and Giannakis (2006), through the parallelization of coordinate descent steps in order to distributedly compute the Linear Minimum Mean Square Error (LMMSE) estimate of an unknown signal. Similar techniques have been used in Mateos et al. (2010), where authors consider three different consensus-based distributed Lasso regression algorithms: the first based on quadratic programming techniques, the second on cyclic coordinate descent steps, and the third on the decomposition of the original cost function into smaller optimization subproblems. Other authors proposed distributed inference schemes based on graphical models, like in Ihler (2005) or in Delouille et al. (2004), where an LMMSE estimator is proposed that exploite a particular implementation of the Gauss-Seidel matrix inversion algorithm.

Parametric modeling of random processes naturally arise in scenarios where it is possible to classify the nature of the random process, and therefore the estimator is searched within a specific class of models such as polynomials or radial basis functions. However, there are problems for which this is difficult and nonparametric estimation has been found to be more suitable and effective. In particular, the nonparametric approaches can be designed to be consistent with a large number of models classes, e.g. Nonlinear AutoRegressive eXogenus (NARX) models (De Nicolao and Ferrari-Trecate, 1999). Within the nonparametric framework, the theory of Reproducing Kernel Hilbert Spaces (RKHSs) (Aronszajn, 1950) have been often used for regression purposes (Rasmussen and Williams, 2006; Schölkopf and Smola, 2001). This theory has been successfully used also in distributed scenarios: for example, Predd et al. (2009) proposes a distributed regularized kernel Least Squares (LS) regression problem based on successive orthogonal projections. Similarly, in Pérez-Cruz and Kulkarni (2010) the authors extend Predd et al. (2009) by proposing modifications reducing the communication burden and synchronization assumptions. In Honeine et al. (2008, 2009) authors propose a reduced order model approach where sensors construct an estimate considering only a subset of the representing functions that would be used in the optimal solution, with a selection method based on the assessment of the potential improvement given to the current solution by adding a new representing function. Other approaches involve message-passing based schemes in graphical models: in Predd et al. (2005) the authors consider a nonparametric distributed regression algorithm that is subject to communication constraints, while Guestrin et al. (2004) considers kernel linear regressors without regularizing terms through the usage of opportune junction and routing trees. nonparametric schemes have been associated also with belief propagation schemes, for example in Çetin et al. (2006), or in Sudderth et al. (2003), where it is used in conjunction with regularizing kernels associated to each particle.

Although the current trend is towards the design of purely distributed algorithms where each agent runs the same algorithm, also hierarchical strategies have been proposed. For example, Yamanishi (1997) offers a distributed Bayesian learning scheme where a supervisor node fuses the results of local outputs. Zheng et al. (2008) proposes an iterative conditional expectation algorithm that distributedly

estimates a deterministic function, while Li et al. (2010) uses a pre-defined cyclic learning schemes based on information routing tables.

An other interesting research field is given by mobile sensor networks, where agents exploit their motion capabilities to perform particular tasks. A first example is Cortés (2009), where the author introduces the so-called Distributed Kriged Kalman Filter, an algorithm used to estimate the distribution of a dynamic Gaussian random field and its gradient. We notice that here sensors estimate their own neighborhood and not to the global scenario. In the same framework, in Choi et al. (2009) the authors develop a distributed learning and cooperative control algorithm where sensors estimate a static field. This field is modeled as a network of radial basis functions that are known in advance by sensors, and this resembles our approach. Nonparametric schemes are applied also in Martínez (2010), where the mobile sensor network distributedly estimates a noisily sampled scalar random field through opportune Nearest-Neighbors interpolation schemes.

Distributed nonparametric techniques have been used also in other frameworks: for example in detection with Nguyen et al. (2005), where the authors consider a decentralized classification framework based on minimization of empirical risks and the concept of marginalized kernels, under communication constraints. In classification, with D'Costa and Sayeed (2003) analytically and numerically comparing three distributed classifiers of objects moving though the sensor field (one based on ML concepts, the others based on data-averagings and different data correlations hypotheses). And also for calibration purposes, for example in Dogandžić and Zhang (2006) through the distributed regression of the realization of a random Markov field. Notice that some schemes considering severe limitations in communications capability have been considered, like for example in Wang et al. (2008) where sensors are allowed to exchange only one bit per time of information.

We briefly recall that sensor networks have been proposed also for fault detection and change detection purposes, both in parametric (Snoussi and Richard, 2006) and in nonparametric frameworks (He et al., 2006; Nasipuri and Tantaratana, 1997).

Statement of contribution: in the first part of this thesis we will characterize some distributed parametric and nonparametric regression algorithms in terms of the tradeoff between estimation performance and communication, computation and memory complexity. In particular we will provide two types of quantitative bounds concerning the estimation performance:

- the first type of bounds, for the parametric scenario, can be computed off-line, i.e. before the measurement processes, but tends to be pessimistic;

- the second type of bounds is derived in the nonparametric scenario but is easily transportable into the parametric one. It need to be computed on-line and collaboratively with the other nodes after the measurement process, thus it adds some extra computational complexity. However it is generally accurate and can be used as a *certificate of quality* attesting if the distributed estimation results are close to the ones that would be obtained using the optimal

centralized strategy.

We show also the practical distributability of the most important regularized regression techniques, the so-called Regularization Networks (RNs) (Poggio and Girosi, 1990; Evgeniou et al., 2000, and references therein). In doing so, we exploit the linear structure of these state-of-the-art algorithms, that is inherited by their quadratic loss functions. Differently, the other most relevant regularization technique, namely the Support Vector Regression based on Vapnik's loss functions (Vapnik, 1995, Chap. 6.1) (Schölkopf and Smola, 2001, Chap. 9), cannot be easily distributed.

We then consider a slight generalization of the previously expressed framework, and consider how to simultaneously perform Time Delay Estimation (TDE) and function regression under Gaussian hypotheses. In the literature, classical TDE techniques work only in a scenario which involves two sensors. Usually, the delay is estimated by maximizing cross-correlation functions or -when proper filtering is applied- generalized cross-correlation functions (Azaria and Hertz, 1984). Other authors use Fast Fourier Transforms (Marple, 1999). In TDE for signals over a discrete domain, additional hypotheses allow efficient interpolation schemes (Boucher and Hassab, 1981; Viola and Walker, 2005). However, classical discrete TDE strategies cannot be usually applied when the sampling period is not constant, and it does not easily generalize to the case of more than two sensors collecting measurements. The algorithm we propose instead can handle non-uniform sampling grids and an arbitrary number of sensors. Moreover, as compared to classical function estimation techniques, developed either in centralized contexts or in distributed ones our approach is also suitable to simultaneously estimate time delays between sensors.

We then propose a generic fully distributed procedure for the estimation of the number of sensors in a network[1] that is based on the generation of random variables and on consensus-based information exchange mechanisms. The advantages with respect to classical schemes are that estimations will be independent on the network structure and on the transmission medium, and that sensors will in general be not required to authenticate, allowing to be insensible to privacy problematics. Our contributions can be summarized in a series of asymptotic analyses and theorems characterizing, from a statistical point of view, the performances of the proposed estimators under general assumptions and communication schemes. We moreover consider the cases where a-priori information on the number of agents in the network is available, showing that Maximum A Posteriori (MAP) estimators may be implemented if some particular conditions are satisfied.

Structure of the work: this thesis is divided as follows. The first part is composed by Chapters 1 and 2, and deals with general distributed regression techniques. More precisely, in Chapter 1) we consider the distributed parametric estimation framework, and offer some answers on the previously posed questions, while in 2 we consider the distributed nonparametric one. The second part of the thesis, composed by Chapter 3, deals with the estimation of the number of agents in a network. We finally

[1] Literature review of this particular problem is given in Chapter 3.1.

draw some concluding remarks and analyze the future works in the conclusions, offered from page 97.

List of acronyms

AIC	Akaike Information Criterion
BIC	Bayesian Information Criterion
BLUE	Best Linear Unbiased Estimator
GP	Gaussian Process
MAP	Maximum A Posteriori
MMSE	Minimum Mean Square Error
ML	Maximum Likelihood
LMMSE	Linear Minimum Mean Square Error
LS	Least Squares
WSN	Wireless Sensor Network
NARX	Nonlinear AutoRegressive eXogenus
NCS	Networked Control System
PEM	Prediction Error Methods
RKHS	Reproducing Kernel Hilbert Space
RN	Regularization Network
TDE	Time Delay Estimation

1 Distributed Parametric Estimation

1.1 Introduction

In this chapter we briefly review some concepts about Bayesian estimation in parametric scenarios. Interested readers can find more details in Kay (1993); Anderson and Moore (1979). We refer to standard textbooks like Feller (1971) for the concepts of probability theory like random vectors, characteristic functions, etc., and to standard textbooks like Nef (1967) for the concepts of linear algebra.

The chapter is divided in two parts: we initially introduce the basic concept of Gaussian random vector (r.v.) among some of its mathematical properties in Section 1.1.1, and then we introduce the Bayesian regression framework that we will use in part of this thesis in Section 1.1.2.

1.1.1 Gaussian random vectors

Let b be an E dimensional real-valued r.v..

Definition 1 (Gaussian random vector)**.** If the characteristic function φ of the r.v. b has the form

$$\varphi(\Theta) = \exp\left(i\Theta^T\mathbf{m} - \frac{1}{2}\Theta^T\Lambda_0\Theta\right) \tag{1.1}$$

with

$$\Theta \in \mathbb{R}^E \qquad \mathbf{m} \in \mathbb{R}^E \qquad \Lambda_0 \in \mathbb{R}^{E\times E} \tag{1.2}$$

then b is said to be a *Gaussian random vector*.

From the properties of the characteristic function it follows that

$$\mathbb{E}[b] = \mathbf{m} \qquad \text{and} \qquad \mathbb{E}\left[(b - \mathbb{E}[b])(b - \mathbb{E}[b])^T\right] = \Lambda_0 \tag{1.3}$$

thus $\mathbf{m}$ is the vector of the mean values and Λ_0 is the covariance matrix. If Λ_0 is invertible, then the probability density function of the r.v. b is well defined and equal to

$$p(\beta) = \frac{1}{\sqrt{(2\pi)^E|\Lambda_0|}}\exp\left(-\frac{1}{2}(\beta - \mathbf{m})^T\Lambda_0^{-1}(\beta - \mathbf{m})\right) \tag{1.4}$$

where $|\Lambda_0|$ indicates the determinant of Λ_0. In the following, to indicate a Gaussian r.v. we will use the standard notation $b \sim \mathcal{N}(\mathbf{m}, \Lambda_0)$.

1.1.1.1 Conditional densities for Gaussian random vectors

Assume

$$\begin{bmatrix} b_1 \\ b_2 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{m}_1 \\ \mathbf{m}_2 \end{bmatrix}, \begin{bmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{bmatrix} \right) \tag{1.5}$$

where vectors and matrices are of consistent dimensions. Then it is possible to show that

$$b_1 \sim \mathcal{N}(\mathbf{m}_1, \Lambda_{11}) \qquad b_2 \sim \mathcal{N}(\mathbf{m}_2, \Lambda_{22}) \tag{1.6}$$

and, more importantly to our purposes, that

$$b_1 | \, b_2 = \beta_2 \sim \mathcal{N}\left(\mathbf{m}_1 + \Lambda_{12}\Lambda_{22}^{-1}(\beta_2 - \mathbf{m}_2), \Lambda_{11} - \Lambda_{12}\Lambda_{22}^{-1}\Lambda_{21}\right) \tag{1.7}$$

where $b_1 | \, b_2 = \beta_2$ indicates the r.v. b_1 conditioned on $b_2 = \beta_2$.

1.1.1.2 Linear models for Gaussian random vectors

In sight of the next derivations, it is important to recall the following basic result. Assume that we want to estimate an unknown r.v. b from a set of noisy measurements $y \in \mathbb{R}^M$ that are linearly related to b through

$$y = Cb + \nu \tag{1.8}$$

with $\nu \in \mathbb{R}^M$ the noise vector s.t. $\nu \sim \mathcal{N}(0, \Sigma_\nu)$ independent on b, and with $C \in \mathbb{R}^{M \times E}$ the known transformation matrix. In this case

$$b | \, y \sim \mathcal{N}\left(\mathbf{m}', \Lambda_0'\right) \tag{1.9}$$

with

$$\mathbf{m}' = \mathbf{m} + \Lambda_0 C^T \left(C\Lambda_0 C^T + \Sigma_\nu\right)^{-1} (y - C\mathbf{m}) \tag{1.10}$$

$$\Lambda_0' = \Lambda_0 - \Lambda_0 C^T \left(C\Lambda_0 C^T + \Sigma_\nu\right)^{-1} C\Lambda_0 \tag{1.11}$$

or, equivalently,

$$\mathbf{m}' = \mathbf{m} + \left(\Lambda_0^{-1} + C^T\Sigma_\nu^{-1}C\right)^{-1} C^T\Sigma_\nu^{-1}(y - C\mathbf{m}) \tag{1.12}$$

$$\Lambda_0' = \left(\Lambda_0^{-1} + C^T\Sigma_\nu^{-1}C\right)^{-1} . \tag{1.13}$$

1.1.2 Bayesian regression

Bayesian regression, and more in general Bayesian inference, is one of the most powerful analysis techniques. In this section we will briefly introduce only the concepts that will be used in the subsequent chapters, and refer to standard textbooks on decision theory like Berger (1985) for the unspecified details.

The Bayesian approach assumes the knowledge of some prior information on the quantity to be estimated b, here assumed to be on the form of a prior density $p(b)$.

Assuming that there exists a statistical relationship between the unknown b and a known r.v. y under the form of the conditional probability density $p(b|y)$, through the well known Bayes rule it is possible to compute the posterior probability density

$$p(b|y) = \frac{p(y|b)\,p(b)}{\int_{\mathbb{R}^E} p(y|b)\,p(b)\,db} \,. \tag{1.14}$$

Even if this is, from a theoretical point of view, the best information we can have on the quantity to be estimated b, in practice this could be even impossible to be computed due to numerical integration difficulties.

It is often the case then to compute a strategy $\widehat{b}(y)$ (in the following abbreviated with $\widehat{b}$ for brevity) that, given in input the known vector y, returns a point-estimate of b. Assume then we are given a loss function

$$L : \mathbb{R}^E \times \mathbb{R}^M \to \mathbb{R} \tag{1.15}$$

associating to each couple $\left(b, \widehat{b}\right)$ a loss $L\left(b, \widehat{b}\right) = L\left(b - \widehat{b}\right)$. This loss has to be intended as a level of disappointment[1] depending on the estimation error $b - \widehat{b}$. With this definition, it is possible to introduce the concept of *risk* (or *Bayes risk*) associated with a particular estimator $\widehat{b}$, defined as

$$\mathcal{R}_{\widehat{b}} := \mathbb{E}\left[L\left(b, \widehat{b}\right)\right] \tag{1.16}$$

where the expectation is on the joint density $p(b, y)$ and with the intuitive meaning of being the average loss in which we will incur using the estimator $\widehat{b}$. Now, given definition (1.16) and given the task of designing a suitable estimator, it is straightforward that the choice is the estimator $\widehat{b}$ with the minimal risk $\mathcal{R}_{\widehat{b}}$.

The structure of the optimal estimator -where optimality has to be meant in terms of minimization of $\mathcal{R}_{\widehat{b}}$- and its statistical properties strongly depend on the loss function L. While from a theoretical point of view the design of L should be application-oriented, from a practical point of view it is often common to choose some well known and predefined L's due their already analyzed mathematical or practical good qualities. Some of the most typical loss functions for the case $E = 1$ are shown in Figure 1.1 and described in its caption.

We recall now a well-known fact about Bayesian analysis: for squared-error loss functions the Bayes risk is minimized by the conditional mean $\mathbb{E}[b \mid y]$, i.e.

$$\mathbb{E}[b \mid y] = \arg\min_{\widehat{b}} \mathcal{R}_{\widehat{b}} = \arg\min_{\widehat{b}} \int_{\mathbb{R}^E} \int_{\mathbb{R}^M} \left(b - \widehat{b}\right)^2 p(b, y)\,db\,dy \,. \tag{1.17}$$

This expresses the fact that conditional means are Minimum Mean Square Error (MMSE) estimators. Conditional means have also some other desirable properties: first of all, as stated in Sections 1.1.1.1 and 1.1.1.2, if b and y are jointly gaussian, then conditional mean can be computed through linear operations. Moreover there is a strong connection between MMSE estimators and classical least squares

[1] The interpretation we use in this thesis is rather restrictive. We send the reader back to (Berger, 1985, Chap. 2) for exhaustive descriptions and interpretations of utility and loss functions.

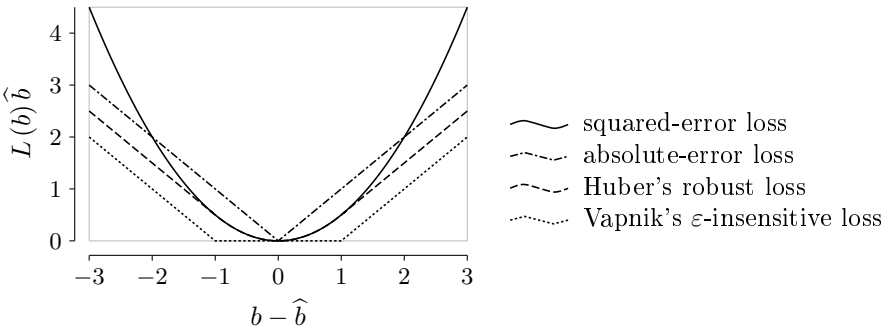


Figure 1.1: Typical loss functions for the case $E = 1$. Properties of estimators based on squared-error losses ($L\left(b-\widehat{b}\right) = \left(b-\widehat{b}\right)^2$) will be described later. Estimators for absolute-error losses ($L\left(b-\widehat{b}\right) = \left|b-\widehat{b}\right|$) correspond to the medians of the posterior densities $p\,(b\,|y\,)$. Vapnik's loss functions (Vapnik, 1995, Chap. 6.1) have the desirable property of returning estimators with compact representations. Huber loss functions (Huber, 1964) combine the sensitivity associated to quadratic loss functions and the robustness to outliers associated to absolute loss functions, and has been proved to be effective in practical cases (e.g. Müller et al. (1997)).

theory (Gelb, 1974, Chap. 4). Due to these desirable qualities, in this thesis we consider only the squared-error loss case, and keep the analyses for other loss functions as future works.

1.2 Case I: Identical Sensors

We consider S distinct sensors each of them taking M scalar noisy measurements on the same input locations. We model this scenario in a parametric framework as

$$y_i = Cb + \nu_i, \qquad i = 1, \dots, S \tag{1.18}$$

where $y_i \in \mathbb{R}^M$ is the measurements vector collected by the i-th sensor, and $b \in \mathbb{R}^E$ is the vector of unknown parameters modeled as a zero-mean Gaussian vector with autocovariance Λ_0, i.e. $b \sim \mathcal{N}(0, \Lambda_0)$. In addition, $\nu_i \in \mathbb{R}^M$ is the noise vector with density $\mathcal{N}\left(0, \sigma^2 I\right)$, independent of b and of ν_j, for $i \neq j$. Finally, $C \in \mathbb{R}^{M \times E}$ is a known matrix identical for all sensors.

1.2.1 Local Bayesian Estimation

Under the assumptions above, the local MMSE estimator of b given y_i, is unbiased and given by

$$\begin{aligned} \widehat{b}_\ell^i &:= \mathbb{E}\left[b \mid y_i\right] = \operatorname{cov}\left(b, y_i\right) \left(\operatorname{var}\left(y_i\right)\right)^{-1} y_i \\ &= \Lambda_0 C^T \left(C \Lambda_0 C^T + \sigma^2 I\right)^{-1} y_i \\ &= \left(\Lambda_0^{-1} + \frac{C^T C}{\sigma^2}\right)^{-1} \frac{C^T y_i}{\sigma^2} \end{aligned} \tag{1.19}$$

while the autocovariance of the local estimation error is

$$\Lambda_\ell^i := \operatorname{var}\left(b - \widehat{b}_\ell^i\right) = \left(\Lambda_0^{-1} + \frac{C^T C}{\sigma^2}\right)^{-1} = \Lambda_\ell \tag{1.20}$$

which is independent of the measurements y_i and sensor index i.

1.2.2 Centralized Bayesian Estimation

If $S \geq 2$ and all measurements $\{y_i\}_{i=1}^S$ are collected by a central unit, the MMSE estimate of b given $\{y_i\}_{i=1}^S$ can be computed as

$$\widehat{b}_{\mathrm{c}} := \operatorname{cov}\left(b, \begin{bmatrix} y_1 \\ \vdots \\ y_S \end{bmatrix}\right) \operatorname{var}\left(\begin{bmatrix} y_1 \\ \vdots \\ y_S \end{bmatrix}\right)^{-1} \begin{bmatrix} y_1 \\ \vdots \\ y_S \end{bmatrix} \tag{1.21}$$

where:

$$\operatorname{var}\left(\begin{bmatrix} y_1 \\ \vdots \\ y_S \end{bmatrix}\right) = \begin{bmatrix} V\left(\sigma^2\right) & \dots & V(0) \\ \vdots & & \vdots \\ V(0) & \dots & V\left(\sigma^2\right) \end{bmatrix} \tag{1.22}$$

where

$$V(\theta) := C \Lambda_0 C^T + \theta I\,. \tag{1.23}$$

Using the matrix inversion lemma and simple algebraic manipulations, (1.21) can be rewritten in two equivalent forms, i.e. as

$$\widehat{b}_{\mathrm{c}} = \Lambda_0 C^T \left(C \Lambda_0 C^T + \frac{\sigma^2}{S} I\right)^{-1} \left(\frac{1}{S} \sum_{i=1}^S y_i\right) \tag{1.24}$$

or as

$$\widehat{b}_{\mathrm{c}} = \left(\frac{1}{S}\Lambda_0^{-1} + \frac{C^T C}{\sigma^2}\right)^{-1} \left(\frac{1}{S}\sum_{i=1}^{S}\frac{C^T y_i}{\sigma^2}\right) . \tag{1.25}$$

Obviously the variance of the estimation error is the same for both the forms, and is given by

$$\Lambda_{\mathrm{c}} := \mathrm{var}\left(\widehat{b}_{\mathrm{c}} - b\right) = \left(\Lambda_0^{-1} + \frac{S}{\sigma^2}\cdot C^T C\right)^{-1} . \tag{1.26}$$

1.2.3 Distributed Bayesian Estimation

Before continuing it is important to highligh the following:

Remark 2. In order to be able to implement the optimal estimation strategies (1.24) or (1.25), all sensors must have perfect knowledge on S, the actual number of agents participating to the consensus process. In fact, in (1.24) S contributes to weight properly the noisiness of the averaged measurements, while in (1.25) it properly weights the contribution of the prior Λ_0.

Remark 3. To compute $\widehat{b}_{\mathrm{c}}$ through (1.24), sensors need to reach an average consensus on their measurements y_i which are M-dimensional vectors, while to compute $\widehat{b}_{\mathrm{c}}$ through (1.25) they need to reach an average consensus on the E-dimensional transformed measurement vectors $C^T y_i/\sigma^2$. In the context of parametric estimation, these vectors are also known as the information vectors associated to the measurements y_i (Anderson and Moore, 1979).

In the rest of the section we will make the natural assumptions

- $E \ll M$;
- S is unknown.

These lead immediately to the following approximated distributed estimation strategy:

$$\widehat{b}_{\mathrm{d}}\left(S_{\mathrm{g}}\right) := \left(\frac{1}{S_{\mathrm{g}}}\Lambda_0^{-1} + \frac{C^T C}{\sigma^2}\right)^{-1} \left(\frac{1}{S}\sum_{i=1}^{S}\frac{C^T y_i}{\sigma^2}\right) \tag{1.27}$$

where S_{g} is an estimate of the number of sensors in the networks. To simplify the notation, in the following we denote $\widehat{b}_{\mathrm{d}}\left(S_{\mathrm{g}}\right)$ as $\widehat{b}_{\mathrm{d}}$ unless differently stated. Simple algebraic manipulations lead to the computation of the corresponding estimation error covariance

$$\begin{aligned} \Lambda_{\mathrm{d}}\left(S_{\mathrm{g}}\right) &:= \mathrm{var}\left(\widehat{b}_{\mathrm{d}} - b\right) \\ &= \left(\frac{1}{S_{\mathrm{g}}}\Lambda_0^{-1} + \frac{C^T C}{\sigma^2}\right)^{-1} \left(\frac{1}{S_{\mathrm{g}}^2}\Lambda_0^{-1} + \frac{1}{S}\frac{C^T C}{\sigma^2}\right) \left(\frac{1}{S_{\mathrm{g}}}\Lambda_0^{-1} + \frac{C^T C}{\sigma^2}\right)^{-1} . \end{aligned} \tag{1.28}$$

Notice that if $S_{\mathrm{g}} = 1$ then

$$\widehat{b}_{\mathrm{d}}\left(1\right) = \frac{1}{S}\sum_{i=1}^{S}\widehat{b}_{\ell}^{i} \tag{1.29}$$

i.e. $\widehat{b}_{\mathrm{d}}(1)$ is equal to the average the local estimators. If $S_{\mathrm{g}} = +\infty$ then

$$\widehat{b}_{\mathrm{d}}(+\infty) = \left(S \cdot C^T C\right)^{-1} \left(\sum_{i=1}^{S} C^T y_i\right) \tag{1.30}$$

i.e. $\widehat{b}_{\mathrm{d}}(+\infty)$ is equal to the least squares solution, which discards the prior information on b. Finally, if $S_{\mathrm{g}} = S$ then $\widehat{b}_{\mathrm{d}}(S)$ is equal to the centralized solution, i.e. $\widehat{b}_{\mathrm{d}}(S) = \widehat{b}_{\mathrm{c}}$. Notice that the same results and the same expression for Λ_{d} would have been obtained also considering the case $E > M$. In this case, the expression of the distributed estimator would be

$$\widehat{b}_{\mathrm{d}}(S_{\mathrm{g}}) := \Lambda_0 C^T \left(C\Lambda_0 C^T + \frac{\sigma^2}{S_{\mathrm{g}}} I\right)^{-1} \left(\frac{1}{S}\sum_{i=1}^{S} y_i\right) . \tag{1.31}$$

1.2.4 Characterization of the Distributed Algorithm

In the following we concern in determining conditions on the parameter S_{g} that guarantee $\Lambda_{\mathrm{d}}(S_{\mathrm{g}}) \leq \Lambda_\ell$, i.e. when a distributed strategy that shares information among nodes is better then the one obtained by using only local information, and in determining the accuracy of the distributed solution as compared to the centralized solution as a function of S_{g}: it is important both to understand when the distributed strategy is beneficial with respect to the local one in order to justify the strain of communication, and also when the approximation represented by considering S_{g} instead of S does not introduce significant performances losses. These two scenarios are addressed separately.

1.2.4.1 Distributed versus Local Estimation

Based on the direct comparison of the distributed estimation error covariance Λ_{d} and the local error covariance Λ_ℓ it is possible to derive sufficient conditions which hold for every prior Λ_0, number of measurements M, number of parameters E, measurement noise variance σ^2, and matrix C:

Theorem 4. If

$$S_{\mathrm{g}} \in [1, 2S-1] \tag{1.32}$$

then the variance of the estimation error of the distributed estimator $\widehat{b}_{\mathrm{d}}(S_{\mathrm{g}})$ is smaller than the one of the local estimators $\widehat{b}_\ell^i$, for every prior Λ_0, number of parameters E, measurement noise variance σ^2, matrix C and sensor i.

Proof. The objective is to find sufficient conditions in terms of the systems parameters $\Lambda_0, S_{\mathrm{g}}, S, C, \sigma$ such that

$$\Lambda_{\mathrm{d}} = \mathrm{var}\left(b - \widehat{b}_{\mathrm{d}}(S_{\mathrm{g}})\right) \leq \mathrm{var}\left(b - \widehat{b}_\ell\right) = \Lambda_\ell \tag{1.33}$$

Recalling the definition $V(\theta) := C\Lambda_0 C^T + \theta I$, it is immediate to verify through the matrix inversion lemma and the equivalence between expressions (1.25) and (1.24)

that

$$\widehat{b}_{\mathrm{d}} = \Lambda_0 C^T V \left(\frac{\sigma^2}{S_{\mathrm{g}}}\right)^{-1} \left(Cb + \frac{1}{S}\sum_{i=1}^{S} \nu_i\right) \tag{1.34}$$

therefore the variance of distributed estimator is given by

$$\Lambda_{\mathrm{d}} = \Lambda_0 - 2\Lambda_0 C^T V \left(\frac{\sigma^2}{S_{\mathrm{g}}}\right)^{-1} C\Lambda_0 + \Lambda_0 C^T V \left(\frac{\sigma^2}{S_{\mathrm{g}}}\right)^{-1} V\left(\frac{\sigma^2}{S}\right) V \left(\frac{\sigma^2}{S_{\mathrm{g}}}\right)^{-1} C\Lambda_0 \tag{1.35}$$

Similarly, for the local estimator we get

$$\Lambda_\ell = \Lambda_0 - \Lambda_0 C^T V \left(\sigma^2\right)^{-1} C\Lambda_0 \tag{1.36}$$

By substituting the previous two equations into (1.33) and by pre and post-multipling by Λ_0^{-1}, we get

$$-2V\left(\frac{\sigma^2}{S_{\mathrm{g}}}\right)^{-1} + V\left(\frac{\sigma^2}{S_{\mathrm{g}}}\right)^{-1} V\left(\frac{\sigma^2}{S}\right) V\left(\frac{\sigma^2}{S_{\mathrm{g}}}\right)^{-1} \leq -2V\left(\sigma^2\right)^{-1} \tag{1.37}$$

which guarantees $\Lambda_{\mathrm{d}} \leq \Lambda_\ell$. Considering now the orthogonal matrix U that diagonalizes $C\Lambda_0 C^T$, i.e.

$$C\Lambda_0 C^T = UDU^T \tag{1.38}$$

s.t. $UU^T = I$, where $D := \mathrm{diag}\left(d_1, \ldots, d_S\right)$, we have that $V(\theta) = U\left(D + \theta I\right)U^T$. Therefore (1.37) can be written as

$$-2U\left(D + \frac{\sigma^2}{S_{\mathrm{g}}} I\right)^{-1} U^T + U\left(D + \frac{\sigma^2}{S_{\mathrm{g}}} I\right)^{-2} \left(D + \frac{\sigma^2}{S} I\right) U^T \leq -U\left(D + \sigma^2 I\right)^{-1} U^T \tag{1.39}$$

where we also used the fact that diagonal matrices commute. Since for invertible matrices U we have that $A \leq 0 \Leftrightarrow UAU^{-1} \leq 0$, so (1.39) is still a sufficient condition for $\Lambda_{\mathrm{d}} \leq \Lambda_\ell$ if we remove all the U's. Now all the remaining matrices are diagonal, so the matricial inequality (1.39) is satisfied if and only if the inequalities are valid component-wise for the diagonal elements. Therefore, 1.33 is equivalent to:

$$\frac{-2}{d_m + \frac{\sigma^2}{S_{\mathrm{g}}}} + \frac{d_m + \frac{\sigma^2}{S}}{\left(d_m + \frac{\sigma^2}{S_{\mathrm{g}}}\right)^2} \leq \frac{-1}{d_m + \sigma^2} \qquad m = 1, \ldots, M \tag{1.40}$$

that can be rewritten as

$$p_m\left(S_{\mathrm{g}}\right) := \left(\sigma^2 + (1 - S)\, d_m\right) S_{\mathrm{g}}^2 - 2\sigma^2 S S_{\mathrm{g}} - \sigma^2 S \leq 0 \tag{1.41}$$

for all m's. Using the following shorthands

$$\dot{p}_m := \frac{\partial p_m}{\partial S_{\mathrm{g}}} \qquad \ddot{p}_m = \frac{\partial^2 p_m}{\partial S_{\mathrm{g}}^2} \tag{1.42}$$

for all m's and d_m's we have that $p_m(0) = \sigma^2 S > 0$ and $p_m(1) = (1 - S)\left(d_m + \sigma^2\right) < (1 - S)\,\sigma^2 < 0$ since we are assuming there are at least two sensors. Moreover we

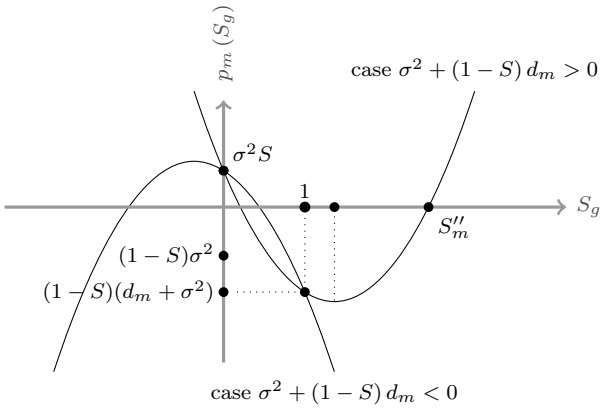


Figure 1.2: Example of possible parabolas $p_m(S_\mathrm{g})$.

also have $\dot{p}_m(0) = -2\sigma^2 S < 0$ and $\dot{p}_m(1) = p_m(1) < 0$. This implies that each $p_m(\cdot)$ has exactly one root in $(0,1)$, referred as S'_m, while the other root, referred as S''_m, can be before 0 or after 1 depending on the sign of $\sigma^2 + (1-S)\,d_m$, as depicted in Figure 1.2.

Now consider a fixed m. Condition (1.41) is assured for $S_\mathrm{g} \in [1, S_m)$, where:

$$S_m := \begin{cases} +\infty & \text{if } S''_m < 0 \\ S''_m & \text{otherwise.} \end{cases} \tag{1.43}$$

Note that this condition still depends on m (i.e. depends on $C\Lambda_0 C^T$). Consider then the parabola with the smallest S_m, say the $\widehat{m}$-th. If p_min is its point of minimum, then $2p_\mathrm{min} - 1 < S_m$ for every m, so if $S_\mathrm{g} \in [1, 2p_\mathrm{min} - 1]$ then condition (1.37) is again satisfied. Now, since $(1-S)d_{\widehat{m}} < 0$ we have:

$$p_\mathrm{min} = \frac{\sigma^2 S}{\sigma^2 + (1-S)d_m} > \frac{\sigma^2 S}{\sigma^2} = S \tag{1.44}$$

and thus $[1, 2S-1] \subset [1, 2p_\mathrm{min} - 1]$. Now we can conclude that if $S_\mathrm{g} \in [1, 2S-1]$ then inequality (1.37) is satisfied, and this proves the theorem. □

The sufficient condition of theorem 4 assures that there exists a large set potential guesses of number of sensors S_g for which the distributed estimator $\widehat{b}_\mathrm{d}$ is performing better than the local one b_ℓ. In particular, this theorem confirms the intuition that the average of the local estimators, i.e. $S_\mathrm{g} = 1$, always produces a better estimate. Moreover, if only rough estimate of S is available, it can be safely used to improve performance. The second sufficient condition (b) implies that the distributed estimator is better than the local for all $S_\mathrm{g} \in [1, +\infty)$. In particular, it confirms the intuition that if the prior information about b is sufficiently small, i.e. Λ_0 is large, and if C is full rank, then the influence of S_g is small on the overall estimator performance.

Assuming now the knowledge of $C\Lambda_0 C^T$ (or equivalently on its eigenvalues d_m), it is possible to enlarge bound (1.32) and find that there could be networks (i.e. S

and σ^2) where, no matter how the guess S_{g} is chosen, distributed estimation leads to a smaller error variance than the local one:

Theorem 5. If $d_{\min}$ is the smallest eigenvalue of $C\Lambda_0 C^T$ and if

$$d_{\min} > \frac{\sigma^2}{S-1} \tag{1.45}$$

then the variance of the estimation error of the distributed estimator $\widehat{b}_{\mathrm{d}}(S_{\mathrm{g}})$ is smaller than the one of the local estimators $\widehat{b}_{\ell}^{i}$, for every sensor i and guess $S_{\mathrm{g}} \in [1, +\infty)$.

Proof. Condition (1.45) assures parabolas $p_m(S_{\mathrm{g}})$ to be all concave, thus $p_{\min} = +\infty$, and this is sufficient for the thesis. □

In this case, the distributed estimator behave better than the local one *also* assuming $S_{\mathrm{g}} = +\infty$, that is equivalent to assume that the averaged measurements have no measurements error. Note that networks with high S or low σ^2 have higher probability to satisfy condition (1.45). The statistical requirement of theorem 5 is that the smallest eigenvalue of $C\Lambda_0 C^T$ has to dominate the resulting noise of the *averaged* measurements.

If S and σ^2 are s.t. theorem 5 is not satisfied, then we can state (as an intermediate consequence of the proof of theorem 4) the following:

Corollary 6. Define:

$$\widehat{d}(S) := \min_{m \in \{1,\dots,M\}} \left\{ d_m \text{ s.t. } \sigma^2 + (1-S)\, d_m > 0 \right\} \tag{1.46}$$

and:

$$\overline{S} := \frac{\sigma^2 S + \sqrt{\sigma^2 S (S-1)\left(\sigma^2 + \widehat{d}(S)\right)}}{\sigma^2 + (1-S)\,\widehat{d}(S)} \,. \tag{1.47}$$

If

$$S_{\mathrm{g}} \in \left[1, 2\overline{S} - 1\right] \tag{1.48}$$

then the variance of the estimation error of the distributed estimator $\widehat{b}_{\mathrm{d}}(S_{\mathrm{g}})$ is smaller than the one of the local estimators $\widehat{b}_{\ell}^{i}$, for every prior Λ_0, number of parameters E, measurement noise variance σ^2, matrix C and sensor i.

Remark 7. Although the conditions in the Theorem 4 are only sufficient, they are nonetheless tight, in the sense that there are scenarios for which if they are not satisfied than $\Lambda_{\mathrm{d}} > \Lambda_{\ell}$. This is in fact the case for the particular scalar system $E = 1$, $M = 1$, $\Lambda_0 = 1$, $C = 1$ and $S = 100$. Exploiting (1.20) and (1.28) it is immediate to check that if $\sigma^2 > 39600$ then $\mathrm{var}\left(b - b_d(2S)\right) > \mathrm{var}\left(b - \widehat{b}_{\ell}\right)$, and thus show that the bound of Theorem 4 can be tight.

In Figure 1.3 we analyze the dependance of the performances of the distributed estimator on the measurement noise level σ^2 and the guess S_{g} for the scalar system introduced in remark 7. In Figure 1.3 we plot $\mathrm{var}\left(b - \widehat{b}_{\mathrm{d}}(S_{\mathrm{g}})\right)$ for different values of σ^2. Noticing that for the considered cases $\mathrm{var}\left(b - \widehat{b}_{\ell}\right) = \frac{\sigma^2}{\sigma^2+1} \approx 1$, we can observe that the importance of a good choice for S_{g} increases with the noisiness level.

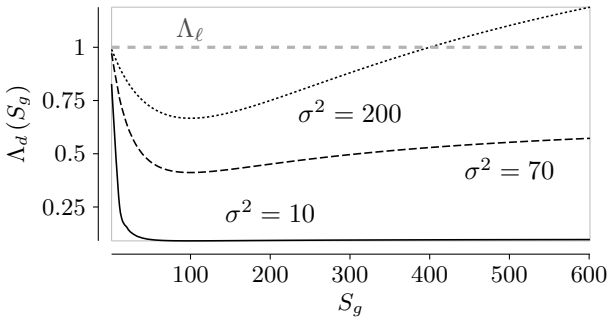


Figure 1.3: Dependency on S_g of the estimation error variance Λ_d of the distributed estimator $\widehat{b}_\mathrm{d}(S_\mathrm{g})$, respectively defined in (1.28) and (1.27), for the particular case $E = 1$, $M = 1$, $\Lambda_0 = 1$, $C = 1$ and $S = 100$, and for different values of σ^2. The dashed gray line approximatively indicates the estimation error variance for the local estimators $\widehat{b}_\ell$.

1.2.4.2 Distributed versus Centralized Estimation

Although we always have $\Lambda_\mathrm{c} \leq \Lambda_\mathrm{d}$, it is relevant to study the influence of the parameter S_g in terms of accuracy between the centralized estimator $\widehat{b}_\mathrm{c}$ and the decentralized estimator $\widehat{b}_\mathrm{d}$. If prior bounds about the unknown parameter S are available, i.e. $S \in [S_{\max}, S_{\min}]$, then the following theorem provides a direct bound on the relative distance of the estimators [2]:

Theorem 8. Under the assumption that $S \in [S_{\min}, S_{\max}]$ then

$$\frac{\left\|\widehat{b}_\mathrm{d} - \widehat{b}_\mathrm{c}\right\|_2}{\left\|\widehat{b}_\mathrm{d}\right\|_2} \leq \frac{S_{\max}}{S_{\min}} - 1 \tag{1.49}$$

for all $S_\mathrm{g} \in [S_{\min}, S_{\max}]$.

Proof. Rewriting (1.25) as

$$\left(\frac{1}{S}\Lambda_0^{-1} + \frac{C^T C}{\sigma^2}\right)\widehat{b}_\mathrm{c} = \frac{1}{S}\sum_{i=1}^{S}\frac{C^T y_i}{\sigma^2} \tag{1.50}$$

and (1.27) as

$$\left(\frac{1}{S}\Lambda_0^{-1} + \frac{C^T C}{\sigma^2}\right)\widehat{b}_\mathrm{d} + \left(\frac{1}{S_\mathrm{g}} - \frac{1}{S}\right)\Lambda_0^{-1}\widehat{b}_\mathrm{d} = \left(\frac{1}{S}\sum_{i=1}^{S}\frac{C^T y_i}{\sigma^2}\right) \tag{1.51}$$

and then subtracting member to member the previous two equations, we obtain

$$\left(\frac{1}{S}\Lambda_0^{-1} + \frac{C^T C}{\sigma^2}\right)\widehat{b}_\mathrm{c} - \widehat{b}_\mathrm{d} = \left(\frac{1}{S_\mathrm{g}} - \frac{1}{S}\right)\Lambda_0^{-1}\widehat{b}_\mathrm{d} \tag{1.52}$$

[2] The following can be considered a connection between the estimation error variance and the square-norm of the error: for a generic vector $\widetilde{b} \in \mathbb{R}^E$ we have

$$\mathrm{tr}\left(\mathrm{var}\left(\widetilde{b}\right)\right) = \mathrm{tr}\left(\mathbb{E}\left[\widetilde{b}\widetilde{b}^T\right]\right) = \mathbb{E}\left[\mathrm{tr}\left(\widetilde{b}\widetilde{b}^T\right)\right] = \mathbb{E}\left[\left\|\widetilde{b}\right\|_2^2\right] .$$

that implies

$$\frac{\left\|\widehat{b}_{\mathrm{c}}-\widehat{b}_{\mathrm{d}}\right\|_2}{\left\|\widehat{b}_{\mathrm{d}}\right\|_2} \leq \left(\frac{1}{S_{\mathrm{g}}}-\frac{1}{S}\right)\left\|\left(\frac{1}{S}\Lambda_0^{-1}+\frac{C^TC}{\sigma^2}\right)^{-1}\Lambda_0^{-1}\right\|_2 . \tag{1.53}$$

Now, since

$$\left(\frac{1}{S_{\mathrm{g}}}-\frac{1}{S}\right) \leq \left(\frac{1}{S_{\min}}-\frac{1}{S_{\max}}\right) \tag{1.54}$$

$$\frac{1}{S}\Lambda_0^{-1}+\frac{C^TC}{\sigma^2} \geq \frac{1}{S_{\max}}\Lambda_0^{-1} \tag{1.55}$$

we have

$$\frac{\left\|\widehat{b}_{\mathrm{c}}-\widehat{b}_{\mathrm{d}}\right\|_2}{\left\|\widehat{b}_{\mathrm{d}}\right\|_2} \leq \left(\frac{1}{S_{\min}}-\frac{1}{S_{\max}}\right)\left\|\left(\frac{1}{S_{\max}}\Lambda_0^{-1}\right)^{-1}\Lambda_0^{-1}\right\|_2 \tag{1.56}$$

and thus (1.49). □

Although the bound provided in the theorem could be improved if additional knowledge about Λ_0 and C is available, it nonetheless suggests that the performance is not strongly dependent on the parameter S_{g}, therefore any sensible choice for this parameter, such as $S_{\mathrm{g}} = (S_{\max}+S_{\min})/2$, is likely to provide a performance close to the centralized solution. This intuition is confirmed in Figure 1.4, where we show the dependency of the relative error $\frac{\left\|\widehat{b}_{\mathrm{d}}-\widehat{b}_{\mathrm{c}}\right\|_2}{\left\|\widehat{b}_{\mathrm{d}}\right\|_2}$ (in this scalar case equal to $\frac{\left|\widehat{b}_{\mathrm{d}}-\widehat{b}_{\mathrm{c}}\right|}{\left|\widehat{b}_{\mathrm{d}}\right|}$) for the system considered in remark 7 and for various strategies, as written in the caption.

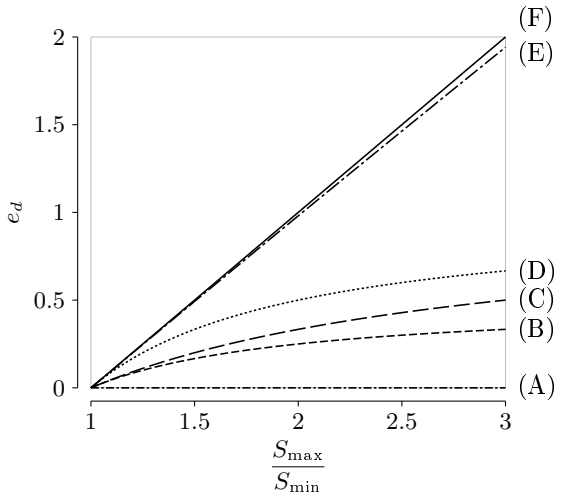


Figure 1.4: Dependency of the relative error $e_d := \frac{\|b_d-b_c\|_2}{\|b_d\|_2}$ on $S_{\max}/S_{\min}$ and σ^2 for various choices of S_g, for the scenario $E=1$, $M=1$, $\Lambda_0=1$, $C=1$. (A): $S_g=S$. (B): $S=S_{\min}$, $S_g=S_{\max}$, $\sigma^2=10^2$. (C): $S=S_{\max}$, $S_g=S_{\min}$, $\sigma^2=10^2$. (D): $S=S_{\min}$, $S_g=S_{\max}$, $\sigma^2=+\infty$. (E): $S=S_{\max}$, $S_g=S_{\min}$, $\sigma^2=10^4$. (F): bound (1.49).

1.3 Case II: Partially Different Sensors

As before, we consider S distinct sensors each of them taking M scalar noisy measurements on the same input locations, and let the measurement model be

$$y_i = Cb + \nu_i, \qquad i = 1, \dots, S \tag{1.57}$$

where S is the number of sensors, $y_i \in \mathbb{R}^M$ is the measurements vector collected by the i-th sensor, $b \in \mathbb{R}^E$ is the vector of unknown parameters modeled as a zero-mean Gaussian vector with autocovariance Λ_0, i.e. $b \sim \mathcal{N}(0, \Lambda_0)$. In addition, $\nu_i \in \mathbb{R}^M$ is the noise vector with density $\mathcal{N}\left(0, \sigma_i^2 I\right)$, independent of b and of ν_j, for $i \neq j$. Finally, $C \in \mathbb{R}^{M \times E}$ is a known matrix, equal for all sensors.

1.3.1 Local Bayesian Estimation

Under the assumptions above, the local MMSE estimator of b given y_i, is unbiased and given by

$$\begin{aligned} \widehat{b}_\ell^i &:= \mathbb{E}\left[b \mid y_i\right] = \operatorname{cov}(b, y_i) \operatorname{var}(y_i)^{-1} y_i \\ &= \Lambda_0 C^T \left(C\Lambda_0 C^T + \sigma_i^2 I\right)^{-1} y_i \\ &= \left(\Lambda_0^{-1} + \frac{C^T C}{\sigma_i^2}\right)^{-1} \frac{C^T y_i}{\sigma_i^2} \,. \end{aligned} \tag{1.58}$$

while the autocovariance of the local estimation error is

$$\Lambda_\ell^i := \operatorname{var}\left(b - \widehat{b}_\ell^i\right) = \left(\Lambda_0^{-1} + \frac{C^T C}{\sigma_i^2}\right)^{-1} \tag{1.59}$$

which is again independent of the measurements y_i but now depends on the noisiness of sensor i.

1.3.2 Centralized Bayesian Estimation

If $S \geq 2$ and all measurements $\{y_i\}_{i=1}^S$ are collected by a central unit, the MMSE estimate of the parameter vector b can be computed as we did in Section 1.2.2 and can be written as

$$\widehat{b}_{\mathrm{c}} = \Lambda_0 C^T \left(C\Lambda_0 C^T + \left(\sum_{i=1}^S \frac{1}{\sigma_i^2}\right)^{-1} I\right)^{-1} \left(\frac{\frac{1}{S}\sum_{i=1}^S \frac{y_i}{\sigma_i^2}}{\frac{1}{S}\sum_{i=1}^S \frac{1}{\sigma_i^2}}\right) \tag{1.60}$$

or as

$$\widehat{b}_{\mathrm{c}} = \left(\frac{1}{S}\Lambda_0^{-1} + \left(\frac{1}{S}\sum_{i=1}^S \frac{1}{\sigma_i^2}\right) \cdot C^T C\right)^{-1} \left(\frac{1}{S}\sum_{i=1}^S \frac{C^T y_i}{\sigma^2}\right) \,. \tag{1.61}$$

Using

$$\alpha := \sum_{i=1}^S \frac{1}{\sigma_i^2} \tag{1.62}$$

as a shorthand for the sum of the precisions $1/\sigma_i^2$, the previous expressions can be written as

$$\widehat{b}_{\mathrm{c}} = \Lambda_0 C^T \left(C\Lambda_0 C^T + \frac{1}{\alpha} I \right)^{-1} \left(\frac{\frac{1}{S}\sum_{i=1}^{S} \frac{y_i}{\sigma_i^2}}{\frac{1}{S}\sum_{i=1}^{S} \frac{1}{\sigma_i^2}} \right) \tag{1.63}$$

and

$$\widehat{b}_{\mathrm{c}} = \left(\frac{1}{S}\Lambda_0^{-1} + \frac{\alpha}{S} \cdot C^T C \right)^{-1} \left(\frac{1}{S} \sum_{i=1}^{S} \frac{C^T y_i}{\sigma^2} \right) . \tag{1.64}$$

Again the variance of the estimation error is the same for both the forms, and is given by

$$\Lambda_{\mathrm{c}} := \mathrm{var}\left(\widehat{b}_{\mathrm{c}} - b \right) = \left(\Lambda_0^{-1} + \left(\sum_{i=1}^{S} \frac{1}{\sigma_i^2} \right) \cdot C^T C \right)^{-1} . \tag{1.65}$$

Connection between the sum of precisions α and the number of sensors S

Let h be the harmonic mean of the measurements noises variances, i.e.:

$$h := h\left(\sigma_1^2, \ldots, \sigma_S^2\right) := \frac{S}{\sum_{i=1}^{S} \frac{1}{\sigma_i^2}} . \tag{1.66}$$

It is evident that average consensus on the quantities $\dfrac{1}{\sigma_i^2}$ corresponds to a distributed estimation of h^{-1}. Thus, exploiting the relation

$$\alpha := \sum_{i=1}^{S} \frac{1}{\sigma_i^2} = \frac{S}{h} , \tag{1.67}$$

after a pre-distributed estimation step for h^{-1}, the knowledge of the number of sensors S is equivalent to the knowledge of the sum of precisions α.

1.3.3 Distributed Bayesian Estimation

Remarks 2 and 3 of Section 1.2.3 should now be modified in the following way:

Remark 9. In order to be able to implement the optimal estimation strategy (1.63), all sensors must have perfect knowledge on α, while to implement strategy (1.64) sensors must have perfect knowledge on S.

Remark 10. To compute $\widehat{b}_{\mathrm{c}}$ through (1.63), sensors need to have a guess run in parallel two average consensi: one on their normalized measurements y_i/σ_i^2, which are M-dimensional vectors, and one on their precisions $1/\sigma_i^2$, that are scalar quantities. Once computed these two quantities, they have to compute their ratio.

To compute $\widehat{b}_{\mathrm{c}}$ through strategy (1.25), sensors need to reach an average consensus on the E-dimensional transformed measurement vectors $C^T y_i/\sigma_i^2$. Moreover they run

in parallel an average consensus to compute the average precision $\frac{1}{S}\sum_i 1/\sigma_i^2$, that is needed to correctly estimate the ratio α/S.

Notice that the two possible strategies have different drawbacks: assuming again $E \ll M$, with (1.24) sensors will exchange more data, while with (1.25) sensors will have to compute the operator $\left(\frac{1}{S}\Lambda_0^{-1} + \frac{\alpha}{S}\cdot C^T C\right)^{-1}$ only after the consensus processes.

In the rest of the section we will make again the natural assumptions

- $E \ll M$;
- S and α are unknown.

We will consider both the strategies

$$\widehat{b}_{\mathrm{d}}\left(\alpha_{\mathrm{g}}\right) := \Lambda_0 C^T \left(C\Lambda_0 C^T + \frac{1}{\alpha_{\mathrm{g}}} I\right)^{-1} \left(\frac{\frac{1}{S}\sum_{i=1}^{S}\frac{y_i}{\sigma_i^2}}{\frac{1}{S}\sum_{i=1}^{S}\frac{1}{\sigma_i^2}}\right) \tag{1.68}$$

and

$$\widehat{b}_{\mathrm{d}}\left(S_{\mathrm{g}}\right) := \left(\frac{1}{S_{\mathrm{g}}}\Lambda_0^{-1} + \left(\frac{1}{S}\sum_{i=1}^{S}\frac{1}{\sigma_i^2}\right)\cdot C^T C\right)^{-1} \left(\frac{1}{S}\sum_{i=1}^{S}\frac{C^T y_i}{\sigma^2}\right) . \tag{1.69}$$

where α_{g} and S_{g} are estimates respectively of α and S. With some algebraic manipulations it is possible to explicitly derive their –obviously identical– estimation error covariance, that is given by

$$\begin{aligned} \Lambda_{\mathrm{d}} &:= \mathrm{var}\left(\widehat{b}_{\mathrm{d}} - b\right) \\ &= \left(\frac{1}{S_{\mathrm{g}}}\Lambda_0^{-1} + \frac{\alpha}{S}\cdot C^T C\right)^{-1} \left(\frac{1}{S_{\mathrm{g}}^2}\Lambda_0^{-1} + \frac{\alpha}{S^2}\cdot C^T C\right) \left(\frac{1}{S_{\mathrm{g}}}\Lambda_0^{-1} + \frac{\alpha}{S}\cdot C^T C\right)^{-1} . \end{aligned} \tag{1.70}$$

1.3.4 Characterization of the Distributed Algorithm

As we did in Section 1.2.4, we now derive conditions that guarantee that the process of sharing and combining the information improves the estimation of b with respect to the local estimation strategy. In other words, we obtain conditions relative to the level of uncertainty on the values of α and S that ensure that the distributed strategy returns a smaller autocovariance (in a matrix sense) of the estimation error than that obtainable by the local one.

1.3.4.1 Distributed versus Local Estimation

We start deriving conditions referred to the level of uncertainity with respect to α, and then transport them into conditions on the uncertainity on S.

Uncertainity with respect to α:

Theorem 11. If

$$\alpha_{\mathrm{g}} \in \left[\alpha - \sqrt{\alpha^2 - \frac{\alpha}{\sigma_i^2}},\ \alpha + \sqrt{\alpha^2 - \frac{\alpha}{\sigma_i^2}}\right] \tag{1.71}$$

then the variance of the estimation error of the distributed estimator $\widehat{b}_{\mathrm{d}}(\alpha_{\mathrm{g}})$ is smaller than the one of the local estimator $\widehat{b}_{\ell}^{i}$, for every prior Λ_0, number of parameters E, sum of precisions α and matrix C.

Proof. Repeating the initial steps of the proof of Theorem 4, we can obtain that a sufficient condition assuring

$$\Lambda_{\mathrm{d}} = \mathrm{var}\left(b - \widehat{b}_{\mathrm{d}}(\alpha_{\mathrm{g}})\right) \leq \mathrm{var}\left(b - \widehat{b}_{\ell}\right) = \Lambda_{\ell} \tag{1.72}$$

is given by

$$\frac{-2}{d_m + \frac{1}{\alpha_{\mathrm{g}}}} + \frac{d_m + \frac{1}{\alpha}}{\left(d_m + \frac{1}{\alpha_{\mathrm{g}}}\right)^2} \leq \frac{-1}{d_m + \sigma_i^2} \qquad m = 1, \ldots, M. \tag{1.73}$$

where d_m is, as previously indicated, an eigenvalue of $C\Lambda_0 C^T$, thus it is $d_m \geq 0$ for all m's since Λ_0 is at least semi-positive definite. Now condition (1.73) can be rewritten as

$$p_{i,m}(\alpha_{\mathrm{g}}) := \left(\sigma_i^2 + \left(1 - \alpha\sigma_i^2\right) d_m\right)\alpha_{\mathrm{g}}^2 - \left(2\alpha\sigma_i^2\right)\alpha_{\mathrm{g}} + \alpha \leq 0 \tag{1.74}$$

for all m. Notice that

$$\alpha\sigma_i^2 = \sum_{j=1}^{S} \frac{\sigma_i^2}{\sigma_j^2} = 1 + \sum_{j \neq i} \frac{\sigma_i^2}{\sigma_j^2} \geq 1 \tag{1.75}$$

thus $\left(1 - \alpha\sigma_i^2\right) d_m \leq 0$, thus parabolas $p_{i,m}(\alpha_{\mathrm{g}})$ can be convex, concave or degenerated depending on σ_i^2. Their roots are in general

$$\begin{aligned} r_{\pm}(i,m) &:= \frac{\alpha\sigma_i^2 \pm \sqrt{\left(\alpha\sigma_i^2 - 1\right)\left(\alpha d_m + \alpha\sigma_i^2\right)}}{\sigma_i^2 + \left(1 - \alpha\sigma_i^2\right) d_m} \\ &= \frac{\alpha}{\alpha\sigma_i^2 \mp \sqrt{\left(\alpha\sigma_i^2 - 1\right)\left(\alpha d_m + \alpha\sigma_i^2\right)}}. \end{aligned} \tag{1.76}$$

Recalling that we have to find the α_{g}'s that assure condition (1.74) independently of i and m, we analyze separately the three cases.

Convex parabolas (i.e. $\sigma_i^2 + \left(1 - \alpha\sigma_i^2\right) d_m > 0$): in this case $r_-(i,m) < r_+(i,m)$ for all i and m. Since

$$r_-(i,m) < \frac{\alpha}{\alpha\sigma_i^2 + \sqrt{\left(\alpha\sigma_i^2 - 1\right)\alpha\sigma_i^2}} =: b_-(i) \tag{1.77}$$

$$r_+(i,m) > \frac{\alpha\sigma_i^2 + \sqrt{\left(\alpha\sigma_i^2 - 1\right)\alpha\sigma_i^2}}{\sigma_i^2} =: b_+(i) \tag{1.78}$$

and since it can be shown by rationalization of $b_-(i)$ that $b_-(i) < b_+(i)$ for all $\sigma_i^2 \geq 0$, we are sure that for any convex parabola $p_{i,m}(\alpha_g)$

$$\alpha_g \in [b_-(i),\ b_+(i)] \quad \Rightarrow \quad p_{i,m}(\alpha_g) \leq 0\ . \tag{1.79}$$

Concave parabolas (i.e. $\sigma_i^2 + \left(1 - \alpha\sigma_i^2\right) d_m < 0$): we check that implication (1.79) is still valid. For doing so it is sufficient to check if $p_{i,m}(b_-(i)) \leq 0$, $p_{i,m}(b_+(i)) \leq 0$ and that

$$\operatorname{sign}\left(\left.\frac{\partial p_{i,m}(\alpha_g)}{\partial \alpha_g}\right|_{b_-(i)}\right) = \operatorname{sign}\left(\left.\frac{\partial p_{i,m}(\alpha_g)}{\partial \alpha_g}\right|_{b_+(i)}\right) \tag{1.80}$$

and by simple algebraic majorizations this can be easily shown to always subsist.

Degenerated parabolas (i.e. $\sigma_i^2 + \left(1 - \alpha\sigma_i^2\right) d_m = 0$): in this case $p_{i,m}(\alpha_g) = -\left(2\alpha\sigma_i^2\right)\alpha_g + \alpha$ is a negatively skewed line. Since it easy to verify that also in this case $p_{i,m}(b_-(i)) \leq 0$, it is true that condition (1.79) is always satisfied, for all m. Now, by simple algebraic manipulations, it can be shown that $\alpha_g \in [b_-(i), b_+(i)]$ is equivalent to condition (1.71). □

Notice that even if α_g is assumed to be the same among all the sensors, the bound (1.71) is different for each sensor i.

Remark 12. Assuming $\sigma_i^2 = \sigma^2$, $i = 1, \ldots, S$, exploiting

$$\alpha = \sum_{i=1}^{S} \frac{1}{\sigma^2} = \frac{S}{\sigma^2} \tag{1.81}$$

we can reformulate bound (1.71) as

$$S_g \in \left[\left(S - \sqrt{S^2 - S}\right),\ \left(S + \sqrt{S^2 - S}\right)\right]\ . \tag{1.82}$$

It is immediate to derive the conditions

$$S - \sqrt{S^2 - S} \leq 1 \Leftrightarrow 1 - \frac{1}{S} \leq \sqrt{1 - \frac{1}{S}} \tag{1.83}$$

and

$$S + \sqrt{S^2 - S} \geq 2S - 1 \Leftrightarrow \sqrt{1 - \frac{1}{S}} \geq 1 - \frac{1}{S} \tag{1.84}$$

thus (since $S \geq 1$) it follows that

$$[1, 2S - 1] \subset \left[\left(S - \sqrt{S^2 - S}\right),\ \left(S + \sqrt{S^2 - S}\right)\right]\ . \tag{1.85}$$

Since from a numerical point of view these two bounds are practically equivalent[3], we prefer to offer also bound (1.32) and its derivation because of its elegance.

[3] This is in a certain sense implied by the facts expressed in remark 7.

Before deriving other results it is interesting to analyze the asymptotic behavior of bound (1.71). For ease of notation we define

$$b_{-}(i) := \alpha - \sqrt{\alpha^2 - \frac{\alpha}{\sigma_i^2}} \qquad b_{+}(i) := \alpha + \sqrt{\alpha^2 - \frac{\alpha}{\sigma_i^2}}\,. \tag{1.86}$$

• if the topology and σ_i^2 are fixed but we vary the noisiness of sensors $j \neq i$, we have that

$$\exists j \text{ s.t. } \sigma_j^2 \to 0 \quad \Rightarrow \quad b_{-}(i) \to \frac{1}{2\sigma_i^2}, \quad b_{+}(i) \to +\infty \tag{1.87}$$

i.e. if there exists a sensor that has "perfect" measurements, then sensor i will improve its estimation with any guess α_{g} that is at least half of its precision $\frac{1}{\sigma_i^2}$. In the contrary, if

$$\forall j\ \sigma_j^2 \to +\infty \quad \Rightarrow \quad b_{-}(i) \to \frac{1}{\sigma_i^2}, \quad b_{+}(i) \to \frac{1}{\sigma_i^2}, \tag{1.88}$$

i.e. if all the sensors have unreliable measures then sensor i should use the local estimator (1.58);

• if the noisiness of all the sensors are the same (i.e. we are in the case analyzed in Section 1.2) but we vary the number of sensors S in the network, we have that

$$S \to +\infty \quad \Rightarrow \quad b_{-}(i) \to 0 \quad b_{+}(i) \to +\infty \tag{1.89}$$

• if the topology and the noisiness of all sensors j are fixed but the one of sensor i, and we vary it, then we have that

$$\sigma_i^2 \to 0 \quad \Rightarrow \quad b_{-}(i) \to +\infty, \quad b_{+}(i) \to +\infty \tag{1.90}$$

i.e. if the measurements of sensor i are "perfect" then sensor i should estimate without caring about the other sensors. In the contrary, if the measurements of sensor i are unreliable we should expect to have an improvement for every guess α_{g}. Unfortunately from bound (1.71) we obtain only the following

$$\sigma_i^2 \to +\infty \quad \Rightarrow \quad b_{-}(i) \to 0, \quad b_{+}(i) \to 2\alpha \tag{1.91}$$

i.e. a subset of the interval we were expecting. This is due to the fact that Theorem 11 gives only a sufficient condition for the optimality we are looking for.

As a general consideration, if sensor i is highly accurate while all the others are not, then bound (1.71) is thight for the sensor i (the accurate one), so it is more probable that the guessed α_{g} falls outside of its bound. Since (1.71) is a sufficient condition, it could be that, if α_{g} falls near outside the indicated interval, then still the distributed estimation is better than the local one also for the accurate sensor i. But if it falls far outside, this could become false.

We continue now stating some conditions that can be referred to the general behavior of the network, and not to the single sensor. The following, for example, assures that each sensor in the network has an advantage from the distributed algorithm:

Corollary 13. Define $\sigma^2_{\min} := \min_i \{\sigma_i^2\}$. Then if

$$\alpha_{\mathrm{g}} \in \left[\alpha - \sqrt{\alpha^2 - \frac{\alpha}{\sigma^2_{\min}}},\ \alpha + \sqrt{\alpha^2 - \frac{\alpha}{\sigma^2_{\min}}}\right] \tag{1.92}$$

then the variance of the estimation error of the distributed estimator $\widehat{b}_{\mathrm{d}}(\alpha_{\mathrm{g}})$ is smaller than the one of the local estimator $\widehat{b}^i_\ell$ for each sensor $i = 1, \ldots, S$.

Since in a distributed scenario it could be interesting to analyze *average* behaviors, it is important to answer to the following question: can we find values of α_{g} s.t. the variance of the error of the distributed strategy is smaller than the average error of the various local strategies, independently of the used prior Λ_0 and of the matrix C? The answer is given in the following:

Theorem 14. Considering the harmonic mean h defined in (1.66), if

$$\alpha_{\mathrm{g}} \in \left[\alpha - \sqrt{\alpha^2 - \frac{\alpha}{h}},\ \alpha + \sqrt{\alpha^2 - \frac{\alpha}{h}}\right] \tag{1.93}$$

then the variance of the estimation error of the distributed estimator $\widehat{b}_{\mathrm{d}}(\alpha_{\mathrm{g}})$ is smaller than the average variance of the estimation errors of the local estimators $\widehat{b}^i_\ell$.

Proof. We are seeking the guesses α_{g} such that

$$\frac{1}{S}\sum_{i=1}^{S} \mathrm{var}\left(b - \widehat{b}_{\mathrm{d}}(\alpha_{\mathrm{g}})\right) \leq \frac{1}{S}\sum_{i=1}^{S} \mathrm{var}\left(b - \widehat{b}^i_\ell\right) \tag{1.94}$$

and, repeating the initial steps of the proof of Theorem 4, we obtain the following sufficient condition:

$$\frac{-2}{d_m + \frac{1}{\alpha_{\mathrm{g}}}} + \frac{d_m + \frac{1}{\alpha}}{\left(d_m + \frac{1}{\alpha_{\mathrm{g}}}\right)^2} \leq \frac{1}{S}\sum_{i=1}^{S} \frac{-1}{d_m + \sigma_i^2} \qquad m = 1, \ldots, S\,. \tag{1.95}$$

We notice that if it is true that

$$\frac{-1}{d_m + h} \overset{?}{\leq} \frac{1}{S}\sum_{i=1}^{S} \frac{-1}{d_m + \sigma_i^2} \qquad \forall m \tag{1.96}$$

then we can repeat the other steps of proof of Theorem 11 to obtain the bound (1.93). Now condition (1.96) can be rewritten as

$$d_m + h \overset{?}{\leq} h\left(d_m + \sigma_1^2, \ldots, d_m + \sigma_S^2\right) \tag{1.97}$$

but, since $h = h\left(\sigma_1^2, \ldots, \sigma_S^2\right)$, this is true for the following Lemma 15. □

Lemma 15. If $a_i \geq 0$, $i = 1, \ldots, S$ and $d \geq 0$, then:

$$h(d + a_1, \ldots, d + a_S) \geq d + h(a_1, \ldots, a_S) \tag{1.98}$$

Proof. Defining:

$$f(d) := h(d+a_1,\ldots,d+a_S) - h(a_1,\ldots,a_S) - d \tag{1.99}$$

we need to prove that $f(d) \geq 0$ for $d \geq 0$. Since $f(0) = 0$, it is sufficient to demonstrate that $\frac{\partial f(d)}{\partial d} \geq 0$. Now this is true if:

$$S\sum_{i=1}^{S}\left(\frac{1}{d+a_i}\right)^2 \geq \left(\sum_{i=1}^{S}\frac{1}{d+a_i}\right)^2 . \tag{1.100}$$

Considering the two vectors $x = \left[\frac{1}{d+a_1},\ldots,\frac{1}{d+a_S}\right]^T$ and $y = [1,\ldots,1]^T$, condition (1.100) corresponds to $\langle x,x\rangle\langle y,y\rangle \geq |\langle x,y\rangle|^2$ that is the well-known Cauchy-Schwarz inequality. □

As expected, since the minimum element of the set of scalars is always smaller than the harmonic mean of this set, the interval described in bound (1.92) is always included in the interval described in bound (1.93), implying that condition (1.92) is sufficient for condition (1.93).

Uncertainity with respect to S: the previous results can be immediately reformulated as follows:

Corollary 16. If

$$S_{\mathrm{g}} \in \left[S - \sqrt{S^2 - \frac{Sh}{\sigma_i^2}},\ S + \sqrt{S^2 - \frac{Sh}{\sigma_i^2}}\right] \tag{1.101}$$

then the variance of the estimation error of the distributed estimator $\widehat{b}_{\mathrm{d}}$ is smaller than the one of the local estimator $\widehat{b}_\ell^i$, for every prior Λ_0, number of parameters E, sum of precisions α and matrix C.

Corollary 17. If

$$S_{\mathrm{g}} \in \left[S - \sqrt{S^2 - \frac{Sh}{\sigma_{\min}^2}},\ S + \sqrt{S^2 - \frac{Sh}{\sigma_{\min}^2}}\right] \tag{1.102}$$

then the variance of the estimation error of the distributed estimator $\widehat{b}_{\mathrm{d}}$ is smaller than the one of the local estimator $\widehat{b}_\ell^i$ for each sensor i.

Corollary 18. If

$$S_{\mathrm{g}} \in \left[S - \sqrt{S^2 - S},\ S + \sqrt{S^2 - S}\right] \tag{1.103}$$

then the variance of the estimation error of the distributed estimator $\widehat{b}_{\mathrm{d}}$ is smaller than the average variance of the estimation errors of the local estimators $\widehat{b}_\ell^i$.

Notice that corollary 18 is not independent of the various noises variances σ_i^2 since it implicitly requires the knowledge on their harmonic mean h.

1.3.4.2 Distributed versus Centralized Estimation

Following exactly the same reasonings made for Theorem 8, and using form (1.69) for estimator $\widehat{b}_{\mathrm{d}}$, it is possible to prove once again that:

Theorem 19. Under the assumption that $S \in [S_{\min}, S_{\max}]$ then

$$\frac{\left\|\widehat{b}_{\mathrm{d}} - \widehat{b}_{\mathrm{c}}\right\|_2}{\left\|\widehat{b}_{\mathrm{d}}\right\|_2} \leq \frac{S_{\max}}{S_{\min}} - 1 \tag{1.104}$$

for all $S_{\mathrm{g}} \in [S_{\min}, S_{\max}]$.

Once again it is possible to compute more refined bounds using the knowledge of C or Λ_0.

1.4 Case III: Totally Different Sensors

In this section we consider S distinct sensors each of them taking M scalar noisy measurements on different input locations, and let the measurement model be

$$y_i = C_i b + \nu_i, \qquad i = 1, \ldots, S \tag{1.105}$$

where S is the number of sensors, $y_i \in \mathbb{R}^M$ is the measurements vector collected by the i-th sensor, $b \in \mathbb{R}^E$ is the vector of unknown parameters modeled as a zero-mean Gaussian vector with autocovariance Λ_0, i.e. $b \sim \mathcal{N}(0, \Lambda_0)$. In addition, $\nu_i \in \mathbb{R}^M$ is the noise vector with density $\mathcal{N}\left(0, \sigma_i^2 I\right)$, independent of b and of ν_j, for $i \neq j$. Finally, $C_i \in \mathbb{R}^{M \times E}$ is a matrix that is in general different among sensors.

1.4.1 Local Bayesian Estimation

Under the assumptions above, the local MMSE estimator of b given y_i, is unbiased and given by

$$\begin{aligned} \widehat{b}_\ell^i &:= \mathbb{E}\left[b \mid y_i\right] = \operatorname{cov}(b, y_i) \operatorname{var}(y_i)^{-1} y_i \\ &= \Lambda_0 C_i^T \left(C_i \Lambda_0 C_i^T + \sigma_i^2 I\right)^{-1} y_i \\ &= \left(\Lambda_0^{-1} + \frac{C_i^T C_i}{\sigma_i^2}\right)^{-1} \frac{C_i^T y_i}{\sigma_i^2} \,. \end{aligned} \tag{1.106}$$

The autocovariance of the local estimation error $\widetilde{b}_i := b - \widehat{b}_\ell^i$ is given by

$$\Lambda_\ell^i := \operatorname{var}\left(b - \widehat{b}_\ell^i\right) = \left(\Lambda_0^{-1} + \frac{C_i^T C_i}{\sigma_i^2}\right)^{-1} \,. \tag{1.107}$$

1.4.2 Centralized Bayesian Estimation

Through computations similar to those of Section 1.4.2, it is possible to show that the MMSE estimate of the parameter vector b can be computed in a centralized way via the following

$$\widehat{b}_{\mathrm{c}} = \left(\frac{1}{S}\Lambda_0^{-1} + \frac{1}{S}\sum_{i=1}^{S} \frac{C_i^T C_i}{\sigma_i^2}\right)^{-1} \left(\frac{1}{S}\sum_{i=1}^{S} \frac{C_i^T y_i}{\sigma_i^2}\right) \tag{1.108}$$

while the variance of the estimation error is given by

$$\Lambda_{\mathrm{c}} := \mathrm{var}\left(\widehat{b}_{\mathrm{c}} - b\right) = \left(\Lambda_0^{-1} + \sum_{i=1}^{S} \frac{C_i^T C_i}{\sigma_i^2}\right)^{-1} . \tag{1.109}$$

Remark 20. We notice that the measurements-based form of estimator $\widehat{b}_{\mathrm{c}}$ is given by

$$\widehat{b}_{\mathrm{c}} = \Lambda_0 \left[C_1^T, \ldots, C_S^T\right] \left(\begin{bmatrix} C_1 \\ \vdots \\ C_S \end{bmatrix} \Lambda_0 \left[C_1^T, \ldots, C_S^T\right] + \mathrm{diag}\left(\sigma_i^2 I\right) \right)^{-1} \begin{bmatrix} y_1 \\ \vdots \\ y_S \end{bmatrix} \tag{1.110}$$

where $\mathrm{diag}\left(\sigma_i^2 I\right)$ indicates a block-diagonal matrix which diagonal blocks are given by $\sigma_i^2 I$ matrices. Computation of $\widehat{b}_{\mathrm{c}}$ through (1.110) involves $M \cdot S \times M \cdot S$ matrices, while in the previous cases of Sections 1.2 and 1.3 it was involving smaller matrices (only $M \times M$-dimensional ones).

1.4.3 Distributed Bayesian Estimation

Given remark 20, it is clear that the distributed estimator $\widehat{b}_{\mathrm{d}}$ of b can be derived only from the innovations-based form of the centralized estimator $\widehat{b}_{\mathrm{c}}$. Given also remark 2, we obtain that $\widehat{b}_{\mathrm{d}}$ is given by

$$\widehat{b}_{\mathrm{d}}\left(S_{\mathrm{g}}\right) := \left(\frac{1}{S_{\mathrm{g}}}\Lambda_0^{-1} + \frac{1}{S}\sum_{i=1}^{S} \frac{C_i^T C_i}{\sigma_i^2}\right)^{-1} \left(\frac{1}{S}\sum_{i=1}^{S} \frac{C_i^T y_i}{\sigma_i^2}\right) . \tag{1.111}$$

We notice that, in order to compute $\widehat{b}_{\mathrm{d}}$ through (1.111), sensors have to perform two average-consensi: one on $C_i^T C_i/\sigma_i^2$ ($E \times E$-dimensional matrices), and one on $C_i^T y_i/\sigma_i^2$ (E-dimensional vectors). We also notice that the inversion of the matrix $\left(\frac{1}{S_{\mathrm{g}}}\Lambda_0^{-1} + \frac{1}{S}\sum_{i=1}^{S} \frac{C_i^T C_i}{\sigma_i^2}\right)$ can be performed only after the consensus processes, thus it has to be performed on-line by the sensors and not in an off-line fashion.

Remark 21. It would have been easy to consider zero-mean measurement noises that are correlated if belonging to the same sensor, or uncorrelated if not, i.e. s.t.

$$\mathbb{E}\left[\nu_i \nu_j^T\right] = \begin{cases} \Sigma_{\nu,i} & \text{if } i = j \\ 0 & \text{otherwise.} \end{cases} \tag{1.112}$$

In this case, in fact, the changes that should be applied to estimator (1.111) would be

$$\frac{C_i^T C_i}{\sigma_i^2} \mapsto C_i^T \Sigma_{\nu,i}^{-1} C_i \qquad \text{and} \qquad \frac{C_i^T y_i}{\sigma_i^2} \mapsto C_i^T \Sigma_{\nu,i}^{-1} y_i \, . \tag{1.113}$$

In the most general case of zero-mean noises that are correlated even if taken by different sensors, it is easy to show that there exist no distributed estimators like the ones proposed up to now that exactly compute the centralized solution, even assuming the knowledge of the number of sensors in the network S, or the sum of the precisions α.

The autocovariance for the estimation error is given in this case by

$$\begin{aligned}\Lambda_{\mathrm{d}} &:= \operatorname{var}\left(\widehat{b}_{\mathrm{d}} - b\right) \\ &= \left(\frac{1}{S_{\mathrm{g}}}\Lambda_0^{-1} + \frac{1}{S}\sum_{i=1}^{S}\frac{C_i^T C_i}{\sigma_i^2}\right)^{-1} \cdot \\ &\quad \cdot \left(\frac{1}{S_{\mathrm{g}}^2}\Lambda_0^{-1} + \frac{1}{S^2}\sum_{i=1}^{S}\frac{C_i^T C_i}{\sigma_i^2}\right) \cdot \\ &\quad \cdot \left(\frac{1}{S_{\mathrm{g}}}\Lambda_0^{-1} + \frac{1}{S}\sum_{i=1}^{S}\frac{C_i^T C_i}{\sigma_i^2}\right)^{-1} .\end{aligned} \tag{1.114}$$

1.4.4 Characterization of the Distributed Algorithm

We seek now to characterize the distributed estimator we just introduced as we did in Section 1.2.4 and 1.3.4.

1.4.4.1 Distributed versus Local Estimation

Unfortunately in this case it is not possible to derive explicit sufficient conditions assuring the distributed estimator to perform better than the local one. In fact both Theorem and rely on finding an orthonormal matrix U diagonalizing the matrix $\left(C\Lambda_0 C^T + \sigma_i^2 I\right)$, that was obtained from the measurements-based forms of the distributed estimators $\widehat{b}_{\mathrm{d}}$ (see 1.31 and 1.68). But this form cannot be derived in this novel case, see remark (20). We are currently investigating other research directions in order to circumvent this fact.

1.4.4.2 Distributed versus Centralized Estimation

Once again it is possible to show that Theorem 8 is still valid, using the same reasonings of the original proof.

2 Distributed Nonparametric Estimation

2.1 Introduction

In regression and system identification the adjective "nonparametric" is usually referred to techniques that do not fix a priori the structure of the result. If not acquainted with these techniques, this lack of structure may initially seem a negative characteristic. On the contrary, years of application on real fields shown that their usage is motivated by various practical and mathematical reasons like:

- if there is a lack of knowledge of the model to be identified or if the model belongs to a family of different parametric models, then nonparametric identification leads to better estimates (Pillonetto and De Nicolao, 2010). A specific example is Pillonetto et al. (2011), where authors prove that in some practical cases the identification of linear systems through combination of classical model selection strategies, like Akaike Information Criterion (AIC) (Akaike, 1974) or Bayesian Information Criterion (BIC) (Schwarz, 1978), and Prediction Error Methods (PEM) strategies (Ljung, 1999; Söderström and Stoica, 1989) performs worse than identification through nonparametric Gaussian regression approaches;

- nonparametric identification approaches can be consistent where parametric approaches fail to be (Smale and Zhou, 2007; De Nicolao and Ferrari-Trecate, 1999);

- in general, nonparametric approaches require the tuning of very few parameters, allowing the implementation of fast line search strategies (Pillonetto and Bell, 2007);

- for some parametric models, the distributed implementation of ML strategies could be infeasible, due to the structure of the likelihood function. An approach is then to convexify -as it will be clear later- the likelihood through the construction a suitable nonparametric approximated model. In general this allows the application of generic distributed optimization techniques (Bertsekas

and Tsitsiklis, 1997). But, under particular choices of the cost and regularization functions, we will show that this can distributedly solve the approximated ML problem through a distributed approximated Regularization Network (RN) requiring small computational, communication and memory requirements.

An other important point is the following: the amount of prior information given with nonparametric techniques (e.g. the kernel functions introduced below, that can be considered as *covariances* whenever using Bayesian approaches based on Gaussian processes, see Section 2.1.4) is far less than the amount of prior information that is given assuming the model to be a certain parametric function. Intuitively the prior of the nonparametric techniques is *weaker* than the parametric one, and this eventually makes the nonparametric strategies more widely applicable and more robust. Nonetheless this is a tricky point: in fact, if an experiment returned a small amount of data, small information is available. In this case, if the parametric model is (from an intuitive point of view) *accurate*, the amount of information could be sufficient to obtain an estimate far better than the one that could be obtained with the less informative nonparametric prior.

After this warning, we focus now on a nonparametric Reproducing Kernel Hilbert Space (RKHS) based approach, one of the most widely used approaches for regression purposes.

2.1.1 Background

From an intuitive point of view, RKHSs are sets of sufficiently-smooth functions with some nice mathematical properties. The theory was founded by (Aronszajn, 1950). See also (Yosida, 1965; Cucker and Smale, 2002; Poggio and Girosi, 1990; Wahba, 1990). For an overview of their uses in statistical signal processing see Weinert (1982).

Definition 22 (Reproducing kernel Hilbert space)**.** Let $\mathcal{H}_K$ be a Hilbert space of functions[1]

$$f(\cdot) : \mathcal{X} \subseteq \mathbb{R}^d \mapsto \mathbb{R} \tag{2.1}$$

endowed with the inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}_K}$ and norm $\|f\|_{\mathcal{H}_K} := \sqrt{\langle f, f \rangle_{\mathcal{H}_K}}$. If there exists a function

$$K(\cdot, \cdot) : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R} \tag{2.2}$$

such that

(a) $K(x, \cdot) \in \mathcal{H}_K$ for every $x \in \mathcal{X}$

(b) $\langle f(\cdot), K(x, \cdot) \rangle_{\mathcal{H}_K} = f(x)$ for every $x \in \mathcal{X}$ and $f \in \mathcal{H}_K$

then $\mathcal{H}_K$ is said to be a *reproducing kernel Hilbert space* with kernel K.

[1] We restrict our analysis only real-valued functions even if the same concepts could be applied to complex-valued functions.

Property (b) is usually called the *reproducing property*. Notice that $\mathcal{L}^2$ is not a RKHS since its representing functions, namely the delta functions, are not in $\mathcal{L}^2$. For the following derivations it is necessary to introduce some definitions.

Definition 23 (Positive-definite kernel)**.** A kernel K is said to be *positive-definite* if, for every $N \in \mathbb{N}_+$ and N-tuple $x_1, \ldots, x_N \in \mathcal{X}$

$$\begin{bmatrix} K(x_1, x_1) & \cdots & K(x_1, x_N) \\ \vdots & & \vdots \\ K(x_N, x_1) & \cdots & K(x_N, x_N) \end{bmatrix} =: \mathbf{K} \geq 0 \tag{2.3}$$

where the inequality has to be intended in a matricial positive-semidefinite sense.

Definition 24 (Symmetric kernel)**.** A kernel K is said to be *symmetric* if $K(x, x') = K(x', x)$ for all $x, x' \in \mathcal{X}$.

Definition 25 (Mercer kernel)**.** A symmetric positive-definite kernel K is said to be a *Mercer kernel* if it is also continuous.

The term *kernel* derives from the theory of integral operators, where, given a non-degenerate measure[2] μ and a function K as in 2.2, it is possible to define the integral operator

$$L_{K,\mu}[g](x) := \int_{\mathcal{X}} K(x, x') g(x') d\mu(x') \ . \tag{2.4}$$

Operator $L_{K,\mu}[\cdot]$ is said to be *positive definite* if K is positive definite.

The following theorem proves the biunivocity between symmetric positive-definite kernels and RKHSs.

Theorem 26 (Moore-Aronszajn (Aronszajn, 1950))**.** For every symmetric positive-definite kernel K there exists an unique RKHS $\mathcal{H}_K$ having K as its reproducing kernel. Viceversa, the reproducing kernel of every RKHS $\mathcal{H}_K$ is unique.

Having in mind our future applications on regression, we focus now on the implications of the spectral theory of compact operators on RKHS theory[3]. Assume then $\mathcal{X}$ to be compact, K to be Mercer on $\mathcal{X} \times \mathcal{X}$, $\mathcal{L}^2(\mu)$ to be the set of the Lebesgue square integrable functions under the non-degenerate measure μ. A function ϕ that obeys the integral equation[4]

$$\lambda \phi(x) = L_{K,\mu}[\phi](x) \tag{2.5}$$

is said to be an *eigenfunction* of $L_{K,\mu}[\cdot]$ with associated eigenvalue λ. The following result holds.

[2] We recall that a Borel measure μ is said to be non-degenerate w.r.t. the Lebesgue measure $\mathcal{L}^2$ if $\mathcal{L}^2(A) > 0 \Rightarrow \mu(A) > 0$ for every A in the Borel σ-algebra.

[3] See (Zhu, 2007, Chap. 1.3) for more details on compact operators on general Hilbert spaces.

[4] In some cases eigenvalues and eigenfunctions can be computed in closed forms, specially in Gaussian cases (Zhu et al., 1998). Often it is necessary to perform numerical computations (De Nicolao and Ferrari-Trecate, 1999), (Rasmussen and Williams, 2006, Chap. 4.3.2).

Theorem 27 (Cucker and Smale (2002), see also König (1986))**.** Let K be a Mercer kernel on $\mathcal{X} \times \mathcal{X}$ and μ a non-degenerate measure. Let $\{\phi_e\}$ be the eigenfunctions of $L_{K,\mu}[\cdot]$ normalized in $\mathcal{L}^2(\mu)$, i.e. s.t.

$$\int_{\mathcal{X}} \phi_e(x)\,\phi_l(x)\,d\mu(x) = \delta_{el} \tag{2.6}$$

with corresponding eigenvalues λ_e ordered s.t. $\lambda_1 \geq \lambda_2 \geq \ldots$. Then

(a) $\lambda_e \geq 0$ for all e;

(b) $\sum_{e=1}^{+\infty} \lambda_e = \int_{\mathcal{X}} K(x,x)\,d\mu(x) < +\infty$

(c) $\{\phi_e\}_{e=1}^{+\infty}$ is an orthonormal basis for $\mathcal{L}^2(\mu)$

(d) the RKHS $\mathcal{H}_K$ associated to $\{\phi_e\}_{e=1}^{+\infty}$ is given by

$$\mathcal{H}_K := \left\{ g \in \mathcal{L}^2(\mu) \;\text{ s.t. }\; g = \sum_{e=1}^{\infty} a_e \phi_e \;\text{ with }\; \{a_e\} \;\text{ s.t. }\; \sum_{e=1}^{\infty} \frac{a_e^2}{\lambda_e} < +\infty \right\} \tag{2.7}$$

(e) K can be expanded via the relation:

$$K(x,x') = \sum_{e=1}^{\infty} \lambda_e \phi_e(x)\,\phi_e(x') \tag{2.8}$$

where the convergence of the series is absolute and uniform[5] in $\mathcal{X} \times \mathcal{X}$.

Remark 28. Condition $\sum_{e=1}^{\infty} \frac{a_e^2}{\lambda_e} < +\infty$ expressed in (2.7) can be seen as a smoothness condition. In fact, since the sequence $\lambda_1, \lambda_2, \ldots$ has to vanish because the associated series is convergent, it follows that a_e^2 must vanish sufficiently fast.

From the same theorem it follows that if $g_1 = \sum_{e=1}^{+\infty} a_e \phi_e$ and $g_2 = \sum_{e=1}^{+\infty} a'_e \phi_e$ then their inner product is

$$\langle g_1, g_2 \rangle_{\mathcal{H}_K} = \sum_{e=1}^{+\infty} \frac{a_e \cdot a'_e}{\lambda_e}\,. \tag{2.9}$$

Notice that, if $g = \sum_{e=1}^{+\infty} a_e \phi_e \in \mathcal{H}_K$ and $\mathbf{a} = [a_1, a_2, \ldots]^T$, orthogonality of eigenfunctions in $\mathcal{L}^2(\mu)$ implies that

$$\|g\|^2_{\mathcal{L}^2(\mu)} = \sum_{e=1}^{+\infty} \sum_{l=1}^{+\infty} a_e a_l \int_{\mathcal{X}} \phi_e(x)\,\phi_l(x)\,d\mu(x) = \|\mathbf{a}\|_2^2\,. \tag{2.10}$$

Moreover orthonormality of eigenfunctions in $\mathcal{L}^2(\mu)$ implies orthogonality in $\mathcal{H}_K$, i.e.

$$\langle \phi_e, \phi_l \rangle_{\mathcal{L}^2(\mu)} = \delta_{el} \qquad \Leftrightarrow \qquad \langle \phi_e, \phi_l \rangle_{\mathcal{H}_K} = \frac{1}{\lambda_e}\delta_{el}\,. \tag{2.11}$$

In the following we will use the shorthands $\|\cdot\|_\mu$ for $\|\cdot\|_{\mathcal{L}^2(\mu)}$ and $\|\cdot\|_K$ for $\|\cdot\|_{\mathcal{H}_K}$.

[5] This has the nice practical implication that it is possible to compute K with the desired level of precision using a finite number of eigenfunctions.

Remark 29. We could have defined $\mathcal{H}_K$ using the so-called *reproducing kernel map construction* (Rasmussen and Williams, 2006, page 131), i.e. starting from the representing functions $K(x, \cdot)$. We preferred to use eigenfunctions-eigenvalues decompositions because these will be heavily used in the following sections.

2.1.2 Examples of RKHSs

In this section we offer some examples of the most commonly used kernels, focusing on the case $\mathcal{X} = [0, 1]$. We send the reader back to (Schölkopf and Smola, 2001, Chap. 13) and references therein for general kernels design techniques.

Gaussian Kernels A Gaussian kernel is described by

$$K(x, x') = \exp\left(-\frac{\|x - x'\|_2^2}{2\sigma^2}\right) \tag{2.12}$$

where $x, x' \in \mathcal{X} \subset \mathbb{R}^d$ ($\mathcal{X}$ is a compact). This kernel may have eigenfunctions and eigenvalues in closed forms, depending on μ, see for example Zhu et al. (1998).

In Figures 2.1 and Figure 2.2 we plot the first 4 eigenfunctions for the cases $\mu = \mathcal{U}[0, 1]$ and $\mu = \mathcal{N}(0.5, 0.01)$, both with $\sigma^2 = 0.01$. We notice how the approximation capability of the eigenfunctions is concentrated where it is more probable to have measurements. In Figure 2.3 we show the behavior of the eigenvalues for the two different μ's. Finally in Figure 2.4 we show 4 different realizations f_μ relative to the kernel just considered, under the assumptions of Section 2.1.4.

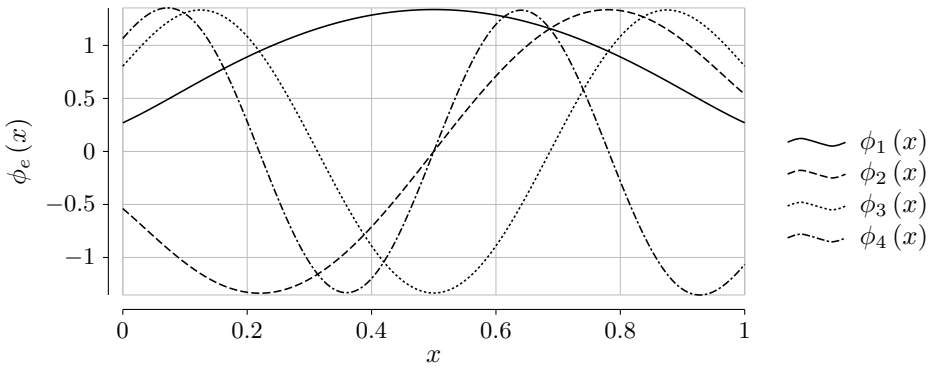


Figure 2.1: First 4 eigenfunctions for the Gaussian kernel (2.12), associated to $\mu = \mathcal{U}[0, 1]$ and $\sigma^2 = 0.01$.

Laplacian Kernels A Laplacian kernel is described by

$$K(x, x') = \exp\left(-\frac{|x - x'\|}{\sigma}\right) \tag{2.13}$$

where $x, x' \in \mathcal{X} \subset \mathbb{R}^d$, $\sigma \in \mathbb{R}_+$.

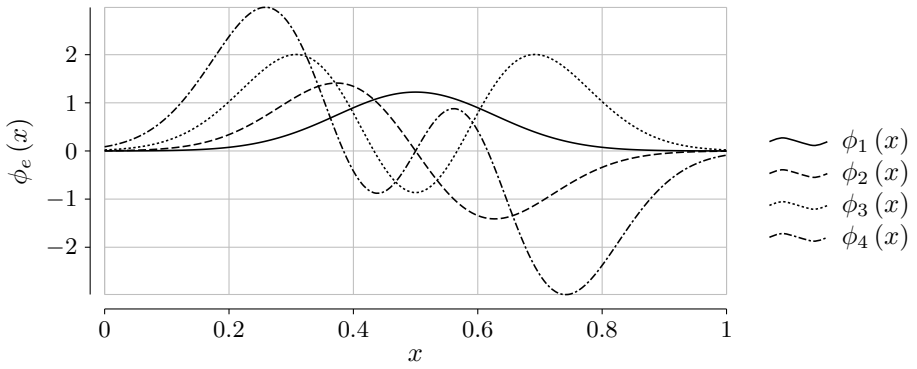


Figure 2.2: First 4 eigenfunctions for the Gaussian kernel (2.12), associated to $\mu = \mathcal{N}(0.5, 0.01)$ and $\sigma^2 = 0.01$.

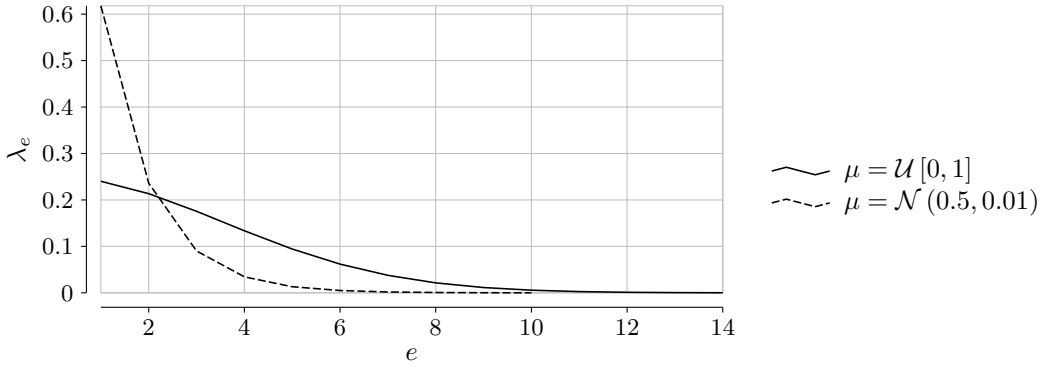


Figure 2.3: Eigenvalues of the Gaussian kernel (2.12), associated to $\sigma^2 = 0.01$ and different measures μ.

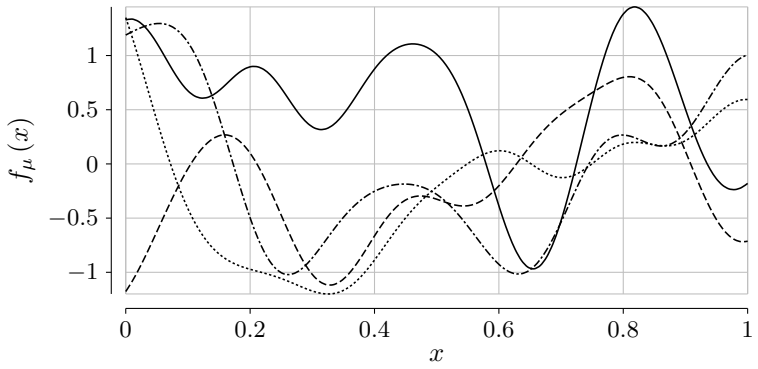


Figure 2.4: Independently generated realizations for the Gaussian kernel (2.12), associated to $\sigma^2 = 0.01$.

In Figure 2.5 we plot the first 4 eigenfunctions for the case $\mu = \mathcal{U}[0,1]$ with $\sigma = 0.1$. In Figure 2.6 we show the behavior of the eigenvalues for this kernel, and in Figure 2.7 we show 4 different realizations f_μ relative to the kernel just considered, again under the assumptions of Section 2.1.4.

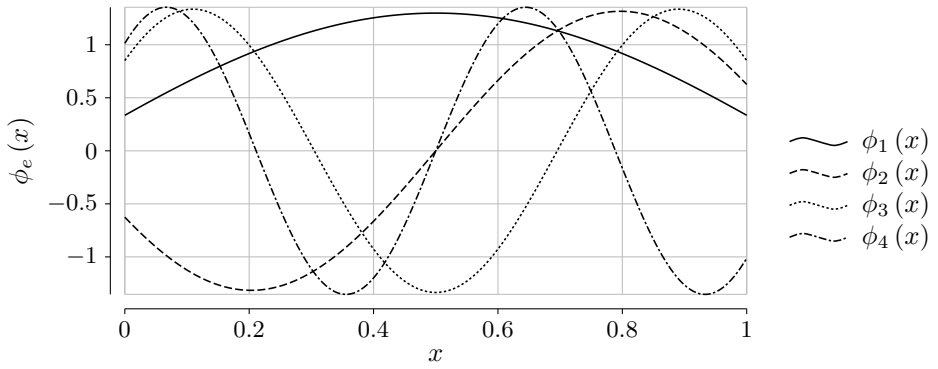


Figure 2.5: First 4 eigenfunctions for the Laplacian kernel (2.13), associated to $\mu = \mathcal{U}[0,1]$ and $\sigma = 0.1$.

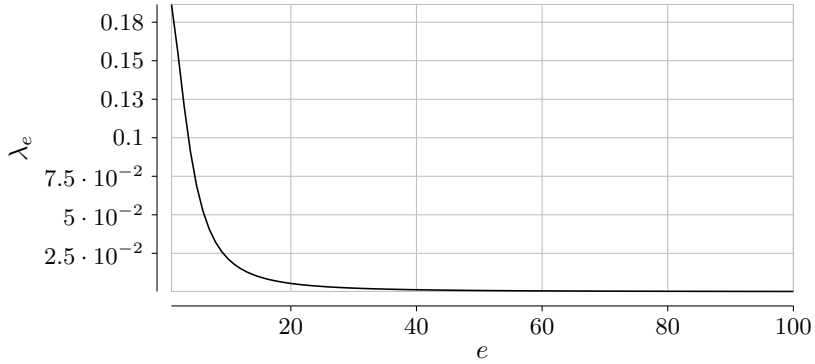


Figure 2.6: Eigenvalues of the Laplacian kernel (2.13), associated to $\mu = \mathcal{U}[0,1]$ and $\sigma = 0.1$.

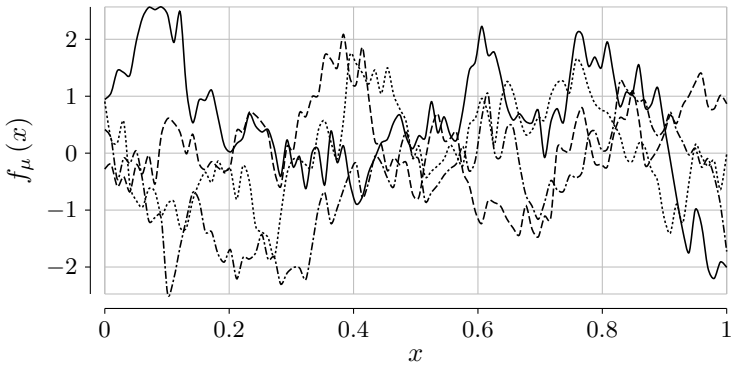


Figure 2.7: Independently generated realizations for the Laplacian kernel (2.13), associated to $\sigma = 0.1$.

Stable-Splines Kernels When f_μ is the impulse response of a BIBO-stable system, it is better to choose a prior incorporating this information. The kernel of equation (2.14) (parametrized with $\beta \geq 0$) has been proved to lead to better approximations of such responses than cubic spline kernel does (Pillonetto and De Nicolao,

2010). Corresponding eigenfunctions for the case $\mu = \mathcal{U}[0,1]$, that are causal and decrease exponentially as time goes to infinity, are plotted in Figure 2.8.

$$K\left(x, x'; \beta\right) = \begin{cases} \dfrac{\exp\left(-2\beta x\right)}{2}\left(\exp\left(-\beta x'\right) - \dfrac{\exp\left(-\beta x\right)}{3}\right) & \text{if } x \leq x' \\ \dfrac{\exp\left(-2\beta x'\right)}{2}\left(\exp\left(-\beta x\right) - \dfrac{\exp\left(-\beta x'\right)}{3}\right) & \text{if } x \geq x' \end{cases} \tag{2.14}$$

In Figure 2.8 we plot the first 4 eigenfunctions for the case $\mu = \mathcal{U}[0,1]$ with $\sigma = 0.1$. In Figure 2.9 we show the behavior of the eigenvalues for this kernel, and in Figure 2.10 we show 4 different realizations f_μ relative to the kernel just considered, again under the assumptions of Section 2.1.4.

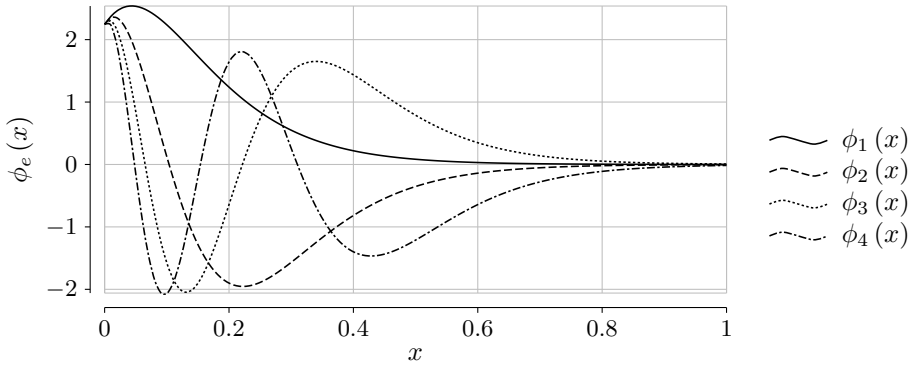


Figure 2.8: First 4 eigenfunctions for the kernel (2.14), associated to $\mu = \mathcal{U}[0,1]$ and $\beta = 5$.

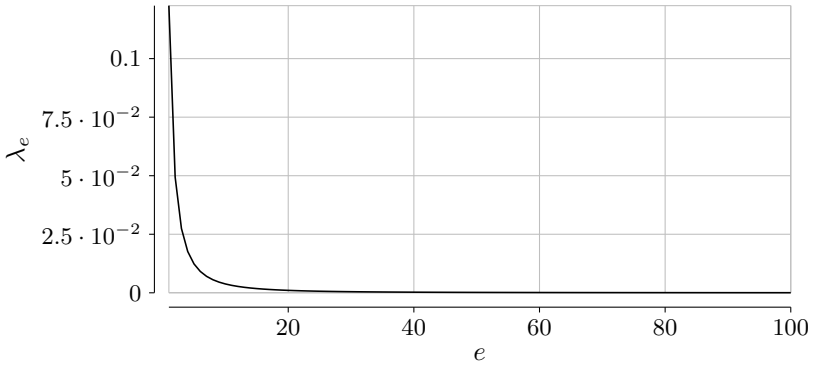


Figure 2.9: Eigenvalues of the kernel (2.14), associated to $\beta = 5$ and the measure $\mu = \mathcal{U}[0,1]$.

2.1.3 Regularized regression

Let $f_\mu : \mathcal{X} \to \mathbb{R}$ denote an unknown deterministic function defined on the compact $\mathcal{X} \subset \mathbb{R}^d$. Assume we have the following S noisy measurements

$$y_i = f_\mu\left(x_i\right) + \nu_i, \qquad i = 1, \ldots, S \tag{2.15}$$

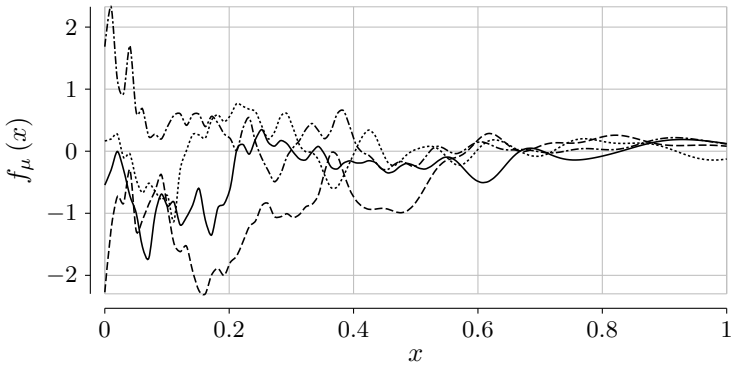


Figure 2.10: Independently generated realizations for the kernel (2.14), associated to $\beta = 5$.

with ν_i white noise and i the measurement index. Without any additional assumption, the problem of inferring f_μ given the data set $\{x_i, y_i\}_{i=1}^S$ is ill-posed in the sense of Hadamard. One of the most used approaches to overcome this problem relies upon the Tikhonov regularization theory[6][7] (Tikhonov and Arsenin, 1977), that relies computing the estimate of the unknown function as

$$\widehat{f}_{\mathrm{c}} := \arg \min_{f \in \mathcal{H}_K} Q(f) \tag{2.16}$$

where the functional $Q(\cdot)$ is defined as

$$Q(f) := L\left(f, \{x_i, y_i\}_{i=1}^S\right) + \gamma \|f\|_K^2 \tag{2.17}$$

and where the hypothesis space $\mathcal{H}_K$ is typically given by the reproducing kernel Hilbert space induced by the Mercer kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. The first term is a loss function accounting for data-fitting properties of f (see Figure 1.1 and related comments), while the second term, usually called *regularizer*, weights the smoothness of f, penalizing thus non-smooth solutions[8]. Finally, γ is the so called *regularization parameter* that trades off empirical evidence and smoothness information on f_μ.

By using the famous *representer theorem* (introduced in Kimeldorf and Wahba (1971), see (Schölkopf and Smola, 2001, Chap. 4.2) for a generalized version) it is possible to show that each minimizer of $Q(f)$ has the form of a linear combination of S basis functions, i.e.

$$\widehat{f}_{\mathrm{c}} = \sum_{i=1}^{S} c_i K(x_i, \cdot) \tag{2.18}$$

i.e. $\widehat{f}_{\mathrm{c}}$ admits the structure of a *Regularization Network* (RN), term introduced in Poggio and Girosi (1990) to indicate estimates of the form (2.18).

[6] Alternatively one could use explicit prior knowledge, and formulate the problem -for example- through Gaussian Processes formalisms.

[7] Finite-dimensional formulation of this approach is also known as *Ridge regression* (Hoerl and Kennard, 2000)

[8] See Girosi et al. (1995) for smoothness functionals involving Fourier transforms of the candidate estimating function.

A graphical intuition of (2.18) is that the optimal estimate is given by a combination of some "slices" of the kernel function, as shown in Figure 2.11.

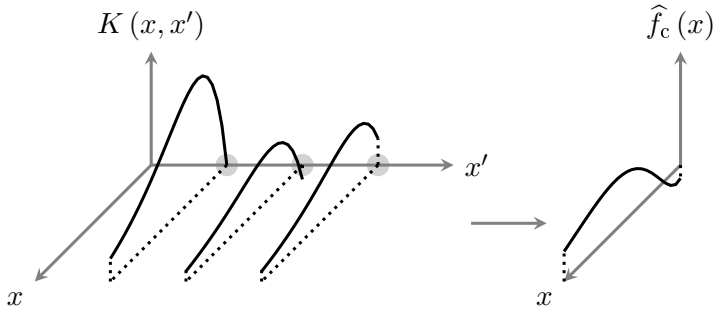


Figure 2.11: Graphical intuition of the structure of the optimal centralized estimate $\widehat{f}_c$. Assume measurements have been taken in correspondence of the input locations highlighted with gray circles in the x' axis: then $\widehat{f}_c$ is given by a linear combination of the kernel function $K(\cdot, \cdot)$ "sliced" along the various input locations contained in the dataset. In the figure, each slice is indicated with a black solid line.

For consistency with Chapter 1 and in sight of the Bayesian interpretation that will be introduced in Section 2.1.4, our choice for the cost function is

$$Q(f) := \sum_{i=1}^{S} (y_i - f(x_i))^2 + \gamma \|f\|_K^2 \tag{2.19}$$

that correspond to obtain the coefficients c_i by means of

$$\begin{bmatrix} c_1 \\ \vdots \\ c_S \end{bmatrix} = (\mathbf{K} + \gamma I)^{-1} \begin{bmatrix} y_1 \\ \vdots \\ y_S \end{bmatrix} \tag{2.20}$$

with

$$\mathbf{K} := \begin{bmatrix} K(x_1, x_1) & \cdots & K(x_1, x_S) \\ \vdots & & \vdots \\ K(x_S, x_1) & \cdots & K(x_S, x_S) \end{bmatrix} . \tag{2.21}$$

2.1.4 Bayesian interpretation

The estimate $\widehat{f}_c$ in (2.16) computed through (2.20) admits also a Bayesian interpretation. In fact, if f_μ is modeled as the realization of a zero-mean, not-necessarily stationary Gaussian random field with covariance K, if the noises ν_i are Gaussian and independent of the unknown function and with variance σ^2, once we set $\gamma = \sigma^2$ it follows that (Kimeldorf and Wahba, 1970; Zhu et al., 1998)

$$\widehat{f}_c(x) = \mathbb{E}[f_\mu(x) \mid x_1, y_1, \ldots, x_S, y_S] . \tag{2.22}$$

Hence, under such Bayesian perspective, the problem of reconstructing f_μ is a generalization of that discussed in Section 1, which increased complexity derives from the

fact that the problem is now infinite-dimensional. We recall that, using the Bayesian point of view and a Gaussian Processs (GPs) based formulation, it is straightforward to derive not only the estimate (to be intended as the maximum a-posteriori of the conditional density), but also to characterize the uncertainity of the prediction by means of the a-posteriori covariance. Moreover GPs formulation is closely related to *Kriging* techniques (Stein, 1999), usually used for interpolation of spatial data.

2.1.5 Computation of Approximated Solutions

The computation of $\widehat{f_c}$ defined in (2.16) through (2.20) requires $O\left(S^3\right)$ operations and requires the processing unit to have stored all the x_i's. This can be impractical in a distributed estimation scenario, where agents may have both limited computational capabilities and constraints in the amount of communicable information. To overcome these problems, we derive an alternative distributed estimation strategy by restricting the hypothesis space to a closed subspace $\breve{\mathcal{H}}_K \subset \mathcal{H}_K$. The following proposition shows that the resulting estimator has favorable theoretical properties. In particular, as the number of measurements S goes to $+\infty$, it returns the best possible approximation of f_μ in $\breve{\mathcal{H}}_K$. The result is obtained by a variation of the arguments used in Smale and Zhou (2007, 2005) to characterize the estimator (2.16).

Proposition 30. Let $0 < \delta < 1$ and the following closed subspace in $\mathcal{H}_K$

$$\breve{\mathcal{H}}_K := \overline{\operatorname{span}_{e\in\mathcal{I}}\{\phi_e\}} \tag{2.23}$$

where the overline denotes the closure in $\mathcal{H}_K$ and $\mathcal{I} \subset \mathbb{N}_+$. Define

$$\widehat{f}_S := \arg\min_{f\in\breve{\mathcal{H}}_K} \sum_{i=1}^{S} \frac{\left(y_i - f\left(x_i\right)\right)^2}{S} + \gamma \left\|f\right\|_K^2 \ . \tag{2.24}$$

Then, assuming $|y_i| \leq \overline{Y}$ a.s., $\widehat{f}_S$ converges in probability to the projection of f_μ onto $\breve{\mathcal{H}}_K$, denoted by $f_\mu^{\breve{\mathcal{H}}_K}$. In particular, let

$$\gamma = \frac{8\overline{K}^2 \log\left(\frac{4}{\delta}\right)}{\sqrt{S}} \tag{2.25}$$

where

$$\overline{K} := \sup_{x,x'\in\mathcal{X}} \sqrt{K\left(x,x'\right)} \tag{2.26}$$

Then, with confidence $1-\delta$ one has

$$\left\|\widehat{f}_S - f_\mu\right\|_\mu \leq \frac{\sqrt{2\log\left(\frac{4}{\delta}\right)}}{S^{\frac{1}{4}}} \left(3\overline{Y} + 2\overline{K}\left\|f_\mu^{\breve{\mathcal{H}}_K}\right\|_K\right) + \left\|f_\mu^{\breve{\mathcal{H}}_K^\perp}\right\|_\mu \tag{2.27}$$

where $f_\mu^{\breve{\mathcal{H}}_K^\perp}$ is the projection of f_μ onto the orthogonal of $\breve{\mathcal{H}}_K$ in $\mathcal{H}_K$.

Proof. We start noticing that it is possible to associate with $\breve{\mathcal{H}}_K$ the restricted kernel $\breve{K}$ and the relative integral operator $L_{\breve{K},\mu}$ defined respectively by

$$\breve{K}\left(x,x'\right) := \sum_{e\in\mathcal{I}} \lambda_e \phi_e\left(x\right)\phi_e\left(x'\right) \tag{2.28}$$

and

$$L_{\breve{K},\mu}\left[g\right]\left(x\right) := \int_{\mathcal{X}} \breve{K}\left(x,x'\right) g\left(x'\right) d\mu\left(x'\right) \ . \tag{2.29}$$

Exploiting RKHS theory, see e.g. Cucker and Smale (2002), one obtains that $\breve{\mathcal{H}}_K$ is exactly the image of $L^{\frac{1}{2}}_{\breve{K},\mu}$ (the square root of $L_{\breve{K},\mu}$) fed with $\mathcal{L}^2\left(\mu\right)$, i.e.

$$\breve{\mathcal{H}}_K = L^{\frac{1}{2}}_{\breve{K},\mu}\left[\mathcal{L}^2\left(\mu\right)\right] \ . \tag{2.30}$$

Now, define

$$\widehat{f}_\gamma := \arg\min_{f\in\breve{\mathcal{H}}_K} \left\|f - f_\mu\right\|^2_\mu + \gamma\left\|f\right\|^2_K \tag{2.31}$$

and notice that

$$\widehat{f}_\gamma = \arg\min_{f\in\breve{\mathcal{H}}_K} \left\|f - f_\mu^{\breve{\mathcal{H}}_K}\right\|^2_\mu + \gamma\left\|f\right\|^2_K \ . \tag{2.32}$$

In addition, it holds that

$$\begin{aligned} \left\|\widehat{f}_S - f_\mu\right\|_\mu &\leq \left\|\widehat{f}_S - \widehat{f}_\gamma\right\|_\mu + \left\|\widehat{f}_\gamma - f_\mu\right\|_\mu \\ &\leq \left\|\widehat{f}_S - \widehat{f}_\gamma\right\|_\mu + \left\|\widehat{f}_\gamma - f_\mu^{\breve{\mathcal{H}}_K}\right\|_\mu + \left\|f_\mu^{\breve{\mathcal{H}}_K^\perp}\right\|_\mu \ . \end{aligned} \tag{2.33}$$

Using theorem 5 in Smale and Zhou (2007), we know that if (2.25) holds, one has

$$\left\|\widehat{f}_S - \widehat{f}_\gamma\right\|_\mu \leq \frac{12\overline{KY}\log\left(\frac{4}{\delta}\right)}{\sqrt{\gamma S}} \ . \tag{2.34}$$

In addition, exploiting theorem 3 in Smale and Zhou (2005) and the definition of $L_{\breve{K},\mu}$, it is easy to obtain that, with confidence $1-\delta$, one has

$$\left\|\widehat{f}_\gamma - f_\mu^{\breve{\mathcal{H}}_K}\right\|_\mu \leq \sqrt{\gamma}\left\|L^{-\frac{1}{2}}_{\breve{K},\mu}\left[f_\mu^{\breve{\mathcal{H}}_K}\right]\right\|_\mu = \sqrt{\gamma}\left\|f_\mu^{\breve{\mathcal{H}}_K}\right\|_K \tag{2.35}$$

where the last equality exploits (2.30) and the fact that $L^{\frac{1}{2}}_{\breve{K},\mu}$ is an isometric map. The proposition is then proved after simple computations once (2.34), (2.35) and (2.25) are substituted into (2.33). □

We consider now a particular finite-dimensional subspace $\breve{\mathcal{H}}_K$ generated by the first E eigenfunctions ϕ_e and denoted by $\mathcal{H}^E_K$, i.e.

$$\mathcal{H}^E_K := \left\{ g\in\mathcal{L}^2\left(\mu\right) \ \text{ s.t. } \ g = \sum_{e=1}^{E} a_e\phi_e \ \text{ where } \ \left[a_1,\ldots,a_E\right]^T\in\mathbb{R}^E \right\} \ . \tag{2.36}$$

The particular choice for $\mathcal{H}_K^E$ is motivated by the presence of the penalty term $\|\cdot\|_K^2$ used to obtain the function estimate. It can be also justified using the Bayesian framework described in Section 2.1.4 under which, before seeing the data, $\mathcal{H}_K^E$ represents the subspace that captures the biggest part of the signal variance among all the subspaces of $\mathcal{H}_K$ of dimension E. This is in accordance with the Rayleigh's principle Zhu et al. (1998); Nef (1967) which underlies Principal Component Analysis.

Using then $\mathcal{H}_K^E$ as hypothesis space, the estimate of f_μ is given by[9]

$$\widehat{f}_{\mathrm{r}} := \arg \min_{f \in \mathcal{H}_K^E} Q(f) \ . \tag{2.37}$$

Exploiting this restriction, we now derive how to explicitly compute $\widehat{f}_{\mathrm{r}}$.

As it will be clear in the sequel, it is now convenient to reformulate the estimates $\widehat{f}_{\mathrm{c}}$ and $\widehat{f}_{\mathrm{r}}$ introduced in (2.37) through the map $T : \mathcal{H}_K \to \mathbb{R}^\infty$ that is induced by definition (2.7) and associates to a function $f = \sum_{e=1}^{+\infty} a_e \phi_e$ the sequence $[a_1, a_2 \ldots]$, i.e.

$$T[f] = [a_1, a_2, \ldots]^T \ . \tag{2.38}$$

With abuse of notation, we also equip $\mathbb{R}^\infty$ with the norm

$$\|\mathbf{a}\|_K^2 := \sum_{e=1}^{+\infty} \frac{a_e^2}{\lambda_e} \tag{2.39}$$

so as to make T an isometric mapping. In what follows, if $A \in \mathbb{R}^{\infty \times \infty}$ and $w \in \mathbb{R}^\infty$, Aw is the vector with i-th element equal to $\sum_{j=1}^{\infty} [A]_{ij} w_j$. In addition, A^{-1} denotes the inverse of the operator induced by A, i.e. we use notation of ordinary algebra to handle infinite-dimensional objects.

Exploiting $T[\cdot]$, the measurement model (2.15) can thus be rewritten as

$$y_i = C_i \mathbf{b} + \nu_i \qquad i = 1, \ldots, S \tag{2.40}$$

where $\mathbf{b} = T[f_\mu]$ and

$$C_i := [\phi_1(x_i), \phi_2(x_i), \ldots] \in \mathbb{R}^\infty \ . \tag{2.41}$$

Notice that C_i is a stochastic i.i.d. sequence whose distribution depends on μ. The following result is obtained, see also Pillonetto and Bell (2007).

Proposition 31. Let

$$Q(\mathbf{a}) := \sum_{i=1}^{S} (y_i - C_i \mathbf{a})^2 + \gamma \|\mathbf{a}\|_K^2 \ . \tag{2.42}$$

Then

$$\begin{aligned} \widehat{\mathbf{b}}_{\mathrm{c}} &:= \arg\min_{\mathbf{a}} Q(\mathbf{a}) \\ &= \left(\mathrm{diag}\left(\frac{\gamma}{\lambda_e}\right) + \sum_{i=1}^{S} C_i^T C_i \right)^{-1} \left(\sum_{i=1}^{S} C_i^T y_i \right) \end{aligned} \tag{2.43}$$

with $\mathrm{diag}(a_e)$ to indicate the matrix with diagonal elements given by $a_1, a_2, \ldots$. Furthermore, $T\left[\widehat{f}_{\mathrm{c}}\right] = \widehat{\mathbf{b}}_{\mathrm{c}}$, where $\widehat{f}_{\mathrm{c}}$ is defined in (2.16).

[9] We use the subscript r to recall that $\widehat{f}_{\mathrm{r}}$ lies in a reduced hypothesis space.

Since (2.43) involves infinite dimensional vectors, approximated solutions in $\mathcal{H}_K^E$ are now searched. To this aim, defining

$$C_i^E := C^E(x_i) := [\phi_1(x_i), \cdots, \phi_E(x_i), 0, 0, \ldots] \tag{2.44}$$

the following proposition is obtained.

Proposition 32. Let

$$Q^E(\mathbf{a}) := \sum_{i=1}^{S} \left(y_i - C_i^E \mathbf{a}\right)^2 + \gamma \|\mathbf{a}\|_K^2 \ . \tag{2.45}$$

Then

$$\begin{aligned} \widehat{\mathbf{b}}_{\mathrm{r}} &:= \arg\min_{\mathbf{a}} Q^E(\mathbf{a}) \\ &= \left(\mathrm{diag}\left(\frac{\gamma}{\lambda_e}\right) + \sum_{i=1}^{S} \left(C_i^E\right)^T C_i^E\right)^{-1} \left(\sum_{i=1}^{S} \left(C_i^E\right)^T y_i\right) \end{aligned} \tag{2.46}$$

and $T\left[\widehat{f}_{\mathrm{r}}\right] = \widehat{\mathbf{b}}_{\mathrm{r}}$ where $\widehat{f}_{\mathrm{r}}$ is defined in (2.37).

Notice that even if $\left(C_i^E\right)^T C_i^E$ is an infinite dimensional matrix, only its $E \times E$ upper-left block can contain non-zero elements. In the same way, every infinite dimensional vector $\left(C_i^E\right)^T y_i$ can have non-zero elements only in its first E components. This implies that also $\widehat{\mathbf{b}}_{\mathrm{r}}$ can have non-zero elements only in its first E components.

Alternative Interpretation of Estimator (2.46)

Under the Bayesian interpretation of Section 2.1.4, we can obtain an alternative interpretation of estimator (2.46). Rewriting measurement model (2.40) exploiting definition (2.44), we obtain the new model

$$y_i = C_i^E [b_1, \ldots, b_E]^T + w_i + \nu_i \qquad i = 1, \ldots, S \tag{2.47}$$

where[10]

$$w_i := \sum_{e=E+1}^{+\infty} [C_i]_e \, b_e \tag{2.48}$$

can be considered the *approximation noise* generated by the choice of considering only the first E eigenfunctions weights. In general these approximation noises are correlated, i.e.

$$\nexists\, \sigma_w^2 \in \mathbb{R}_+ \quad \text{s.t.} \quad \Sigma_{\mathbf{w}} := \mathbb{E}\left[[w_1, \ldots, w_S] \begin{bmatrix} w_1 \\ \vdots \\ w_S \end{bmatrix}\right] = \sigma_w^2 I \tag{2.49}$$

[10] The notational abuses we are committing by considering finite dimensional objects instead of infinite dimensional ones will not affect the following considerations.

thus the Bayesian estimator of the first E eigenfunctions weights is given by

$$\begin{aligned}&\mathbb{E}\left[b_1,\dots,b_E \mid x_1,y_1,\dots,x_S,y_S\right] = \\ &= \Lambda_0 \begin{bmatrix} C_1^E \\ \vdots \\ C_S^E \end{bmatrix}^T \left(\begin{bmatrix} C_1^E \\ \vdots \\ C_S^E \end{bmatrix} \Lambda_0 \begin{bmatrix} C_1^E \\ \vdots \\ C_S^E \end{bmatrix}^T + \Sigma_{\mathbf{w}} + \sigma^2 I \right)^{-1} \begin{bmatrix} y_1 \\ \vdots \\ y_S \end{bmatrix}\end{aligned} \tag{2.50}$$

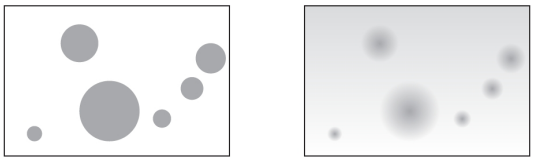
where

$$\Lambda_0 := \operatorname{diag}\left(b_1,\dots,b_E\right) \ . \tag{2.51}$$

It is immediate to check that estimator (2.50) does not coincide with estimator (2.46): the latter would be the Bayesian estimator of $b_1,\dots,b_E$ only if $\Sigma_{\mathbf{w}} = 0$. The interpretation of (2.46) can thus be the optimal estimator for measurements models where the approximation noises w_i are dismissed.

2.2 Distributed Regression

We assume now and in the following that each measurement y_i has been taken from a single[11] sensor: in particular, sensor i takes measurement y_i. We assume moreover that each input location x_i has been independently drawn from the known measure μ. Examples of probability measures μ are shown in Figure 2.12.



(a) (b)

Figure 2.12: Examples of probability measures μ on a compact $\mathcal{X} \subset \mathbb{R}^2$. Levels of gray indicate different values for the probability density, with white indicating zero. (a) x_i's can be located only around certain locations. (b) x_i's can be located everywhere, even if certain locations are preferred.

We then begin with two considerations about the non-distributability of strategies (2.18) and (2.46):

non-distributability of strategy (2.18)**:** $\widehat{f}_{\mathrm{c}}$ is composed by a linear combination of at most S different representing functions, which weights have to be computed through (2.20). This implies that, in the most general case, there has to be at least one unit collecting the whole dataset and performing the computations, thing that we want to avoid.

[11] The restriction of one measurement per sensor has been made only for ease of notation. The results presented in the following can be immediately extended to the case where sensors take multiple measurements.

non-distributability of strategy (2.46) the relation for the computation of $\widehat{\mathbf{b}}_{\mathrm{r}}$ can now be rewritten in a form suited to distributed estimation, i.e.

$$\widehat{\mathbf{b}}_{\mathrm{r}} = \left(\frac{1}{S}\mathrm{diag}\left(\frac{\gamma}{\lambda_e}\right) + \frac{1}{S}\sum_{i=1}^{S}\left(C_i^E\right)^T C_i^E\right)^{-1}\left(\frac{1}{S}\sum_{i=1}^{S}\left(C_i^E\right)^T y_i\right) . \tag{2.52}$$

This shows that, assuming each sensor knows S, $\widehat{\mathbf{b}}_{\mathrm{r}}$ can be distributedly computed through two parallel average consensi (one on $\left(C_i^E\right)^T C_i^E$ and one on $\left(C_i^E\right)^T y_i$), plus multiplications and inversions of $E \times E$ matrices and E-dimensional vectors, implying $O\left(E^2\right)$-communication and $O\left(E^3\right)$-computational costs.

We notice that practical implementation of (2.52) may still be problematic. In fact, not only the agents must know the exact number of measurements/sensors S (assumption that has been considered too strict also in Section 1.2), but also the amount of information that needs to be transmitted could be too elevated, since it scales with the square of E. In the following we will then propose an effective strategies that overcomes these problems and simultaneously and distributedly compute an approximation of both $\widehat{f}_{\mathrm{c}}$ and $\widehat{f}_{\mathrm{r}}$ (or, thanks to the equivalence relation induced by operator $T[\cdot]$, $\widehat{\mathbf{b}}_{\mathrm{c}}$ and $\widehat{\mathbf{b}}_{\mathrm{r}}$).

From a practical point of view, the situation will be like the one depicted in Figure 2.13. Due to the physical constraints imposed to the distributed estimation scenario, neither $\widehat{\mathbf{b}}_{\mathrm{c}}$ nor $\widehat{\mathbf{b}}_{\mathrm{r}}$ (that can be interpreted as the centralized estimate $\widehat{\mathbf{b}}_{\mathrm{c}}$ that would be obtained truncating the prior, i.e. setting $\lambda_e = 0$ for $e > E$) can be exactly computed. We then propose to estimate $\mathbf{b} = T[f_\mu]$ still by means of the hypothesis space $\mathcal{H}_K^E$: the objective will be then to bound the distance between this estimate $\widehat{\mathbf{b}}_{\mathrm{d}}$ and the optimal centralized estimator $\widehat{\mathbf{b}}_{\mathrm{c}}$.

Remark 33. Despite of the possible stochastic interpretation of the problem recalled in Section 2.1.4, in all this section f_μ always represents an unknown but deterministic function.

The approximation of $\widehat{\mathbf{b}}_{\mathrm{c}}$ (and $\widehat{\mathbf{b}}_{\mathrm{r}}$) we define is then the following:

$$\widehat{\mathbf{b}}_{\mathrm{d}} := \left(\frac{1}{S_{\mathrm{g}}}\mathrm{diag}\left(\frac{\gamma}{\lambda_e}\right) + I\right)^{-1}\left(\frac{1}{S}\sum_{i=1}^{S}\left(C_i^E\right)^T y_i\right) . \tag{2.53}$$

Notice that $\widehat{\mathbf{b}}_{\mathrm{d}}$ is an approximation of $\widehat{\mathbf{b}}_{\mathrm{r}}$ since

1. parameter S weighting the regularization term $\mathrm{diag}\left(\gamma/\lambda_e\right)$ is replaced with a guess (or estimate) S_{g};
2. $\frac{1}{S}\sum_{i=1}^{S}\left(C_i^E\right)^T C_i^E$ is replaced with $\mathbb{E}_\mu\left[\left(C_i^E\right)^T C_i^E\right]$. In fact,

$$\mathbb{E}_\mu\left[\left(C_i^E\right)^T C_i^E\right] = I \tag{2.54}$$

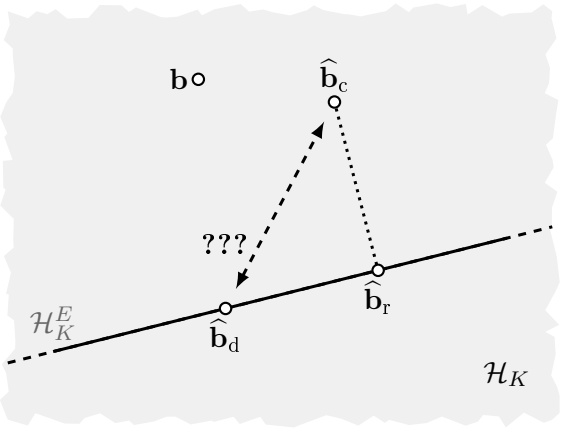


Figure 2.13: Summary of the problem discussed in this section. $\widehat{\mathbf{b}}_{\mathrm{c}}$ is the centralized optimal estimator of $\mathbf{b}$, while $\widehat{\mathbf{b}}_{\mathrm{r}}$ belongs to the reduced hypothesis space $\mathcal{H}_K^E$ and correspond to the optimal estimate that would be obtained truncating the prior by setting $\lambda_e = 0$ for $e > E$. Since both $\widehat{\mathbf{b}}_{\mathrm{c}}$ and $\widehat{\mathbf{b}}_{\mathrm{r}}$ cannot be distributedly compute, we seek strategies computing an estimate $\widehat{\mathbf{b}}_{\mathrm{d}}$ that is sufficiently close to $\widehat{\mathbf{b}}_{\mathrm{c}}$ (and not to $\mathbf{b}$, since this is unknown). We remark that in the context of Bayesian regression discussed in Section 2.1.4, in general $\mathbb{P}[\mathbf{b} \in \mathcal{H}_K] = 0$, but $\mathbb{P}[\mathbb{E}[\mathbf{b} \mid x_1, y_1, \ldots, x_S, y_S] \in \mathcal{H}_K] = 1$.

because one has

$$\left[\frac{1}{S}\sum_{i=1}^{S}\left(C_i^E\right)^T C_i^E\right]_{mn} = \frac{1}{S}\sum_{i=1}^{S}\phi_m(x_i)\,\phi_n(x_i) \tag{2.55}$$

and in addition

$$\frac{1}{S}\sum_{i=1}^{S}\phi_m(x_i)\,\phi_n(x_i) \xrightarrow{S\to+\infty} \int_{\mathcal{X}}\phi_i(x)\,\phi_j(x)\,d\mu(x) = \delta_{ij} \tag{2.56}$$

due to the orthogonality of eigenfunctions in $\mathcal{L}^2(\mu)$ and the fact that the x_i's are i.i.d and extracted from μ.

Computational and communication costs needed to compute $\widehat{\mathbf{b}}_{\mathrm{d}}$ are both $O(E)$, thus dramatically smaller than those associated with $\widehat{\mathbf{b}}_{\mathrm{c}}$ and $\widehat{\mathbf{b}}_{\mathrm{r}}$. Table 2.1 summarizes the computational, communication and memory costs associated with the introduced estimators.

estimator	*comput. cost*	*commun. cost*	*memory cost*
$\widehat{\mathbf{b}}_{\mathrm{c}}$ (Eqn. (2.43))	$O(S^3)$	$O(S)$	$O(S)$
$\widehat{\mathbf{b}}_{\mathrm{r}}$ (Eqn. (2.46))	$O(E^3)$	$O(E^2)$	$O(E^2)$
$\widehat{\mathbf{b}}_{\mathrm{d}}$ (Eqn. (2.53))	$O(E)$	$O(E)$	$O(E)$

Table 2.1: Summary of the computational, communication and memory costs per node associated to the introduced estimators.

2.2.1 On-line bounds computation

Once $\widehat{\mathbf{b}}_{\mathrm{d}}$ is obtained, it is crucial to assess its reliability in terms of closeness to the optimal centralized estimate $\widehat{\mathbf{b}}_{\mathrm{c}}$. We describe now how to simultaneously and distributedly bound the norm of the approximation error $\widehat{f}_{\mathrm{d}} - \widehat{f}_{\mathrm{c}}$ and $\widehat{f}_{\mathrm{d}} - \widehat{f}_{\mathrm{r}}$ with additive $O\left(E^3\right)$-computational and $O\left(1\right)$-communication costs. Under the working hypotheses, these distributedly computed bounds do not depend on the statistical modeling errors of f, and can be considered in some sense "universal". Notice that computation of approximation $\widehat{f}_{\mathrm{d}}$ can be effective and useful also in centralized estimation scenarios whenever the inversion of $S \times S$-dimensional matrices is impractical, and one wants to characterize the estimation performances losses. To this regard, the following two results, namely Algorithm 1 and the related Proposition 34, represent the main results of this section. They provide a way to compute, in a distributed fashion, statistical bounds for the relative errors

$$\left\|\widehat{\mathbf{b}}_{\mathrm{d}} - \widehat{\mathbf{b}}_{\mathrm{c}}\right\|_2 / \left\|\widehat{\mathbf{b}}_{\mathrm{d}}\right\|_2, \qquad \left\|\widehat{\mathbf{b}}_{\mathrm{d}} - \widehat{\mathbf{b}}_{\mathrm{r}}\right\|_2 / \left\|\widehat{\mathbf{b}}_{\mathrm{d}}\right\|_2$$

which, in view of (2.10) and letting $\widehat{f}_{\mathrm{d}} = T^{-1}\left[\widehat{\mathbf{b}}_{\mathrm{d}}\right]$, coincide respectively with

$$\left\|\widehat{f}_{\mathrm{d}} - \widehat{f}_{\mathrm{c}}\right\|_\mu / \left\|\widehat{f}_{\mathrm{d}}\right\|_\mu, \qquad \left\|\widehat{f}_{\mathrm{d}} - \widehat{f}_{\mathrm{r}}\right\|_\mu / \left\|\widehat{f}_{\mathrm{d}}\right\|_\mu .$$

In the sequel, let

$$C_i^{\setminus E} := \left[0, \ldots, 0, \phi_{E+1}\left(x_i\right), \phi_{E+2}\left(x_i\right), \ldots\right] . \tag{2.57}$$

Furthermore, to compact the notation, let

$$V_r := \left(\frac{1}{S}\mathrm{diag}\left(\frac{\gamma}{\lambda_e}\right) + \frac{1}{S}\sum_{i=1}^{S}\left(C_i^E\right)^T C_i^E\right)^{-1} \tag{2.58}$$

and

$$V_d := \left(\frac{1}{S_{\mathrm{g}}}\mathrm{diag}\left(\frac{\gamma}{\lambda_e}\right) + I\right)^{-1} . \tag{2.59}$$

Proposition 34. Consider Algorithm 1 and the definitions therein. Then, conditional on $\mathcal{Z}$ and r_{ave}, up to Monte Carlo approximations, it holds that

$$\mathbb{P}\left(\frac{\left\|\widehat{f}_{\mathrm{r}} - \widehat{f}_{\mathrm{d}}\right\|_\mu}{\left\|\widehat{f}_{\mathrm{d}}\right\|_\mu} \geq \overline{d}_{|dr|}\right) \leq \delta \tag{2.76}$$

$$\mathbb{P}\left(\frac{\left\|\widehat{f}_{\mathrm{c}} - \widehat{f}_{\mathrm{d}}\right\|_\mu}{\left\|\widehat{f}_{\mathrm{d}}\right\|_\mu} \geq \overline{d}_{|dc|}\right) \leq \delta . \tag{2.77}$$

Algorithm 1 Distributed estimation and approximation quality evaluation

Off-line work: Sensors are given a level of confidence $1-\delta$, e.g. $\delta = 0.1$, and know $S_{\min}$, $S_{\max}$, S_{g}, μ, E as well as the quantities

$$\gamma_a := \sup_{x\in\mathcal{X}} \left\| \operatorname{diag}\left(\frac{\lambda_e}{\gamma}\right) \left(C^{\setminus E}(x)\right)^T \right\|_2 \tag{2.60}$$

$$\gamma_b := \sup_{x\in\mathcal{X}} \left\| \operatorname{diag}\left(\frac{\lambda_e}{\gamma}\right) \left(C^{\setminus E}(x)\right)^T C^E(x) \right\|_2 \tag{2.61}$$

$$U_S^* := \left(\frac{1}{S_{\min}} - \frac{1}{S_{\max}}\right) \operatorname{diag}\left(\frac{\gamma}{\lambda_e}\right) . \tag{2.62}$$

In addition, each sensor i stores a particular scenario of the network, i.e. it locally generates $S_{\min}$ independent virtual input locations $x_{i,j}$ by means of density μ

$$x_{i,j} \sim \mu \quad \text{where} \quad j = 1, \ldots, S_{\min} \tag{2.63}$$

and then compute the following quantities

$$C_{i,j}^E := \left[\phi_1(x_{i,j}), \ldots, \phi_E(x_{i,j})\right] \tag{2.64}$$

$$V_{r,i}^* := \left(\frac{1}{S_{\max}} \operatorname{diag}\left(\frac{\gamma}{\lambda_e}\right) + \frac{1}{S_{\max}} \sum_{j=1}^{S_{\min}} \left(C_{i,j}^E\right)^T C_{i,j}^E\right)^{-1} \tag{2.65}$$

$$U_{C,i}^* := \left(I - \frac{1}{S_{\min}} \sum_{j=1}^{S_{\min}} \left(C_{i,j}^E\right)^T C_{i,j}^E\right) . \tag{2.66}$$

▷ continues in the next page

▷ continuation of Algorithm 1

On-line and distributed work:

1: (distributed step) sensors distributedly compute, by means of average consensus protocols, the E-dimensional vector

$$\mathcal{Z} := \frac{1}{S}\sum_{i=1}^{S}\left(C_i^E\right)^T y_i \tag{2.67}$$

2: (local step) each sensor i computes the estimate $\widehat{\mathbf{b}}_{\mathrm{d}} = V_d\mathcal{Z}$

3: (local step) each sensor i computes the auxiliary scalars

$$r_i := \frac{\left|y_i - C_i^E\widehat{\mathbf{b}}_{\mathrm{d}}\right|}{\left\|\widehat{\mathbf{b}}_{\mathrm{d}}\right\|_2} \tag{2.68}$$

$$d^*_{|dr|,i} := \frac{\left\|V^*_{r,i}U^*_S\widehat{\mathbf{b}}_{\mathrm{d}}\right\|_2}{\left\|\widehat{\mathbf{b}}_{\mathrm{d}}\right\|_2} + \frac{\left\|V^*_{r,i}U^*_{C,i}\widehat{\mathbf{b}}_{\mathrm{d}}\right\|_2}{\left\|\widehat{\mathbf{b}}_{\mathrm{d}}\right\|_2} \tag{2.69}$$

4: (distributed step) sensors distributedly compute, by means of average consensus protocols, the scalars

$$r_{\mathrm{ave}} := \frac{1}{S}\sum_{i=1}^{S} r_i \tag{2.70}$$

$$d^*_{|dr|,\mathrm{ave}} := \frac{1}{S}\sum_{i=1}^{S} d^*_{|dr|,i} \tag{2.71}$$

$$d^*_{|dr|,\mathrm{sq}} := \frac{1}{S}\sum_{i=1}^{S}\left(d^*_{|dr|,i}\right)^2 \tag{2.72}$$

5: (local step) each sensor i computes

$$d^*_{|dr|,\mathrm{var}} := d^*_{|dr|,\mathrm{sq}} - \left(d^*_{|dr|,\mathrm{ave}}\right)^2 \tag{2.73}$$

$$\overline{d}_{|dr|} := d^*_{|dr|,\mathrm{ave}} + \sqrt{\left(\frac{1}{\delta} - 1\right) d^*_{|dr|,\mathrm{var}}} \tag{2.74}$$

$$\overline{d}_{|dc|} := \gamma_a S_{\max} r_{\mathrm{ave}} + \left(1 + \gamma_b S_{\max}\right)\overline{d}_{|dr|}\ . \tag{2.75}$$

Proof. Notice that

$$\begin{aligned}\frac{\left\|\widehat{f}_{\mathrm{c}}-\widehat{f}_{\mathrm{d}}\right\|_{\mu}}{\left\|\widehat{f}_{\mathrm{d}}\right\|_{\mu}}=\frac{\left\|\widehat{\mathbf{b}}_{\mathrm{c}}-\widehat{\mathbf{b}}_{\mathrm{d}}\right\|_{2}}{\left\|\widehat{\mathbf{b}}_{\mathrm{d}}\right\|_{2}} &\leq \frac{\left\|\widehat{\mathbf{b}}_{\mathrm{c}}-\widehat{\mathbf{b}}_{\mathrm{r}}\right\|_{2}}{\left\|\widehat{\mathbf{b}}_{\mathrm{d}}\right\|_{2}}+\frac{\left\|\widehat{\mathbf{b}}_{\mathrm{r}}-\widehat{\mathbf{b}}_{\mathrm{d}}\right\|_{2}}{\left\|\widehat{\mathbf{b}}_{\mathrm{d}}\right\|_{2}}= \\ &= \frac{\left\|\widehat{\mathbf{b}}_{\mathrm{c}}-\widehat{\mathbf{b}}_{\mathrm{r}}\right\|_{2}}{\left\|\widehat{\mathbf{b}}_{\mathrm{d}}\right\|_{2}}+\frac{\left\|\widehat{f}_{\mathrm{r}}-\widehat{f}_{\mathrm{d}}\right\|_{2}}{\left\|\widehat{f}_{\mathrm{d}}\right\|_{2}}\end{aligned} \tag{2.78}$$

thus to prove (2.76) and (2.77) it is sufficient to characterize $\left\|\widehat{\mathbf{b}}_{\mathrm{r}}-\widehat{\mathbf{b}}_{\mathrm{d}}\right\|_{2}/\left\|\widehat{\mathbf{b}}_{\mathrm{d}}\right\|_{2}$ and $\left\|\widehat{\mathbf{b}}_{\mathrm{c}}-\widehat{\mathbf{b}}_{\mathrm{r}}\right\|_{2}/\left\|\widehat{\mathbf{b}}_{\mathrm{d}}\right\|_{2}$, that will be analyzed separately in the following.

Case $\left\|\widehat{\mathbf{b}}_{\mathbf{r}}-\widehat{\mathbf{b}}_{\mathbf{d}}\right\|_{2}/\left\|\widehat{\mathbf{b}}_{\mathbf{d}}\right\|_{2}$: we start rewriting (2.46) as

$$V_r^{-1}\widehat{\mathbf{b}}_{\mathrm{r}}=\mathcal{Z} \tag{2.79}$$

and (2.53) as

$$\left(V_r^{-1}+V_d^{-1}-V_r^{-1}\right)\widehat{\mathbf{b}}_{\mathrm{d}}=\mathcal{Z}\,. \tag{2.80}$$

Subtracting (2.80) to (2.79) we obtain

$$\widehat{\mathbf{b}}_{\mathrm{r}}-\widehat{\mathbf{b}}_{\mathrm{d}}=V_r\left(V_d^{-1}-V_r^{-1}\right)\widehat{\mathbf{b}}_{\mathrm{d}} \tag{2.81}$$

from which it immediately follows that

$$\frac{\left\|\widehat{\mathbf{b}}_{\mathrm{d}}-\widehat{\mathbf{b}}_{\mathrm{r}}\right\|_{2}}{\left\|\widehat{\mathbf{b}}_{\mathrm{d}}\right\|_{2}}=\frac{\left\|V_r\left(V_d^{-1}-V_r^{-1}\right)\widehat{\mathbf{b}}_{\mathrm{d}}\right\|_{2}}{\left\|\widehat{\mathbf{b}}_{\mathrm{d}}\right\|_{2}}\,. \tag{2.82}$$

Defining then U_C and U_S by means of (2.101) and (2.102), it is immediate to check that

$$V_d^{-1}-V_r^{-1}=U_S+U_C \tag{2.83}$$

and thus that

$$\frac{\left\|\widehat{\mathbf{b}}_{\mathrm{d}}-\widehat{\mathbf{b}}_{\mathrm{r}}\right\|_{2}}{\left\|\widehat{\mathbf{b}}_{\mathrm{d}}\right\|_{2}}=\frac{\left\|V_rU_S\widehat{\mathbf{b}}_{\mathrm{d}}+V_rU_C\widehat{\mathbf{b}}_{\mathrm{d}}\right\|_{2}}{\left\|\widehat{\mathbf{b}}_{\mathrm{d}}\right\|_{2}} \tag{2.84}$$

from which inequality (2.103) immediately follows.

Defining then

$$d_{|dr|}:=\frac{\left\|V_rU_S\widehat{\mathbf{b}}_{\mathrm{d}}\right\|_{2}}{\left\|\widehat{\mathbf{b}}_{\mathrm{d}}\right\|_{2}}+\frac{\left\|V_rU_C\widehat{\mathbf{b}}_{\mathrm{d}}\right\|_{2}}{\left\|\widehat{\mathbf{b}}_{\mathrm{d}}\right\|_{2}} \tag{2.85}$$

we notice that $d_{|dr|}$ is a random variable since V_r and U_C are random operators. It is then clear that to characterize $d_{|dr|}$ corresponds to characterize the relative error between $\widehat{\mathbf{b}}_{\mathrm{r}}$ and $\widehat{\mathbf{b}}_{\mathrm{d}}$. Notice that it is possible to prove that the probability density

of the random variable $d_{|dr|}$ conditioned on $\widehat{\mathbf{b}}_{\mathrm{d}}$ has compact support. In fact it is immediate to check that

$$0 \geq V_r^{-1} \geq \frac{1}{S_{\max}} \mathrm{diag}\left(\frac{\gamma}{\lambda_e}\right) \quad , \qquad I \geq U_C \; . \tag{2.86}$$

Moreover, the rank of $\left(C_i^E\right)^T C_i^E$ is one, thus if $\rho\left(A\right)$ indicates the spectral radius of A it follows that

$$\rho\left(\left(C_i^E\right)^T C_i^E\right) = \rho\left(C_i^E \left(C_i^E\right)^T\right) = \left\|C_i^E\right\|_2^2 \; . \tag{2.87}$$

Exploiting now the continuity of eigenfunctions on the compact $\mathcal{X}$ we have that

$$\left\|C_i^E\right\|_2^2 \leq E \cdot \sup_{x \in \mathcal{X},\, e=1,\ldots,E} \left|\phi_e\left(x\right)\right|^2 =: \gamma_c < +\infty \tag{2.88}$$

thus

$$U_C \geq -\frac{1}{S} \sum_{i=1}^{S} \left(C_i^E\right)^T C_i^E \geq -\gamma_c I \; . \tag{2.89}$$

From (2.85), (2.86) and (2.89) it then follows that

$$0 \leq d_{|dr|} \leq \left\| S_{\max} \mathrm{diag}\left(\frac{\lambda_e}{\gamma}\right) U_S \widehat{\mathbf{b}}_{\mathrm{d}} \right\|_2 + \left\| S_{\max} \mathrm{diag}\left(\frac{\lambda_e}{\gamma}\right) \max\left(1, \sqrt{\gamma_c}\right) \widehat{\mathbf{b}}_{\mathrm{d}} \right\|_2 \tag{2.90}$$

and thus we can claim that the support of the density of $d_{|dr|}$ is compact.

From (2.90) it follows that $\mathrm{var}\left(d_{|dr|}\right) < +\infty$, and this allows us to use Cantelli's inequality, i.e. to state that

$$\mathbb{P}\left[d_{|dr|} - \mathbb{E}\left[d_{|dr|}\right] \geq \sqrt{\left(\frac{1}{\delta} - 1\right) \mathrm{var}\left(d_{|dr|}\right)} \right] \leq \delta \; . \tag{2.91}$$

Notice now that the knowledge of $\mathcal{Z}$ does not provide a-posteriori information on μ, while knowledge of residuals r_i provide little information[12]: this imply that samples $d^*_{|dr|,i}$ generated in step 3 of Algorithm 1 are approximatively generated from the same density of $d_{|dr|}$. For this reason we can claim that

$$d^*_{|dr|,\mathrm{ave}} \approx \mathbb{E}\left[d_{|dr|}\right] \qquad \text{and} \qquad d^*_{|dr|,\mathrm{var}} \approx \mathrm{var}\left(d_{|dr|}\right) \tag{2.92}$$

that eventually proves (2.76).

- **Case $\left\|\widehat{\mathbf{b}}_{\mathbf{c}} - \widehat{\mathbf{b}}_{\mathbf{r}}\right\|_2 / \left\|\widehat{\mathbf{b}}_{\mathbf{d}}\right\|_2$:** rewriting (2.46) as

$$\begin{aligned} \left(\mathrm{diag}\left(\frac{\gamma}{\lambda_e}\right) + \sum_{i=1}^{S} C_i^T C_i\right) \widehat{\mathbf{b}}_{\mathrm{r}} + \left(\sum_{i=1}^{S} \left(C_i^E\right)^T C_i^E - \sum_{i=1}^{S} C_i^T C_i\right) \widehat{\mathbf{b}}_{\mathrm{r}} = \\ = \sum_{i=1}^{S} C_i^T y_i - \sum_{i=1}^{S} \left(C_i^{\setminus E}\right)^T y_i \end{aligned} \tag{2.93}$$

[12] It can be proven that if $\frac{\lambda_1}{\sigma^2 + \lambda_1 S_{\max}} \leq \sum_{i=1}^{S} \left\|C_i^E\right\|_2$ then the various r_i do not provide information on μ.

and (2.43) as

$$\left(\operatorname{diag}\left(\frac{\gamma}{\lambda_e}\right)+\sum_{i=1}^{S}C_i^TC_i\right)\widehat{\mathbf{b}}_{\mathrm{c}}=\sum_{i=1}^{S}C_i^Ty_i \tag{2.94}$$

after subtracting (2.94) to (2.93), we obtain

$$\begin{aligned}&\left(\operatorname{diag}\left(\frac{\gamma}{\lambda_e}\right)+\sum_{i=1}^{S}C_i^TC_i\right)\left(\widehat{\mathbf{b}}_{\mathrm{c}}-\widehat{\mathbf{b}}_{\mathrm{r}}\right)=\\&=\left(\sum_{i=1}^{S}\left(C_i^E\right)^TC_i^E-\sum_{i=1}^{S}C_i^TC_i\right)\widehat{\mathbf{b}}_{\mathrm{r}}+\sum_{i=1}^{S}\left(C_i^{\setminus E}\right)^Ty_i\,.\end{aligned} \tag{2.95}$$

Substituting now each C_i in the right side of (2.95) with $C_i^E+C_i^{\setminus E}$, exploiting the fact that $C_i^{\setminus E}\widehat{\mathbf{b}}_{\mathrm{r}}=0$ (where 0 is in $\mathbb{R}^\infty$), and properly collecting the various terms, we obtain

$$\widehat{\mathbf{b}}_{\mathrm{c}}-\widehat{\mathbf{b}}_{\mathrm{r}}=\left(\operatorname{diag}\left(\frac{\gamma}{\lambda_e}\right)+\sum_{i=1}^{S}C_i^TC_i\right)^{-1}\sum_{i=1}^{S}\left(C_i^{\setminus E}\right)^T\left(y_i-C_i\widehat{\mathbf{b}}_{\mathrm{r}}\right)\,. \tag{2.96}$$

Since $\operatorname{diag}\left(\frac{\gamma}{\lambda_e}\right)+\sum_{i=1}^{S}C_i^TC_i\geq\operatorname{diag}\left(\frac{\gamma}{\lambda_e}\right)$ (in a matricial positive definite sense), we obtain

$$\left\|\widehat{\mathbf{b}}_{\mathrm{c}}-\widehat{\mathbf{b}}_{\mathrm{r}}\right\|_2\leq\sum_{i=1}^{S}\left\|\operatorname{diag}\left(\frac{\lambda_e}{\gamma}\right)\left(C_i^{\setminus E}\right)^T\left(y_i-C_i\widehat{\mathbf{b}}_{\mathrm{r}}\right)\right\|_2\,. \tag{2.97}$$

Rewriting $y_i-C_i\widehat{\mathbf{b}}_{\mathrm{r}}$ as $y_i-C_i^E\widehat{\mathbf{b}}_{\mathrm{d}}+C_i^E\widehat{\mathbf{b}}_{\mathrm{d}}-C_i^E\widehat{\mathbf{b}}_{\mathrm{r}}$ and using definitions (2.60), (2.61), (2.68), (2.70) and (2.84) it follows immediately that

$$\begin{aligned}\frac{\left\|\widehat{\mathbf{b}}_{\mathrm{c}}-\widehat{\mathbf{b}}_{\mathrm{r}}\right\|_2}{\left\|\widehat{\mathbf{b}}_{\mathrm{d}}\right\|_2}&\leq\gamma_a\sum_{i=1}^{S}\frac{\left\|y_i-C_i\widehat{\mathbf{b}}_{\mathrm{d}}\right\|_2}{\left\|\widehat{\mathbf{b}}_{\mathrm{d}}\right\|_2}+\gamma_b\sum_{i=1}^{S}\frac{\left\|\widehat{\mathbf{b}}_{\mathrm{r}}-\widehat{\mathbf{b}}_{\mathrm{d}}\right\|_2}{\left\|\widehat{\mathbf{b}}_{\mathrm{d}}\right\|_2}\\&\leq\gamma_aS_{\max}r_{\mathrm{ave}}+\gamma_bS_{\max}\frac{\left\|\widehat{\mathbf{b}}_{\mathrm{r}}-\widehat{\mathbf{b}}_{\mathrm{d}}\right\|_2}{\left\|\widehat{\mathbf{b}}_{\mathrm{d}}\right\|_2}\,.\end{aligned} \tag{2.98}$$

Notice that γ_a is finite since, for every $x\in\mathcal{X}$ it holds that

$$\left\|\operatorname{diag}\left(\frac{\lambda_e}{\gamma}\right)C^{\setminus E}(x)\right\|_2^2\leq\sup_{x\in\mathcal{X},e\in\mathbb{N}_+}\phi_e(x)\cdot\sum_{e=E+1}^{+\infty}\frac{\lambda_e}{\gamma} \tag{2.99}$$

with $\sup_{x\in\mathcal{X},e\in\mathbb{N}_+}\phi_e(x)<+\infty$ because eigenfunctions are continuous on a compact, and also with $\sum_{e=E+1}^{+\infty}\frac{\lambda_e}{\gamma}<+\infty$ since K is Mercer. In the same way it is possible to show that also γ_b is finite.

Recalling now (2.78), it immediately follows that

$$\frac{\left\|\widehat{\mathbf{b}}_{\mathrm{c}}-\widehat{\mathbf{b}}_{\mathrm{d}}\right\|_2}{\left\|\widehat{\mathbf{b}}_{\mathrm{d}}\right\|_2}\leq\gamma_aS_{\max}r_{\mathrm{ave}}+(1+\gamma_bS_{\max})\frac{\left\|\widehat{\mathbf{b}}_{\mathrm{r}}-\widehat{\mathbf{b}}_{\mathrm{d}}\right\|_2}{\left\|\widehat{\mathbf{b}}_{\mathrm{d}}\right\|_2} \tag{2.100}$$

and thus that if (2.76) holds, then also (2.77) does. □

The proof of Proposition 34 is reported in Appendix and clarifies the quantities that influence the approximation error. In particular, here we just recall that defining

$$U_C := I - \frac{1}{S}\sum_{i=1}^{S}\left(C_i^E\right)^T C_i^E \tag{2.101}$$

and

$$U_S := \left(\frac{1}{S_{\mathrm{g}}} - \frac{1}{S}\right)\mathrm{diag}\left(\frac{\gamma}{\lambda_e}\right) \tag{2.102}$$

it is shown that

$$\left\|\widehat{f}_{\mathrm{d}} - \widehat{f}_{\mathrm{r}}\right\|_{\mu} \leq \left\|V_r U_S \widehat{\mathbf{b}}_{\mathrm{d}}\right\|_2 + \left\|V_r U_C \widehat{\mathbf{b}}_{\mathrm{d}}\right\|_2 \tag{2.103}$$

i.e. the error between $\widehat{f}_{\mathrm{d}}$ and $\widehat{f}_{\mathrm{r}}$ decomposes into two distinct parts, one involving U_S, proportional to the uncertainty on the number of sensors, and one involving U_C[13], related to the uncertainty on the actual input locations x_i.

Moreover, for what regards the distance of $\widehat{f}_{\mathrm{d}}$ from the optimal estimate $\widehat{f}_{\mathrm{c}}$, it holds that

$$\left\|\widehat{f}_{\mathrm{d}} - \widehat{f}_{\mathrm{c}}\right\|_{\mu} \leq (\gamma_b S_{\max} + 1)\left\|\widehat{\mathbf{b}}_{\mathrm{d}} - \widehat{\mathbf{b}}_{\mathrm{r}}\right\|_2 + \sum_{i=1}^{S}\gamma_a\left\|y_i - C_i^E\widehat{\mathbf{b}}_{\mathrm{d}}\right\|_2 \tag{2.104}$$

i.e. the error between $\widehat{f}_{\mathrm{d}}$ and $\widehat{f}_{\mathrm{c}}$ contains the two components described in (2.103) (scaled by a multiplicative factor always greater than one) plus a term dependent on the sum of the residuals, that accounts for the approximation error deriving from replacing $\mathcal{H}_K$ with $\breve{\mathcal{H}}_K$.

For what regards possible extensions of the algorithm described above, first notice that sensors aiming for more precision on the bounds may locally generate several instances of $d^*_{|dr|,i}$ and then estimate the bounds with the desired level of accuracy. In addition, we also remark that the assumptions on the independence of the various x_i's can be relaxed. In particular, Algorithm 1 and Proposition 34 can be easily extended to handle the case of sensors moving according to an ergodic Markov chain (e.g. generated by the Metropolis-Hastings algorithm Gilks et al. (1996)) having as invariant measure the desired distribution μ. Finally, it is worth stressing that the entire numerical procedure here described can be also easily modified to permit the regularization parameters present in (2.43), (2.46) and (2.53) to be different. For example, assume that the value of γ entering the definition of $\widehat{\mathbf{b}}_{\mathrm{c}}$ and $\widehat{\mathbf{b}}_{\mathrm{r}}$ is fixed. Then, one could estimate the value of the regularization parameter defining $\widehat{\mathbf{b}}_{\mathrm{d}}$ making it vary on a grid common to all the sensors and determining the "optimal" one as that minimizing the distance bounds presented above.

Remark 35. The assumptions on the independence of the various x_i's can be relaxed. In particular, with minor modifications, it is possible to handle the case where the sensors move according to an ergodic Markov chain, e.g. generated by the Metropolis-Hastings scheme, having as invariant measure the desired distribution μ.

[13] For an interesting bound on the norm of matrices of the type U_C, as a function of the number of sensors S and of the dimension E, the reader is also referred to Lemma 1 in Oliveira (2010).

We also stress that it is also possible to find a priori bounds (i.e. independent of $\mathcal{Z}$ and r_{ave}) for the relative error between $\widehat{f}_{\text{d}}$ and $\widehat{f}_{\text{c}}$. Unfortunately, our experience suggests that these bounds tend to be pessimistic and much less useful for practical purposes than those derived above.

It is possible also to construct a distributed strategy for the estimation of the relative error between $\widehat{\mathbf{b}}_{\text{d}}$ and $\widehat{\mathbf{b}}_{\text{r}}$ that does not rely on the knowledge of $\mathcal{Z}$ and residuals r_i. This is based on the fact that it can be shown that there exist a finite $\gamma_d \in \mathbb{R}_+$ s.t.:

$$d_{|dr|} \leq \left\| S_{\max} \text{diag}\left(\frac{\lambda_e}{\gamma}\right)(U_S + \gamma_d I) \right\|_2 . \tag{2.105}$$

Algorithm 1 could then be modified in order to generate and consider instances of U_S in order to compute the bound. Unfortunately also in this case the results tend to be pessimistic and unuseful for practical purposes.

It is worth stressing that in the deterministic scenario we used, the bounds that we obtained, regarding the performance of the proposed estimator, are robust since are not affected by errors in the statistical modeling of f_μ. This is in accordance with the modern statistical learning theory as described e.g. in Vapnik (1995). An example of this has been already provided in Proposition 30 where the validity of (2.27) just requires f_μ to belong to an infinite-dimensional space that may contain a very wide class of functions. For instance, popular choices for $\mathcal{H}_K$ are Sobolev spaces or spaces induced by the Gaussian kernel which are all known to be dense in the space of continuous functions, e.g. see Micchelli et al. (2006).

2.2.2 Simulations

We consider $f_\mu : \mathcal{X} = [0,1] \to \mathbb{R}$ to be given by

$$f_\mu(x) = \sum_{n=1}^{100} \alpha_n \sin(\omega_n x) \tag{2.106}$$

with $\alpha_n \sim \mathcal{N}(0, 0.01)$ i.i.d., $\omega_n \sim \mathcal{U}[0,25]$ i.i.d., μ to be uniform on $[0,1]$ and a measurement noise standard deviation $\sigma = 0.33$ such that, in average, SNR $:= \frac{\text{var}(f_\mu)}{\sigma^2} \approx 5$. Moreover we consider the Gaussian kernel

$$K(x, x') = \exp\left(-\frac{(x-x')^2}{0.02}\right) \tag{2.107}$$

associated to the meaningful but non-optimized regularization parameter $\gamma = 0.3$.

To show the effectiveness of the proposed algorithm, we independently generate 500 different f_μ (say $f_{\mu,j}$ with $j = 1, \ldots, 500$) sampled by $S = 1000$ sensors and estimated using $E = 40$ eigenfunctions. For each Monte Carlo run we apply Algorithm 1 and obtain $d^*_{|dr|,\text{ave},j}$ and $d^*_{|dr|,\text{var},j}$, the indexed versions of the quantities computed in (2.71) and (2.73). Then we compute the following two versions of bound (2.74):

$$\overline{d}_{|dr|,j}(0) := d^*_{|dr|,\text{ave},j} \tag{2.108}$$

$$\overline{d}_{|dr|,j}(3) := d^*_{|dr|,\text{ave},j} + 3\sqrt{d^*_{|dr|,\text{var},j}} \tag{2.109}$$

from which we compute the two versions $\overline{d}_{|dc|,j}(0)$ and $\overline{d}_{|dc|,j}(3)$ of bound (2.75). Notice that $\overline{d}_{|dc|,j}(3)$ corresponds to a confidence level of at least 0.9, while $\overline{d}_{|dc|,j}(0)$ corresponds to use only the means $d^*_{|dr|,\text{ave},j}$. Figures 2.14 and 2.15 then plot the points $\left(\frac{\|f_{d,j}-f_{c,j}\|_\mu}{\|f_{d,j}\|_\mu}, \overline{d}_{|dc|,j}(k)\right)$, $j = 1,\ldots,500$, $k = 0,3$, where $f_{c,j}$ and $f_{d,j}$ are the centralized and distributed estimates of $f_{\mu,j}$. In Figure 2.14 the uncertainty on S is given by $S_{\max} = 1100$ and $S_{\min} = 900$, while in Figure 2.15 it is given by $S_{\max} = 1200$ and $S_{\min} = 800$.

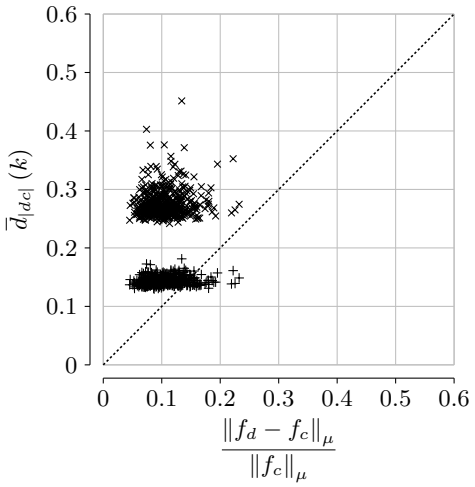


Figure 2.14: Scatter plot of the points $\left(\frac{\|f_{d,j}-f_{c,j}\|_\mu}{\|f_{d,j}\|_\mu}, \overline{d}_{|dc|,j}(k)\right)$ for $k = 3$ (black crosses) and $k = 0$ (black plus-symbols), with $S_{\text{g}} = S_{\max} = 1100$, $S_{\min} = 900$, $S = 1000$ and $E = 40$.

Since points corresponding to $k = 3$ are near the bisector of the first quadrant (black dashed line), bound (2.75) is significative even if we choose an high level of confidence. Notice that its conservativeness, graphically given by the distance of the points with the bisector, is inherited by the requested high level of confidence. Notice also that points corresponding to $k = 0$ are close to the bisector, thus it follows that the quantities $d^*_{|dr|,\text{ave},j}$ are informative with respect to the true error.

In Figure 2.16 we focus on the first Monte Carlo run ($j = 1$) and graphically show the effectiveness of the estimation strategy (2.53) through plotting the true $f_{\mu,1}$ and its relative estimates $f_{c,1}$ and $f_{d,1}$ ($S_{\text{g}} = 1200$). We then show in Figure 2.17, considering again the first Monte Carlo run $f_{\mu,1}$, the qualitative dependence of $\overline{d}_{|dc|,1}(3)$ on S and E. We notice that, as expected, the tightness of the bound generally increases with S and E. Finally we show in Figure 2.18 the dependency of the quality of the estimates $f_{d,j}$ ($j = 1,\ldots,500$) with respect to the accuracy of the guess S_{g}. Notice that by construction estimates $f_{c,j}$ do not depend on S_{g}. Since the actual regularization parameter is inversely proportional to S_{g}, Figure 2.18 can be considered the so-called regularization path and shows the robustness of the proposed estimator

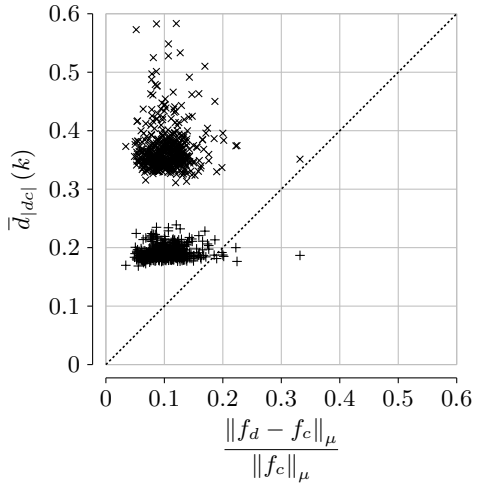


Figure 2.15: Scatter plot of the points $\left(\frac{\|f_{d,j}-f_{c,j}\|_\mu}{\|f_{d,j}\|_\mu}, \overline{d}_{|dc|,j}(k)\right)$ for $k = 3$ (black crosses) and $k = 0$ (black plus-symbols), with $S_{\mathrm{g}} = S_{\max} = 1200$, $S_{\min} = 800$, $S = 1000$ and $E = 40$.

with respect to accuracy on S_{g}. From a practical point of view, sensible performances worsenings can be obtained only with big variations of S_{g}.

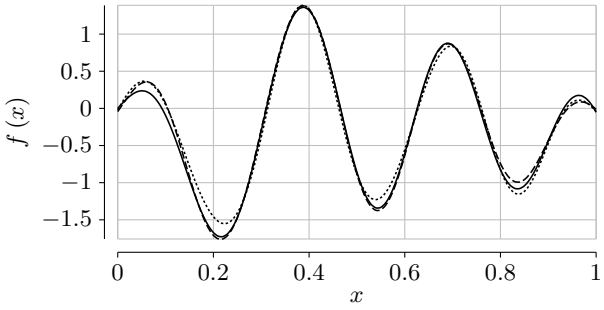


Figure 2.16: Results of the estimation procedure applied to the noisily sampled $f_{\mu,1}$ function in black solid line, with $f_{c,1}$ in black dashed line, $f_{d,1}$ in black dotted line, $S_{\mathrm{g}} = 1200$, $S = 1000$ and $E = 40$.

Remark 36. In some cases E could be sufficiently small to allow sensors to perform an average consensus also on matrices $\left(C_i^E\right)^T C_i^E$. In this case, approximation 2 described in page 58 will not be implemented, and equations presented in this section have to be modified accordingly to the new situation.

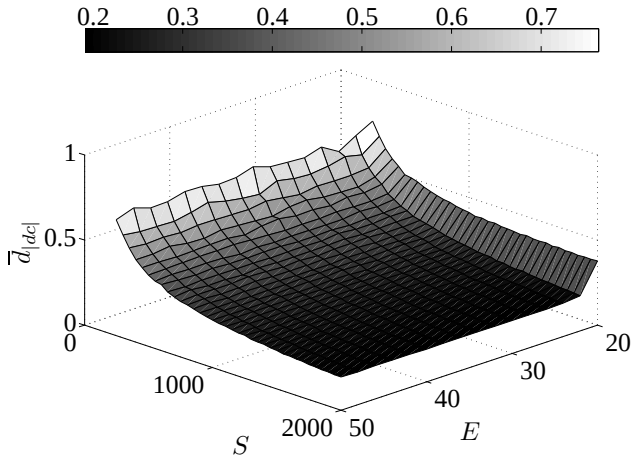


Figure 2.17: Dependence of $\overline{d}_{|dc|,1}$ (3) on S and E. $S_{\mathrm{g}} = S_{\max} = 1.1 \cdot S$, $S_{\min} = 0.9 \cdot S$.

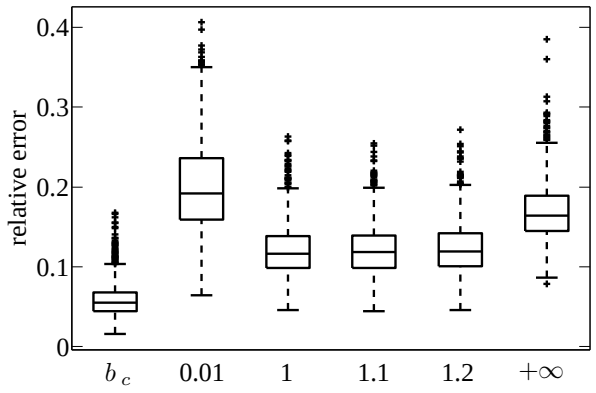


Figure 2.18: Boxplots relative to the relative errors $\frac{\|f_{\mu,j} - f_{c,j}\|_\mu}{\|f_{\mu,j}\|_\mu}$ (leftmost boxplot) and $\frac{\|f_{\mu,j} - f_{d,j}\|_\mu}{\|f_{\mu,j}\|_\mu}$ (the other boxplots) for different ratios S_{g}/S, with $S = 1000$ and $E = 40$.

2.3 Distributed Regression under Unknown Time Delays

2.3.1 Problem formulation

Assume there are S different synchronized sensors that noisily sample S differently shifted versions of the same signal $f_\mu : \mathbb{R} \to \mathbb{R}$, i.e.

$$y_i^m = f_i\left(x_i^m\right) + \nu_i^m \tag{2.110}$$

where

$$f_i(x) := f_\mu(x - d_i) \tag{2.111}$$

and where $i = 1, \ldots, S$ is the index of the sensor and $m = 1, \ldots, M_i$ is the index of the measurement. We define $\mathcal{S}_i := \{(x_i^m, y_i^m)\}$ to be the data set of sensor i, and to be composed of M_i non-uniformly sampled measurements.

Shifts $\{d_i\}$ are uncorrelated random variables, and only poorly informative prior on them is specified. *Time Delay Estimation* between signals $f_i(\cdot)$ and $f_j(\cdot)$ is then the problem of estimating the difference $d_{i,j} := d_i - d_j$ using data sets $\mathcal{S}_i$ and $\mathcal{S}_j$. Given S differently shifted signals, *multiple*-TDE is the attempt to solve simultaneously TDE problem for each couple f_i, f_j. In this section we want to *simultaneously and distributely estimate* f and delays $d_{i,j}$ using the measurements available to each agent.

Notice that we will continue assuming the additive noises ν_i^m to be independent, zero-mean Gaussian, with fixed and equal variance σ^2 and independent also of f. We also assume that sensors are subject to a communication graph, have high computation capabilities and can reliably communicate data.

2.3.2 Regression under Fixed Time Delays

Consider the situation of Figure 2.19.

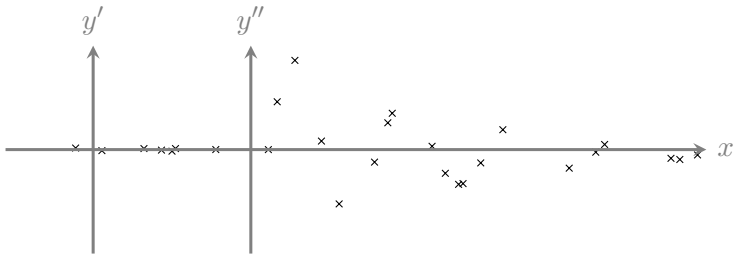


Figure 2.19: A certain set of measurements (black crosses) can be referred to different reference systems. It could be expedient to shift the original reference system in order to achieve better estimations.

Intuitively, in this case the regression using the eigenfunctions of Figure 2.8 with origin in x' will lead to poorer estimates than regression on the same data set using the same eigenfunctions with origin in x''. In fact poorly informative measurements

(i.e. measurements due only to noise) should be discarded, and it would be better to concentrate the approximation capability of $\mathcal{H}_K^E$ where the signal to noise ratio is significant. In order to eliminate such measurements we can define the *shifted data set*:

$$\mathcal{S}_i^\alpha := \{(x_i^m - \alpha,\ y_i^m)\} \tag{2.112}$$

where $\alpha \in \mathbb{R}$ is the time shift, and then compute the estimate of the unknown signal f_i by applying Algorithm 2.

Algorithm 2 Regression using translated reference systems

1: let α be a translation to be applied to $\mathcal{S}_i$
2: compute the shifted data set $\mathcal{S}_i^\alpha$ using Equation (2.112)
3: estimate f_μ through the methods described in Section 2.1.5 using the dataset $\mathcal{S}_i^\alpha$, and thus obtain an estimate $\widehat{f}_i'(x') \in \operatorname{span}\langle\phi_e(x')\rangle_{e=1}^E$ where eigenfunctions ϕ_e are referred to the translated reference system $x' = x - \alpha$
4: shift back the reference system obtaining $\widehat{f}_i(x) \in \operatorname{span}\langle\phi_e(x-\alpha)\rangle_{e=1}^E$.

Assume now α to be fixed. The estimate of the unknown function $\widehat{f}_i$ for sensor i can be written as

$$\widehat{f}_i(x) = \sum_{e=1}^{E} a_e \phi_e(x - \alpha)\ . \tag{2.113}$$

Assume that sensors i and j share the knowledge of functions $\{\phi_e\}$. If sensor j receive noiseless information about $[a_1, \ldots, a_E, \alpha]$, then it can exactly reconstruct $\widehat{f}_i(x)$.

This data shifting mechanism (that does not change the approximating properties of the regression method) constitutes the core of the regression algorithm developed in this section. Informally speaking, the time translation α corresponds to the origin of the eigenfunctions ϕ_e used to estimate f_μ, i.e. if $\alpha = 4$ then eigenfunctions will "start" at $x = 4$. It is important to remark that the eigenfunctions' weights $[a_1, \ldots, a_E]$ computed using Algorithm 2 are in general *different* from the weights computed without translating reference systems. For this reason we will indicate them using the notation $\widehat{\mathbf{a}}_i(\alpha)$, where it is highlighted both the dependance on α (the generic translation applied to the data set) and the sensor index i.

As mentioned above, the time shift α should be chosen in order to concentrate the regression only around points with a high signal to noise ratio. A natural choice is to define the *arrival time* as follows:

$$\alpha_i^0 := \min_m \{x_i^m \in \mathcal{S}_i \ \text{ s.t. } \ |y_i^m| \geq y_{\min}\} \tag{2.114}$$

where threshold $y_{\min}$ is a design parameter to be chosen based on the measurement noise variance σ^2.

2.3.3 Classic Time Delay Estimation

Consider two given signals f_i and f_j satisfying relation (2.111). A plausible estimation strategy for the relative time delay $d_{i,j} := d_i - d_j$ is to maximize the cross

correlation function[14] $\psi(\tau)$, i.e. to solve

$$\widehat{d}_{i,j} := \arg\max_{\tau} \ \psi(\tau) \tag{2.115}$$

where

$$\psi(\tau) := \int_{\mathcal{X}} f_i(x) f_j(x-\tau)\, dx \ . \tag{2.116}$$

Notice that when dealing with versions of the signals that are sampled with a fixed and equal sampling period T, then Equation (2.116) reduces to

$$\psi(\tau) := \sum_{k} f_i(x+kT) f_j(x+(k-\tau)T) \ . \tag{2.117}$$

with $\tau \in \mathbb{I}$. In continuous time, the resolution of the function $\psi(\tau)$ can be set arbitrarily small at the price of higher computational cost. In discrete time, the resolution of the relative delay τ is limited to multiples of the sampling period T, i.e. $\tau = \ell T$ where ℓ is an integer. To obtain a finer resolution, a it is possible to interpolate $\psi(\tau)$ between samples.

We send the readers interested in practical algorithms for the solution of (2.115) back to the specialized literature (for example Jacovitti and Scarano (1993), Boucher and Hassab (1981), Viola and Walker (2005)).

2.3.4 Time Delay Estimation in RKHSs

Equations (2.115) and (2.116) correspond to minimize an inner product in $\mathcal{L}^2$. This concept can be transferred into our RKHS framework by minimizing the inner product in $\mathcal{H}_K^E$ instead of in $\mathcal{L}^2$. Assuming then that sensor i owns its estimate $\widehat{f}_i$ of f_μ, and sensor j owns its estimate $\widehat{f}_j$, they can estimate their time delay by means of

$$\widehat{d}_{i,j} := \arg\min_{\tau} \ \langle \widehat{f}_i(x), \widehat{f}_j(x-\tau) \rangle_{\mathcal{H}_K^E} \ . \tag{2.118}$$

Given a fixed τ, computation of $\langle \widehat{f}_i(x), \widehat{f}_j(x-\tau) \rangle_{\mathcal{H}_K^E}$ through (2.9) can be performed only if

$$\widehat{f}_i(x), \ \widehat{f}_j(x-\tau) \in \operatorname{span} \langle \phi_e(x-\alpha) \rangle_{e=1}^{E} \tag{2.119}$$

i.e. only if both the estimated signals have been computed using eigenfunctions *with the same origin* α. If this is assured, then inner products can be computed using a finite number of operations: the problem is that to solve the minimization problem (2.118) it is required to solve the regression problem for each different τ (since for each different τ there is a differently shifted dataset). The computational complexity of this TDE strategy is then $O\left(\#\tau \cdot \left(E^3 + E^2 M_i + E M_i^2\right)\right)$ where $\#\tau$ indicates how many different τ's are computed. We remark that classical TDE has a computational complexity that depends on the precision used to solve Equation (2.116).

[14] It is also possible to use opportunely filtered versions of the signals in order to suppress the frequencies with low signal-to-noise ratio, see Azaria and Hertz (1984).

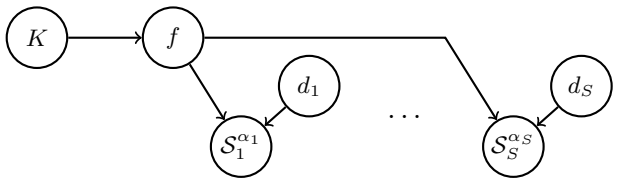


Figure 2.20: Bayesian network describing the relationships between the various random variables used in the current framework. The various d_i correspond to the true delays applied to the measured functions, while the various α_i correspond to estimated versions of those quantities.

2.3.5 Function and Time Delay Estimation for Multiple Signals: Centralized Joint Scenario

In this section we introduce a centralized strategy for the multiple-TDE problem. Assume that a central processing unit can access to all data sets $\mathcal{S}_i$ for all sensors i. Let be given a set of translations $\mathbf{\Lambda} := [\alpha_1 \dots \alpha_S]^T \in \mathbb{R}^S$ and a function f. Once the statistical properties of the measurement noise ν are known, it is possible to compute the probability distribution $\mathbb{P}\left[\mathcal{S}_1^{\alpha_1}, \dots, \mathcal{S}_S^{\alpha_S} \mid \mathbf{\Lambda},\ f\right]$. Once the kernel K is fixed this probability measure induces a likelihood function $\mathcal{L}'$ on the time delay vector $\mathbf{\Lambda}$ and the unknown signal f corresponding to the negative logarithm of the joint density, a part of constant terms. Instead of $\mathcal{L}'$ we consider the following:

$$\begin{aligned} \mathcal{L} &:= -\ln\left(\mathbb{P}\left[\mathcal{S}_1^{\alpha_1}, \dots, \mathcal{S}_S^{\alpha_S} \;\middle|\; \mathbf{\Lambda},\ f\right]\right) + S\,\|f\|^2_{\mathcal{H}_K^E} \\ &= \sum_{i=1}^{S}\left(\sum_{m=1}^{M_i} \frac{\left(f\left(x_i^m - \alpha_i\right) - y_i^m\right)^2}{\sigma^2} + \|f\|^2_{\mathcal{H}_K^E}\right) \end{aligned} \tag{2.120}$$

where σ^2 is the variance of the measurement noise of y_i^m. This function can be used in order to obtain an estimate:

$$\left(\widehat{\mathbf{\Lambda}}_{ML},\ \widehat{f}^{ML}\right) := \arg \min_{\mathbf{\Lambda},\, f \in \mathcal{H}_K^E} \mathcal{L} \tag{2.121}$$

corresponding the expectation of the signal f given all the data sets and the fact that all the functions f_i are shifted versions of the same function f. In its formulation it overweights the regularization factor with respect to $\mathcal{L}'$: notice then that the minimum estimation error variance strategy corresponding to the Bayesian network of Figure 2.20 can be derived from (2.120) once the actual number of sensors S is known.

The problem defined in Equation (2.121) can now be decomposed into two sequential optimization problems. The first is:

$$\widehat{f}\left(\mathbf{\Lambda}\right) := \arg \min_{f \in \mathcal{H}_K^E} \ \mathcal{L}\left(\mathcal{S}_1^{\alpha_1}, \dots, \mathcal{S}_S^{\alpha_S} \;\middle|\; \mathbf{\Lambda},\ f\right), \tag{2.122}$$

which is convex. In fact, for any fixed $\mathbf{\Lambda}$, the different data sets can be combined in an *unique* big data set $\mathcal{S} = \cup_{i=1}^{S} S_i^{\alpha_i}$, and the optimization problem reduces to the regression of a to-be-estimated function. The second problem is:

$$\widehat{\mathbf{\Lambda}}_{ML} = \arg\min_{\mathbf{\Lambda}} \ \mathcal{L}\left(\mathcal{S}_1^{\alpha_1}, \dots, \mathcal{S}_S^{\alpha_S} \;\middle|\; \mathbf{\Lambda},\ \widehat{f}\left(\mathbf{\Lambda}\right)\right) \tag{2.123}$$

which is not convex, and in general has multiple local minima and large domain regions for which the likelihood is constant. Once $\widehat{\mathbf{\Lambda}}_{ML} = [\widehat{\alpha}_{ML,1}, \ldots, \widehat{\alpha}_{ML,S}]^T$ has been computed, the estimate of the function f is given by:

$$\widehat{f}^{ML} = \widehat{f}\left(\widehat{\mathbf{\Lambda}}_{ML}\right) \tag{2.124}$$

and solution of the multiple-TDE problem is simply $\widehat{d}_{i,j} = \widehat{\alpha}_{ML,i} - \widehat{\alpha}_{ML,j}$ for each couple i, j.

It is important to remark that numerical solution of Equation (2.123) through gradient descent algorithms is strongly dependent on the quality of the initialization. We noted that initial guess $\widehat{\mathbf{\Lambda}}(0) = \left[\alpha_1^0, \ldots, \alpha_S^0\right]^T$, where α_1^0 are defined in Equation (2.114), drastically reduces the convergence time and the probability of reaching local minima. Moreover, algorithms described in Section 2.3.3 can be used as line-searches for the updates of the translations instead of gradient-based steps.

2.3.6 Function and Time Delay Estimation for Multiple Signals: Distributed Joint Scenario

In this section we assume that TDE and function estimation has to be performed by a sensor network where each sensor can communicate directly only to a small number of neighboring sensors, i.e. communication is constrained to comply with the so-called communication graph. The centralized solution proposed in Section 2.3.5 can be too expensive in terms of bandwidth requirement, since all nodes must communicate the entire data sets $\mathcal{S}_i$ to some leader node. Inspired by Schizas and Giannakis (2006), we derive a distributed algorithm which provides the same solution of the centralized estimation problem defined in the previous section. This algorithm has three main features: it requires only limited data exchange among sensors, it distributes computation load among all sensors, and allows each sensor to compute the best maximum likelihood estimation of the unknown function.

The sensor network is represented by a graph $\mathcal{G} := (\mathcal{N}, \mathcal{E})$, where $\mathcal{E}$ indicates the communication links, $\mathcal{N}$ indicates the set of nodes, $\mathcal{N}_i$ indicates the set of neighbors of node i including the node itself, i.e. $i \in \mathcal{N}_i$. We assume that our network is a *bridged sensor network* Schizas and Giannakis (2006):

Definition 37 (bridged sensor network)**.** A sensor network is said to be *bridged* if there exist a subset $\mathcal{B} \subseteq \mathcal{N}$ of so-called *bridge nodes* satisfying:

1. each node has at least one bridge as neighbor (i.e. $\mathcal{N}_i \cap \mathcal{B} \neq \emptyset \quad \forall i \in \mathcal{N}$);
2. if two nodes can communicate directly then they must communicate directly with at least a common bridge (i.e. $\mathcal{E}(i,j) \neq 0 \Rightarrow \mathcal{N}_i \cap \mathcal{N}_j \cap \mathcal{B} \neq \emptyset \quad \forall i, j \in \mathcal{N}$).

In the following $\mathcal{B}$ will be the set of bridge nodes. We furthermore assume that communication graph $\mathcal{G}$ is undirected and connected, communications are reliable and single-hop, and no communication-delays are present. An example of such a network is drawn in Figure 2.21.

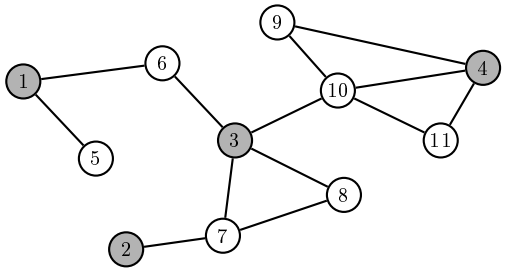


Figure 2.21: Example of *bridged* sensor network (def. 37). Greyed nodes correspond to bridge nodes, while white one to normal nodes.

In order to distributely solve Equation (2.121) using results from Schizas and Giannakis (2006), we must request observations of different nodes to be independent given the function f and translations $\mathbf{\Lambda}$:

$$\mathbb{P}\left[\mathcal{S}_1^{\alpha_1}, \ldots, \mathcal{S}_S^{\alpha_S} \;\middle|\; \mathbf{\Lambda},\ f\right] = \prod_{i=1}^{S} \mathbb{P}_i\left[\mathcal{S}_i^{\alpha_i} \;\middle|\; \alpha_i,\ f\right]. \tag{2.125}$$

Proposition 38 (Schizas and Giannakis (2006): equivalence between constrained and centralized optimizations)**.** Consider the following *constrained* optimization:

$$\{\widehat{\alpha}_{ML,i},\ \widehat{\mathbf{a}}_{ML,i}\}_{i=1}^{S} := \arg \min_{\{\alpha_i,\, \mathbf{a}_i\}_{i=1}^{S}} \quad \sum_{i=1}^{S} \mathcal{L}_i \tag{2.126}$$

$$\mathcal{L}_i := -\ln\left(\mathbb{P}_i\left[\mathcal{S}_i|\ \alpha_i,\ \mathbf{a}_i\right]\right) + \|\mathbf{a}_i\|^2_{\mathcal{H}_K^E} \tag{2.127}$$

$$\text{subject to } \mathbf{a}_i = \mathbf{b}_b \quad \forall b \in \mathcal{B} \text{ and } i \in \mathcal{N}_b. \tag{2.128}$$

If Equation (2.125) holds then solution of Equation (2.126) coincides with the solution of Equation (2.121) (centralized optimization).

Informally speaking, bridge nodes b force consensus of $\mathbf{a}_i$, and hence of $\widehat{f}$, among all nodes through constraints (2.128). Once equivalence between problems given by Equation (2.126) and Eqn (2.121) is assured, the optimization problem can be solved by finding saddle points of the *augmented Lagrangian* Γ relative to Equation (2.126) Bertsekas and Tsitsiklis (1997):

$$\begin{aligned} \Gamma := \quad & -\textstyle\sum_{i=1}^{S} \ln\left(\mathbb{P}_i\left[\mathcal{S}_i^{\alpha_i}\,|\alpha_i,\ \mathbf{a}_i\right]\right) \\ & + \|\mathbf{a}_i\|^2_{\mathcal{H}_K^E} \\ & + \textstyle\sum_{b\in\mathcal{B}} \sum_{i\in\mathcal{N}_b} \left[v_{i,b}\right]^T \cdot \left[\mathbf{a}_i - \mathbf{b}_b\right] \\ & + \textstyle\sum_{b\in\mathcal{B}} \sum_{i\in\mathcal{N}_b} \frac{c_i}{2} \|\mathbf{a}_i - \mathbf{b}_b\|^2_{\mathcal{H}_K^E} \end{aligned} \tag{2.129}$$

where: **(a)** $v_{i,b}$ are the Lagrange multipliers relative to the constraints expressed in proposition 38 and **(b)** c_i's are penalty terms Bertsekas and Tsitsiklis (1997). Note that there is no possibility to assure the existence of a single local minimum. Again, it will be useful to choose initial guesses as in Section 2.3.5. In Schizas and Giannakis (2006) a distributed solution of this problem has been proposed, via the

iterative procedure on time index h described in algorithm 3. Note that step 4 of this algorithm in general can be refined with a line-search step, see Bertsekas and Tsitsiklis (1997); Fiacco and Cormick (1968).

Algorithm 3 distributed optimization

1: (initialization)

(a) choose random lagrange multipliers $v_{i,b}(0)$

(b) choose random normal nodes estimates $\mathbf{a}_i(0)$

(c) choose random bridge nodes estimates $\mathbf{b}_b(0)$

(d) choose initial data set shifts $\alpha_i(0)$ as the various arrival times defined in Equation (2.114)

2: **for** $h = 1, \ldots$ **do**

3: **for** each link *normal node* $i \rightarrow$ *bridge node* b **do**

4: $v_{i,b}(h) = v_{i,b}(h-1) + c_i \left[\mathbf{a}_i(h) - \mathbf{b}_b(h)\right]$;

5: **for** each normal node i **do**

6:

$$\mathbf{a}_i(h+1) = \arg\min_{\mathbf{a}_i} (\Gamma_i)$$

where

$$\begin{aligned} \Gamma_i := \quad & -\ln\left(\mathbb{P}_i\left[\mathcal{S}_i^{\alpha_i(h)}\middle|\ \alpha_i(h), \mathbf{a}_i\right]\right) \\ & + \|\mathbf{a}_i\|^2_{\mathcal{H}_K^E} \\ & + \textstyle\sum_{b\in\mathcal{N}_i\cup\mathcal{B}} \left[v_{i,b}(h)\right]^T \left[\mathbf{a}_i - \mathbf{b}_b(h)\right] \\ & + \textstyle\sum_{b\in\mathcal{N}_i\cup\mathcal{B}} \frac{c_i}{2} \left\|\mathbf{a}_i - \mathbf{b}_b(h)\right\|^2_{\mathcal{H}_K^E} \end{aligned}$$

7: **for** each normal node i **do**

8:

$$\alpha_i(h+1) = \arg\min_{\alpha_i} \ -\ln\left(\mathbb{P}_i\left[\mathcal{S}_i^{\alpha_i}\middle|\ \alpha_i, \mathbf{a}_i(h+1)\right]\right)$$

9: **for** each bridge node b **do**

10:

$$\mathbf{b}_b(h+1) = \sum_{i\in\overline{\mathcal{N}}_b} \frac{1}{\sum_{j\in\overline{\mathcal{N}}_b} c_j} \left[v_{i,b}(h) + c_i \mathbf{a}_i(h+1)\right]$$

where

$$\overline{\mathcal{N}}_b := \mathcal{N}_b \cap (\mathcal{N} - \mathcal{B})$$

.

Differences between algorithm 3 and the original one are: **(a)** we apply it for estimation of functions instead of parameters vectors, **(b)** we have a separate step for likelihood maximization of the set of translations (step 8), **(c)** there are some additional Tikhonov factors (in step 6), **(d)** original algorithm is assured to converge to the global minimum since some convexity hypotheses are assumed, while algorithm 3

is assured only to converge to a local minimum.

2.3.7 Simulations

We simulated the algorithms of Sections 2.3.5 and 2.3.6 for the network of Figure 2.21, with $S = 11$ and $E = 3$. As expected, distributed algorithm's results converge to the centralized one's (compare Figures 2.22 and 2.23; comparable results have been obtained for the eigenfunctions weights $\widehat{\mathbf{a}}_i$), but we noticed that convergence velocity and stability of distributed strategy strongly depends on penalty terms c_i Bertsekas and Tsitsiklis (1997).

The usefulness of joint identification is shown in Figures 2.24 and 2.25. Sensor i may "miss" some important pieces of the signal, but sensor j helps i to reconstruct the missing part using its data set. Note that in this case classical TDE algorithms using local data sets would lead to a bigger estimation error in $\widehat{d}_{i,j}$.

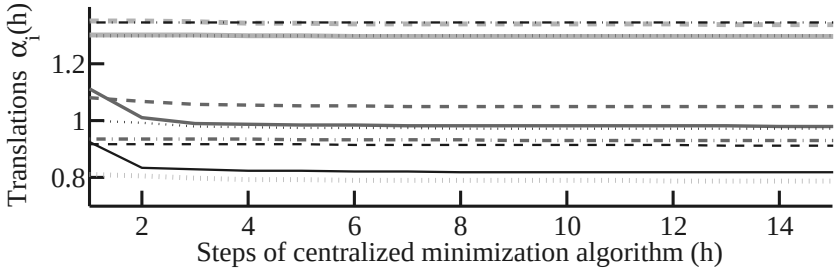


Figure 2.22: Translations $\alpha_i(h)$ applied to the various data sets $\mathcal{S}_i$ during the Newton-Raphson minimization of centralized optimization problem (2.121).

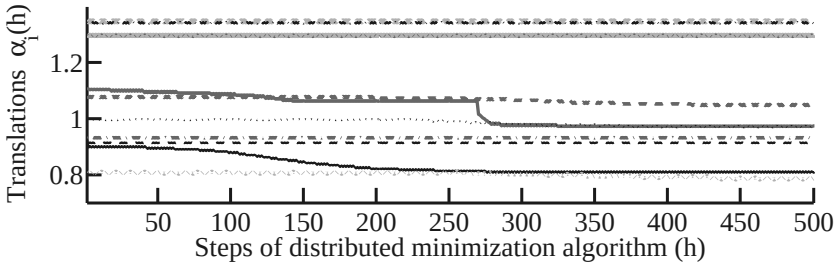


Figure 2.23: Translations $\alpha_i(h)$ applied to the various data sets $\mathcal{S}_i$ during the minimization of distributed optimization problem using algorithm 3.

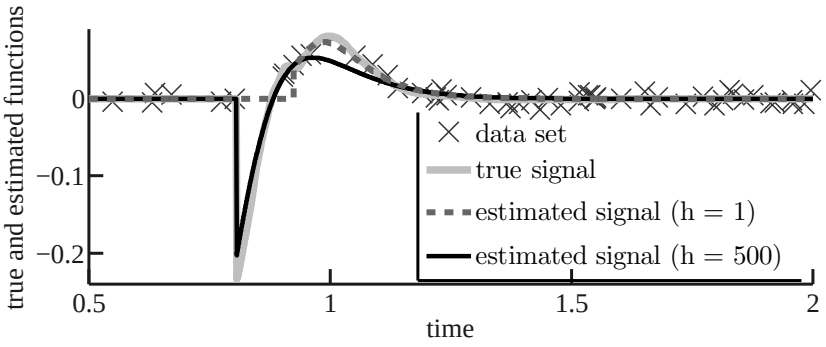


Figure 2.24: Example of data set $\mathcal{S}_i$ (gray crosses) and relative estimated functions for different steps h of minimization algorithm 3 (dotted and solid black lines). The reconstruction of the negative part of the function for $h = 500$ has been possible since other sensors measured it (see Figure 2.25).

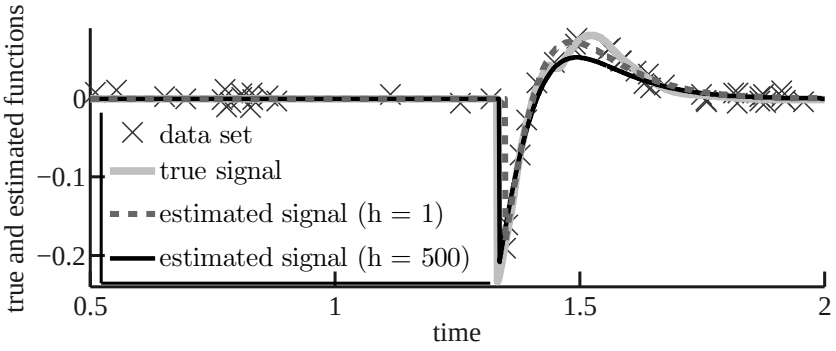


Figure 2.25: Example of data set $\mathcal{S}_j$ (gray crosses) and relative estimated functions for different steps h of minimization algorithm 3 (dotted and solid black lines). This sensor can help other sensors to reconstruct the parts of the signal where measurements are missing (see Figure 2.24).

3

Distributed Estimation of the Number of Sensors

3.1 Introduction

In this chapter we focus on the development of distributed techniques that increase the knowledge of the number of agents participating to the estimation processes.

A common way of performing this task is to use a mobile access point moving through the network. In this context, authors of Budianu et al. (2006) analyze an algorithm based on the Good-Turing estimator of the missing mass (Good, 1953) given vectors of observed sensors IDs, while in Leshem and Tong (2005) other authors propose a probabilistic sequential polling protocol associated to a sensor identification mechanism, and show that the number of transmissions per sensor required to obtain an arbitrarily desired level of estimation accuracy is logarithmically bounded. In Howlader et al. (2008) authors consider underwater communications networks, and provide a probabilistic estimation procedure for counting the number of neighbors (and not of the agents in the network) nodes with a certain accuracy. An other interesting field that has been studied is the resource inventory application. Usually in this scenario hierarchized structures are used: a certain hand-portable sensor is moved through the environment, polling for certain kinds of objects and then returning the information to a centralized server (Huang et al., 2009). There have been proposed also estimators based on the physical properties of the medium within information is transmitted (as in Huang and Barket (1991)).

The estimation of the number of active agents is important also for peer-to-peer networks. In this case there are mainly three estimation techniques (Le Merrer et al., 2006, and references therein): *randomized reports* (Kostoulas et al., 2005); *epidemic algorithms*(Jelasity and Montresor, 2004); *random walks* (Massoulié et al., 2006). These methods are generally based on concepts like distances between nodes and rely on the structure of peer-to-peer networks.

Previous works already considered similar schemes, specially using minima of collections of exponential or uniform random variables. For example, in Cohen (1997), the author obtains results on estimation confidences and accuracies levels, while in Mosk-Aoyama and Shah (2008) the authors describe how to distributedly estimate the outcomes of generic separable functions and thus to estimate the number

of agents, relating topological properties of the network and the speed of decay of the estimation error.

3.2 Problem formulation

We model a network with a graph $\mathcal{G} = \{\mathcal{N}, \mathcal{E}\}$, where $\mathcal{N} = \{1, \dots, S\}$ is the set of the sensors composing the network and $\mathcal{E} \subseteq \mathcal{N} \times \mathcal{N}$ is the set of the communication links between the sensors. We assume that the graph $\mathcal{G}$ is undirected, i.e. $(i, j) \in \mathcal{E} \Leftrightarrow (j, i) \in \mathcal{E}$, and not time-varying.

The proposed distributed strategy is such that each sensor will estimate the number of sensors in the network S only through local communications and with limited coordination among sensors, and is based on 3 steps: 1) sensors locally generate a set of random data, 2) they distributedly compute a function that takes as inputs the locally generated data, 3) from the results of this computation sensors locally estimate S. More formally:

1. each sensor $i = 1, \dots, S$ locally generates a vector of $M \in \mathbb{N}_+$ i.i.d. random values[1] $y_i^m \in \mathbb{R}$, $m = 1, \dots, M$, using a probability density $p(\cdot)$ that is the same among all sensors; does *not* depend on the actual number of sensors S, does *not* depend on the number of generated values M;

2. sensors distributedly compute the vector $\mathbf{f} \in \mathbb{R}^M$ through the function $F : \mathbb{R}^S \to \mathbb{R}$ as follows

$$\mathbf{f} := [f_1, \dots, f_M], \qquad f_m := F(y_1^m, \dots, y_S^m). \tag{3.1}$$

   F must involve only computationally simple operations and local communications among the sensors. Some examples of such computable functions are: the arithmetic mean, the maximum, the minimum and the variance (of the set of data $y_1^m, \dots, y_S^m$);

3. each sensor locally computes an estimate $\widehat{S^{-1}}$ of S^{-1} based on the vector $\mathbf{f} \in \mathbb{R}^M$, through a function $\Psi : \mathbb{R}^M \to \mathbb{R}_+$, i.e.

$$\widehat{S^{-1}} := \Psi(f_1, \dots, f_M) \ . \tag{3.2}$$

The reason for estimating S^{-1} rather than S is motivated by the fact that under general conditions the performance results will be more natural, as will be shown below. Nonetheless, we will give performance results also for estimators of S rather than S^{-1}. The strategy is illustrated in Figure 3.1.

Hypothesizing a lack of knowledge of a prior on S, a natural measure of performance is given by the conditioned Mean Square Error (MSE), namely

$$Q(p, F, \Psi) := \mathbb{E}\left[\left(S^{-1} - \widehat{S^{-1}}\right)^2\right] \tag{3.3}$$

[1] A more realistic scenario would be to consider discrete random variables, but simulations shown us that using floating point precision, from a practical point of view to use more than 12 bits does not improve the estimation performances. The mathematical analysis of the effects of discretization is then beyond the scope of this paper and kept as a future work.

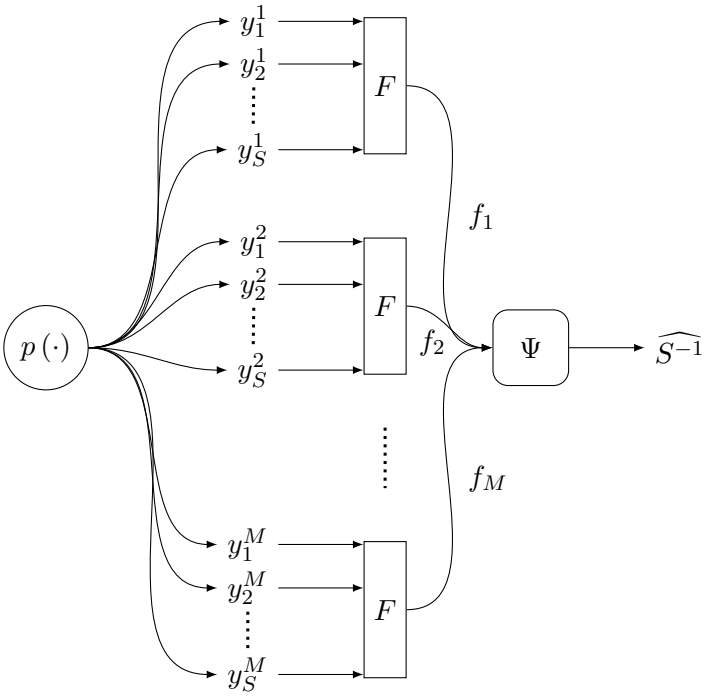


Figure 3.1: Graphical representation of the estimation strategy for the inverse of the number of sensors S^{-1}.

where we explicitly indicated the dependence on the generating p.d.f. $p(\cdot)$, the consensus function F and the estimator Ψ. Ideally we would like to minimize such error over all the possible choices of the triple (p, F, Ψ), but this is a formidable infinite dimensional problem, given the hypotheses previously posed in points 1), 2) and 3). In this work we focus on special classes of the triple (p, F, Ψ) and study the behavior of index (3.3) in these classes, to get some insights on the optimization problem for the general case. We start by looking at two simple classes examples.

3.3 Motivating Examples

3.3.1 Motivating Example 1: Gaussian + Average + Maximum Likelihood

Consider a zero-mean normal distribution for the generation of the data y_i^m, i.e. $p(y_i^m) = \mathcal{N}(0, 1)$; the average for the consensus function, i.e.

$$F(y_1^m, \ldots, y_S^m) := \frac{1}{S} \sum_{i=1}^{S} y_i^m =: f_m \quad ; \tag{3.4}$$

the Maximum Likelihood (ML) estimate for S^{-1} as the estimation function $\widehat{S^{-1}} = \Psi(f_1, \ldots, f_M)$, i.e.

$$\Psi(f_1, \ldots, f_M) := \arg\max_{S^{-1}} p\left(f_1, \ldots, f_M \left| S^{-1} \right.\right) \ . \tag{3.5}$$

Clearly $f_m \sim \mathcal{N}\left(0, S^{-1}\right) \ \ \forall m$ since all the y_i^m are i.i.d. This imply that also all the f_m are i.i.d., therefore

$$p\left(f_1, \ldots, f_M \left| S^{-1} \right.\right) = \frac{1}{\sqrt{2\pi \left(S^{-1}\right)^M}} \exp\left(-\frac{\sum_{m=1}^M f_M^2}{2S^{-1}}\right) \tag{3.6}$$

and thus, after some simple computations

$$\Psi := \arg\max_{S^{-1}} p\left(f_1, \ldots, f_M \left| S^{-1} \right.\right) = \frac{1}{M} \sum_{m=1}^M f_M^2 \ . \tag{3.7}$$

Considering $\widehat{S^{-1}} = \Psi\left(f_1, \ldots, f_M\right)$, since $\sqrt{S} f_M \sim \mathcal{N}(0,1)$, we have that

$$\sum_{m=1}^M \left(\sqrt{S} f_m\right)^2 \sim \chi^2(M) \ , \tag{3.8}$$

that can be finally traduced in

$$\frac{M}{S^{-1}} \widehat{S^{-1}} \sim \chi^2(M) \ . \tag{3.9}$$

This provides the analytic expression for the density $p\left(\widehat{S^{-1}} \left| S \right.\right)$, from which we obtain mean and variance

$$\mathbb{E}\left[\widehat{S^{-1}}\right] = S^{-1}, \qquad \operatorname{var}\left(\widehat{S^{-1}}\right) = S^{-2} \frac{2}{M} \ . \tag{3.10}$$

Hence, the estimator (3.7) is unbiased and its performance index (3.3) coincides with its variance, namely

$$Q(p, F, \Psi) = \mathbb{E}\left[\left(S^{-1} - \widehat{S^{-1}}\right)^2\right] = S^{-2} \frac{2}{M} \ . \tag{3.11}$$

As a remark, the previous expression implies that the relative estimation error $\frac{S^{-1} - \widehat{S^{-1}}}{S^{-1}}$ is independent of S.

For this example it is easy to compute the performance of the ML estimator of S rather that S^{-1}, since

$$\widehat{S} := \arg\max_S p\left(f_1, \ldots, f_M \left| S \right.\right) = \frac{M}{\sum_{m=1}^M f_m^2} = \frac{1}{\widehat{S^{-1}}} \tag{3.12}$$

therefore

$$\frac{1}{SM} \widehat{S} \sim \text{Inv-}\chi^2(M) \tag{3.13}$$

and thus

$$p\left(\widehat{S} \left| S \right.\right) = \Gamma\left(\frac{M}{2}\right)^{-1} \frac{1}{\widehat{S}} \left(\frac{M}{2} \frac{S}{\widehat{S}}\right)^{\frac{M}{2}} \exp\left(-\frac{M}{2} \frac{S}{\widehat{S}}\right) \tag{3.14}$$

where $\Gamma(\cdot)$ is the Gamma function. From this it follows that

$$\mathbb{E}\left[S - \widehat{S}\right] = \frac{S}{M-2}, \tag{3.15}$$

$$\mathrm{var}\left(\widehat{S}\right) = \frac{2S^2}{M}\frac{M^3}{(M-2)^2(M-4)} \tag{3.16}$$

and therefore the mean square error for $\widehat{S}$ is

$$\mathbb{E}\left[\left(S-\widehat{S}\right)^2\right] = S^2\frac{2M^3+M\left(M-4\right)}{M\left(M-2\right)^2\left(M-4\right)}\ . \tag{3.17}$$

Notice now that asymptotically estimators $\widehat{S}$ and $\widehat{S^{-1}}$ have the same relative estimation error, since

$$\lim_{M\to+\infty}\mathbb{E}\left[\left(S-\widehat{S}\right)^2\right] = S^2\frac{2}{M}\ . \tag{3.18}$$

Order statistics

Before proceeding, we recall some basic results relating order statistics, see David and Nagaraja (2003). Assume S to be the number of elements of the sample $y_1^m,\ldots,y_S^m$, and $f_m^{(k)}$ to be its k-th order statistic. Let every y_i^m be i.i.d. and let $p\left(a\right)$ be its probability density evaluated in a, and $P\left(a\right)$ be its probability distribution evaluated in a. Then

$$p_{f_m^{(k)}}\left(a\right) = \frac{S!\,P\left(a\right)^{(k-1)}\left(1-P\left(a\right)\right)^{(S-k)}p\left(a\right)}{(k-1)!\,(S-k)!} \tag{3.19}$$

while the joint density $p_{f_m^{(k)}f_m^{(j)}}\left(a_1a_2\right)$ is given by

$$\begin{aligned} p_{f_m^{(k)}f_m^{(j)}}\left(a_1a_2\right) = & \frac{S!}{(k-1)!\,(j-k-1)!\,(S-j)!} \\ & \cdot\left(P\left(a_2\right)-P\left(a_1\right)\right)^{(j-k-1)} \\ & \cdot\left(1-P\left(a_2\right)\right)^{(S-j)}P\left(a_1\right)^{(k-1)} \\ & \cdot p\left(a_1\right)p\left(a_2\right)\ . \end{aligned} \tag{3.20}$$

Order statistics can be defined also for K-uples: let

$$\mathcal{K} := \left\{k_1,\ldots,k_K\right\}\ , \qquad k_j\in\mathbb{N}_+ \tag{3.21}$$

be a set of indexes of order statistics, where 1 denotes the minimal element. Given the set of order statistics indexes (3.21), and defining $a_0 := -\infty$, $a_{K+1} := +\infty$, $k_0 := 0$, $k_{K+1} := K+1$, the joint density of $f_m^{(k_1)},\ldots,f_m^{(k_K)}$ is given by

$$p_{f_m^{(k_1)},\ldots,f_m^{(k_K)}}\left(a_1,\ldots,a_K\right) = S!\left[\prod_{j=1}^{K}p\left(a_{k_j}\right)\right]\left[\prod_{j=0}^{K}\frac{\left(P\left(a_{k_{j+1}}\right)-P\left(a_{k_j}\right)\right)^{k_{j+1}-k_j-1}}{(k_{j+1}-k_j-1)!}\right]\ . \tag{3.22}$$

3.3.2 Motivating example 2: Uniform + maximum + ML

Consider now data y_i^m to be uniformly distributed, i.e. $p\left(y_i^m\right) = \mathcal{U}\left[0,1\right]$ and define the consensus function F to be the maximum, i.e.

$$F\left(y_1^m,\ldots,y_S^m\right) := \max_i\left\{y_i^m\right\} =: f_m\ . \tag{3.23}$$

Again the ML estimator for S^{-1} is used to define Ψ (see Equation (3.5)). The probability density of the S-th order statistic f_m is known and in general given by Equation (3.19). In this case

$$p\left(f_m \left| S\right.\right) = S f_m^{S-1} \qquad \forall m \; . \tag{3.24}$$

Therefore

$$p\left(f_1, \ldots f_M \left| S\right.\right) = \prod_{m=1}^{M} p\left(f_m \left| S\right.\right) = S^M \prod_{m=1}^{M} f_m^{S-1} \quad \forall m \; . \tag{3.25}$$

Again, after some simple computations

$$\Psi := \arg\max_{S^{-1}} p\left(f_1, \ldots, f_M \left| S^{-1}\right.\right) = -\frac{1}{M} \sum_{m=1}^{M} \log\left(f_M\right) \; . \tag{3.26}$$

Now, defining $z := -\log\left(f_m\right)$, it is immediate to check that z is an exponential random variable with rate S, i.e.

$$p\left(z \left| S\right.\right) = \begin{cases} S\exp\left(-Sz\right) & \text{if } z \geq 0 \\ 0 & \text{otherwise.} \end{cases} \tag{3.27}$$

The sum of M i.i.d. exponential random variables with rate S is a Gamma random variable with shape M and scale $\frac{1}{S}$. Considering then that $\widehat{S^{-1}} = \Psi\left(f_1, \ldots, f_M\right)$ is thus a scaled version of this sum of exponentials, it follows that

$$\frac{M}{S^{-1}} \widehat{S^{-1}} \sim \text{Gamma}\left(M, 1\right) \tag{3.28}$$

from which it is immediate to compute mean and variance

$$\mathbb{E}\left[\widehat{S^{-1}}\right] = S^{-1}, \quad \text{var}\left(\widehat{S^{-1}}\right) = S^{-2}\frac{1}{M} \; . \tag{3.29}$$

This implies that the estimator $\widehat{S^{-1}}$ is unbiased and that its performance index (3.3) coincides with its variance, namely

$$Q\left(p, F, \Psi\right) = \mathbb{E}\left[\left(S^{-1} - \widehat{S^{-1}}\right)^2\right] = S^{-2}\frac{1}{M} \; . \tag{3.30}$$

By considerations similar to those of Section 3.3.1, one obtains that $\widehat{S}$ is asymptotically unbiased, with asymptotic variance equal to that of $\widehat{S^{-1}}$.

Notice that, given a fixed M and comparing Equations (3.11) and (3.30), the performance index of the estimation scheme of this section is exactly half as large as the one of Section 3.3.1.

3.3.3 Discussion on the motivating examples

We remark some points regarding the previous two examples. First, the estimator $\widehat{S^{-1}} = \Psi\left(f_1, \ldots, f_M\right)$ can be decomposed into simpler blocks, as shown in Figure 3.2. Following that scheme, all the quantities f_m are passed through the same

nonlinear function $\psi : \mathbb{R} \to \mathbb{R}$ transforming each f_m into an unbiased estimate $\widehat{S_m^{-1}} := \psi(f_m)$, $m = 1, \ldots, M$ of S^{-1}. Now, since the f_m are uncorrelated, also the $\widehat{S_m^{-1}}$ are uncorrelated. This implies that, to obtain the global estimate using all the available information, the various $\widehat{S_m^{-1}}$ have simply to be combined through an arithmetic mean

$$\widehat{S^{-1}} = \frac{1}{M} \sum_{m=1}^{M} \widehat{S_m^{-1}} \,. \tag{3.31}$$

In fact, in Section 3.3.1 we had $\psi(\cdot) = (\cdot)^2$, while in Section 3.3.2 we had $\psi(\cdot) = -\log(\cdot)$.

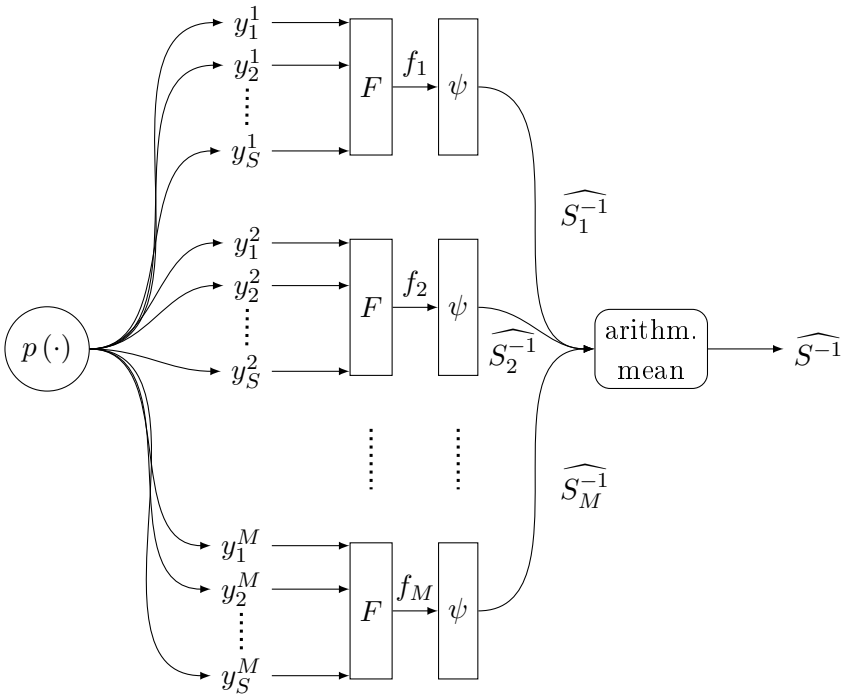


Figure 3.2: Alternative graphical representation of the estimation strategies for S^{-1} proposed in Section 3.3.1 and Section 3.3.2.

The second point is that being each $\widehat{S_m^{-1}}$ an unbiased estimate, the variance of the combined estimate $\widehat{S^{-1}}$ will decrease as $\frac{1}{M}$, and the quality of this variance will depend on the variance of the single estimates $\widehat{S_m^{-1}}$. Moreover, comparing Equations (3.11) and (3.30), we can say that for a fixed M the variance of the error associated to strategy of Section 3.3.2 is half the one associated Section 3.3.1. This is particularly positive since the distributed computation of the maximum of a set of values is much faster than the computation of its average (even if both depend on the size of the network). Sensors can compute maxima simply broadcasting their values and updating them (when receiving messages) via elementwise maximum operations. Under mild conditions it can be proven that each node will correctly compute maxima after a finite number of messages exchange. Also averages can be computed in a

distributed fashion (through average consensus algorithms), but the rate of convergence to the actual average is exponential. For example, in a circular network where each node has only two neighbors (left and right), the number of steps necessary to correctly compute the maximum is $T_{\max} = S/2$, while the number of steps required to achieve a 1%-error in the estimation of the average is

$$T_{\text{ave}} = \frac{\log(0.01)}{\log(1 - 2\pi^2/S^2)} \gg T_{\max} . \tag{3.32}$$

In the next sections we will continue to consider the average and the maximum for F, aiming to understand for which p.d.f. $p(\cdot)$'s generating the various y_i^m the scheme of Figure 3.2 continues to be applicable. Moreover we seek to understand when the ML estimation scheme is optimal (i.e. minimize the performance index (3.3)) and which p.d.f. $p(\cdot)$ (or class of p.d.f.'s) leads to the smallest estimation error.

Considerations on the discrete nature of S

Being $S \in \mathbb{N}_+$, its estimates must also be in $\mathbb{N}_+$. Despite this fact, in the following we will consider estimators that assume $S \in \mathbb{R}_+$. The natural step that have to be performed beyond the estimators that we will analyze, is then to check which one between $\lfloor \widehat{S} \rfloor$ and $\lceil \widehat{S} \rceil$ is the most likely estimate, in order to have an overall solution in $\mathbb{N}_+$.

3.4 Special case: average consensus

Let

$$F_{\text{ave}} := F(y_1^m, \dots, y_S^m) = \frac{1}{S} \sum_{i=1}^{S} y_i^m =: f_m . \tag{3.33}$$

Assume data y_i^m to be generic Gaussian r.v.'s, i.e. $y_i^m \sim p(y_i^m) = \mathcal{N}(\mu, \sigma^2)$. It is easy to show, following the same steps of Section 3.3.1, that the ML estimator for S^{-1} in this case is given by

$$\widehat{S^{-1}} := \Psi_{\text{ML}}(f_1, \dots, f_M) = \frac{1}{M} \sum_{m=1}^{M} \frac{(f_m - \mu)^2}{\sigma^2} \tag{3.34}$$

$$\widehat{S^{-1}} \sim \frac{S^{-1}}{M} \chi^2(M) . \tag{3.35}$$

It is possible to derive the following

Proposition 39. Let $\mathcal{N}$ be the class of all Gaussian random variables with positive variance, i.e. $p \in \mathcal{N}$ if $p = \mathcal{N}(\mu, \sigma^2)$ for some μ and $\sigma^2 > 0$. Then Ψ_{ML} is the Minimum-Variance Unbiased Estimator (MVUE) for S^{-1} within this class. Moreover we have

$$\min_{\substack{p(\cdot) \in \mathcal{N} \\ \Psi \text{ s.t. } \mathbb{E}[\Psi] = S^{-1}}} Q(p, F_{\text{ave}}, \Psi) = Q(\mathcal{N}(0,1), F_{\text{ave}}, \Psi_{\text{ML}}) = \frac{2}{M} . \tag{3.36}$$

Proof. This proposition can be proven with reasonings similar to those followed in that of Proposition 40 (e.g., completeness can be proved just repeating the same argument relative to (3.49) except that the chi-square in place of the Gamma distribution has to be considered). We thus refer to that proof and omit this one. □

Given the definition of Equation (3.3) and a generic density $p(\cdot)$, it is not obvious whether the ML strategy is minimizing Q. Moreover, given a certain $p(\cdot)$ and restricting Ψ to be an ML estimator, it is not easy to find an analytic expression for $\widehat{S^{-1}}$ nor its distribution. One little step forward we can make is to notice that translations and scaling of a certain random variable do not affect the performance of the optimal estimator: in fact, assume $p_x(a)$ to be a generic probability density with mean μ and variance $\sigma^2 > 0$, and y to be the zero-mean unit-variance random variable

$$y = \frac{x - \mu}{\sigma} \quad , \qquad x \sim p_x \tag{3.37}$$

with corresponding density $p_y(a) = \sigma p_x(\sigma a + \mu)$. Using the invariance of the average function F_{ave} with respect to translation and scaling (it is a linear function), it is immediate to show that

$$\min_{\Psi} Q(p_x, F_{\text{ave}}, \Psi) = \min_{\Psi} Q(p_y, F_{\text{ave}}, \Psi) \ , \tag{3.38}$$

and this allows us to restrict to distributions $p(\cdot)$ with zero mean and unit variance.

When we choose the average F_{ave} as the network function, it is not evident how to optimally choose the density $p(\cdot)$ and the estimator function Ψ to minimize the index Q. In case of large networks, with large S, we can still exploit the central limit theorem: if y_i^m s.t. $y_i^m \sim p(\cdot)$ and $\mathbb{E}[y_i^m] = 0$, $\text{var}(y_i^m) = 1$, then $f_m = F_{\text{ave}}(y_1^m, \ldots, y_S^m)$ has in general the following probability distribution

$$p_{f_m}(a) = \underbrace{(p * \cdots * p)}_{S \text{ times}}(a) \tag{3.39}$$

where the symbol $*$ indicates the convolution operator. But from the central limit theorem it follows that, in distribution,

$$\lim_{S \to +\infty} p_{f_m}(\cdot) = \mathcal{N}\left(0, S^{-1}\right) \ . \tag{3.40}$$

As a consequence, it is likely that for large S there is no advantage of using probability distributions $p(\cdot)$ and estimator functions Ψ different from the unit normal distribution and the ML, respectively, i.e.

$$\lim_{S \to +\infty} \min_{p, \Psi} Q(p, F_{\text{ave}}, \Psi) = Q(\mathcal{N}(0,1), F_{\text{ave}}, \Psi_{\text{ML}}) \tag{3.41}$$

although this claim should be rigorously proven. For small S we currently do not have optimality results and we are exploring if there are non-Gaussian distributions $p(\cdot)$ leading to better estimation performance.

3.5 Special case: max consensus

Let

$$F_{\max} := F\left(y_1^m, \ldots, y_S^m\right) = \max_i \left\{y_i^m\right\} =: f_m \ . \tag{3.42}$$

We can notice immediately that, if y_i^m has probability density $p\left(a\right)$ and distribution $P\left(a\right)$, then Equation (3.19) leads to a joint density on $f_1, \ldots, f_M$ of the form

$$p_{f_1,\ldots,f_M}\left(a_1, \ldots, a_M\right) = S^M \prod_{m=1}^{M} P\left(a_m\right)^{S-1} p\left(a_m\right) \ . \tag{3.43}$$

The generic ML estimator for S^{-1} is thus

$$\widehat{S^{-1}} = \Psi_{\mathrm{ML}}\left(f_1, \ldots, f_M\right) := -\frac{1}{M} \sum_{m=1}^{M} \log\left(P\left(f_m\right)\right) \ . \tag{3.44}$$

It is immediate to show that the relative ML estimator $\widehat{S}$ of S is given by $\widehat{S} = 1/\Psi_{\mathrm{ML}} = 1/\widehat{S^{-1}}$. Define $\mathcal{P}$ as the class of densities $p\left(\cdot\right)$ whose relative distribution $P\left(\cdot\right)$ is strictly monotonic and continuous. Then the estimators $\widehat{S^{-1}}$ and $\widehat{S}$ are characterized by the following propositions

Proposition 40. $\forall\, P\left(a\right) \in \mathcal{P}$, Ψ_{ML} is the MVUE of S^{-1}.

Proof. We start considering Proposition 40. Define

$$T\left(f_1, \ldots, f_M\right) := -\sum_{m=1}^{M} \log\left(P\left(f_m\right)\right) \tag{3.45}$$

and (with little abuse of notation)

$$\widehat{S^{-1}}\left(T\right) := \widehat{S^{-1}}\left(T\left(f_1, \ldots, f_M\right)\right) := \widehat{S^{-1}}\left(f_1, \ldots, f_M\right) \ . \tag{3.46}$$

Since $p_{f_1,\ldots,f_M}\left(a_1, \ldots, a_M \,|S\right)$ can be rewritten as

$$\left(\prod_{m=1}^{M} p\left(a_m\right)\right) \left(S^M \exp\left(-\left(S-1\right) T\left(a_1, \ldots, a_M\right)\right)\right) \tag{3.47}$$

for the Fisher-Neyman factorization theorem $T\left(f_1, \ldots, f_M\right)$ is a sufficient statistic for S. From the Lehmann-Scheffé theorem, we know that *if* T is also complete and $\mathbb{E}\left[\widehat{S^{-1}}\right] = S^{-1}$, *then* $\widehat{S^{-1}}$ it is MVUE for S^{-1}. Considering now that the p.d.f. of the r.v. $P\left(f_m\right)$ is $S f_m^{S-1}$, for the same reasonings made in Equation (3.27) we have that $-\log\left(P\left(f_m\right)\right)$ is an exponential r.v. This implies that

$$M\widehat{S^{-1}} = -\sum_{m=1}^{M} \log\left(P\left(f_m\right)\right) \sim \mathrm{Gamma}\left(M, \frac{1}{S}\right) \tag{3.48}$$

and thus condition $\mathbb{E}\left[\widehat{S^{-1}}\right] = S^{-1}$ is satisfied. The completeness of $T\left(f_1, \ldots, f_M\right)$ can be proved showing that if $g\left(T\right)$ is a generic measurable function s.t. $\mathbb{E}\left[g\left(T\right) \mid S\right] =$

0 independently of S, then it must be $g(\cdot) \equiv 0$ almost everywhere (a.e.). Considering that T is Gamma $\left(M, \frac{1}{S}\right)$, the previous condition on the expectation can be rewritten as

$$\Gamma(M)^{-1} S^M \int_0^{+\infty} g(T) T^{M-1} \exp(-TS)\, dT \equiv 0 \,. \tag{3.49}$$

This is equivalent to say that the Laplace transform of $g(T) T^{M-1}$ has to be zero a.e., and this happens if and only if $g(T)$ is zero a.e. This proves the completeness of T and thus the proposition. □

This means that if we restrict Ψ to be unbiased, then $\widehat{S^{-1}} = \Psi_{\mathrm{ML}}$ is optimal with respect to index (3.3). In addition, it is possible to prove that the performance of the estimator is independent of the adopted density

Proposition 41. It holds that

$$\begin{aligned} \min_{p(\cdot) \in \mathcal{P}} \quad & Q(p, F_{\max}, \Psi) = Q(\mathcal{U}[0,1], F_{\max}, \Psi_{\mathrm{ML}}) \,. \\ \Psi \text{s.t. } & \mathbb{E}[\Psi] = S^{-1} \end{aligned} \tag{3.50}$$

Proof. This is a consequence of (3.49) that shows that the distribution of the estimator is independent of the density p which is adopted. □

Uniform generation with min consensus

For symmetry reasons, the usage of max or min consensus strategies lead to the same performance results. In case of min consensus, the general expression for ML estimators of S^{-1} is

$$\Psi(f_1, \ldots, f_M) = -\frac{1}{M} \sum_{m=1}^{M} \log(1 - P(f_m)) \,. \tag{3.51}$$

3.6 Special case: range consensus

Running max and min consensus in parallel it is possible to find simultaneously the statistics of order 1 and S of $\{f_m\}$:

$$\overline{f_m} := \max_i \{y_i^m\} \qquad\qquad \underline{f_m} := \min_i \{y_i^m\} \,. \tag{3.52}$$

If y_i^m has probability density $p(a)$ and distribution $P(a)$, then Equation (3.20) leads to a joint density on $\overline{f_m}, \underline{f_m}$ of the form

$$p_{\overline{f_m}\underline{f_m}}(a_1\ a_2) = \left(S^2 - S\right)(P(a_1) - P(a_2))^{S-2} \cdot p(a_1)\, p(a_2) \tag{3.53}$$

whenever $a_1 \geq a_2$, while $p_{\overline{f_m}\underline{f_m}}(a_1\ a_2) = 0$ otherwise. The joint density on $\overline{f_1}, \ldots, \underline{f_M}$ can be immediately computed and minimized in S, in order to obtain a general ML estimator of the form

$$\widehat{S} = \frac{1}{2} - L^{-1} + \sqrt{\frac{1}{4} + L^{-2}} \tag{3.54}$$

where

$$L := \frac{1}{M}\sum_{m=1}^{M} \log\left(P\left(\overline{f_m}\right) - P\left(\underline{f_m}\right)\right) \ . \tag{3.55}$$

Considering again $y_i^m \sim \mathcal{U}[0,1]$, the joint density is thus

$$p\left(\overline{f_1}, \ldots, \underline{f_M} \,|S\right) = S^M (S-1)^M \prod_{m=1}^{M} \left(\overline{f_m} - \underline{f_m}\right)^{S-2} \tag{3.56}$$

and Equation (3.55) becomes

$$L := \frac{1}{M}\sum_{m=1}^{M} \log\left(\overline{f_m} - \underline{f_m}\right) \ . \tag{3.57}$$

Notice that the p.d.f. of $r_m := \overline{f_m} - \underline{f_m}$ is

$$\begin{aligned} p\left(r_m \,|S\right) &= \frac{\partial}{\partial r_m}\left(1 - \int_{r_m}^{1}\int_{0}^{\overline{f_m}-r_m} p\left(\overline{f_m}, \underline{f_m} \,|S\right) d\underline{f_m} d\overline{f_m}\right) \\ &= S(S-1)\left(r_m^{S-2} - r_m^{S-1}\right) \end{aligned} \tag{3.58}$$

which transformation $\log(r_m)$ is anymore a r.v. which p.d.f. is analytically known. For this reason we are not able to find the density of L (and thus of $\widehat{S}|S$) in a closed form. In this case $p\left(\widehat{S}\,|S\right)$ shall be estimated with Monte Carlo simulations, as we made in Figure 3.3. Rather surprisingly, simulations show that, with the same amount of information exchanged among the sensors and with $y_i^m \sim \mathcal{U}[0,1]$, the approach performs slightly better than the max consensus one (notice that for a given M the range-consensus scheme actually computes $2M$ sensible data).

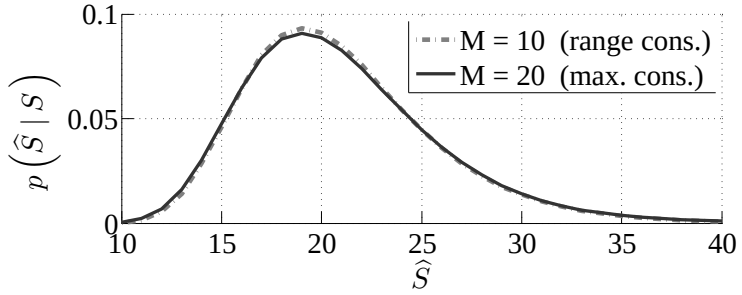


Figure 3.3: Empirical $p\left(\widehat{S}\,|S\right)$'s relative to max and range consensus strategies ($2\cdot 10^6$ number of samples, $S = 20$).

3.6.1 Range consensus with a generic number of samples

It is possibile to generalize the strategy proposed in Section 3.6 to consider a general combination of max and min consensus techniques. If $f_m^{(k)}$ is then the statistic of

order k of the set of samples $\{y_1^m, \ldots, y_S^m\}$, and if $\mathcal{K} := \{k_1, \ldots, k_K\}$ is a set of different order statistics indexes, sensors can distributely compute the matrix

$$\mathcal{F} := \begin{bmatrix} f_1^{(k_1)} & \cdots & f_1^{(k_K)} \\ \vdots & & \vdots \\ f_M^{(k_1)} & \cdots & f_M^{(k_K)} \end{bmatrix} \tag{3.59}$$

using opportune combinations of max and min consensus algorithms.

We notice that to estimate the statistic $f_m^{(S-k)}$ using max consensus algorithms, it is necessary to estimate also the statistics $f_m^{(S)}, \ldots, f_m^{(S-k+1)}$, and a simmetric consideration applies for the min consensus. Thus, assuming to have a constraint on the number of estimable statistics K, it follows immediately that the structure of $\mathcal{K}$ must be

$$\mathcal{K} = \{1, 2, \ldots, k_{\min}, k_{\max}, \ldots, S-1, S\} \tag{3.60}$$

where $k_{\max} = S - \beta$ (with β an opportune integer) and with $k_{\min} + \beta + 1 = K$.

Remark 42. If the number of used statistics is K, then (ignoring quantization effects) all the S s.t. $S < K$ will be almost surely (a.s.) correctly identified, since the matrix $\mathcal{F}$ will be not entirely filled. Even if for ease of notation we will ignore this in our mathematical derivations, practical implementation of the proposed estimators should consider this important property.

Assume now the set $\mathcal{K}$ to be as in Equation (3.60), where the statistics up to $k_{\min}$ are computed using min consensus strategies, while the other ones are computed using max consensus ones. Constraining Ψ to be an ML estimator, from Equation (3.22) we have that the ML condition is given by

$$\sum_{j=0}^{k_{\min}+\beta} \frac{1}{S-j} = -\frac{1}{M} \sum_{m=1}^{M} \log\left(P\left(f_m^{(k_{\max})}\right) - P\left(f_m^{(k_{\min})}\right)\right) \tag{3.61}$$

where we impose that:

- if $k_{\min} = 0$ (i.e. no min consensus are applied) then $P\left(f_m^{(k_{\min})}\right) = 0$;
- if $k_{\max} = S + 1$ (i.e. no max consensus are applied) then $P\left(f_m^{(k_{\max})}\right) = 1$.

Rewriting now condition (3.61) as

$$\mathcal{S}^{-1} = \frac{1}{M} \sum_{m=1}^{M} \psi\left(f_m^{(k_{\max})}, f_m^{(k_{\min})}\right) \tag{3.62}$$

with

$$\mathcal{S}^{-1} := \sum_{j=0}^{k_{\min}+\beta} \frac{1}{S-j} \tag{3.63}$$

and

$$\psi\left(f_m^{(k_{\max})}, f_m^{(k_{\min})}\right) := -\log\left(P\left(f_m^{(k_{\max})}\right) - P\left(f_m^{(k_{\min})}\right)\right) , \tag{3.64}$$

we notice the following:

1. of the generic m-th row of matrix $\mathcal{F}$, only $f_m^{(k_{\max})}$ and $f_m^{(k_{\min})}$ are considered: this implies that the values of all the other statistics are *ininfluent* for the ML estimator;

2. each of these couples is transformed through the same nonlinear function $\psi(\cdot) : \mathbb{R} \to \mathbb{R}$ into an estimate $\mathcal{S}_m^{-1}$ of $\mathcal{S}^{-1}$. due to the independence (in m) of the various $f_m^{(k_{\max})}$ and $f_m^{(k_{\min})}$, the $\mathcal{S}_m^{-1}$ are uncorrelated. This implies that the global estimate is obtained simply combining the various $\mathcal{S}_m^{-1}$'s through an arithmetic mean, like in Equation (3.62). Now, being each $\psi(\cdot)$ an unbiased estimator of $\mathcal{S}^{-1}$, once again the variance of the combined estimate will decrease as $\frac{1}{M}$, and the quality of this variance will depend on the variance of the single estimates $\mathcal{S}_m^{-1}$.

We have thus shown that also in this general case the estimation process involves arithmetic averages of the "local" estimates, as depicted in Figure 3.4.

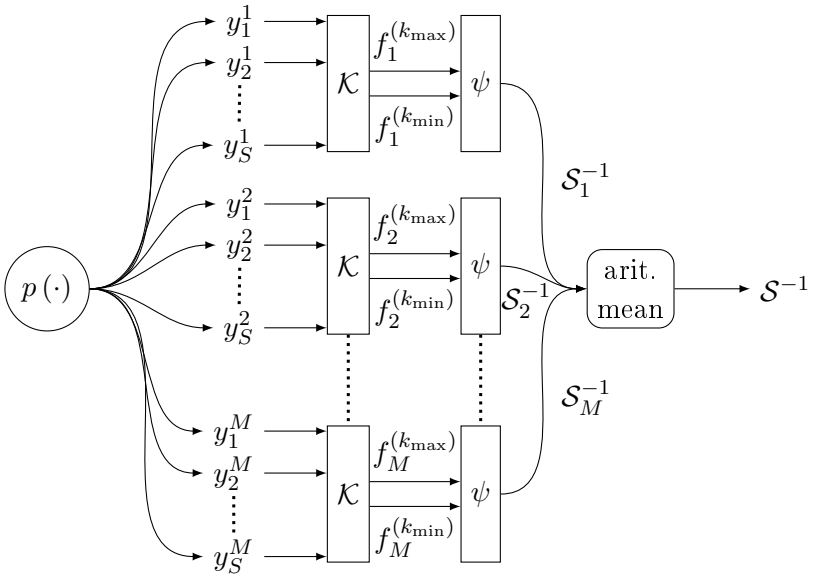


Figure 3.4: Graphical representation of the estimation strategy for generic range-consensus based estimators.

3.7 Bayesian modeling

In some cases it is possible to combine the ML strategies developed before with possible prior information on S. An interesting case is the following: assume the a priori mass probability of S^{-1} can be approximated with the Inverse-Gamma prior

$$p\left(\frac{1}{S}|\alpha,\ \beta\right) \propto \left(\frac{1}{S}\right)^{-\alpha-1} \exp\left(-\beta S\right)\ , \tag{3.65}$$

dependent on the (known) hyperparameters α (shape) and β (scale), with mean $\frac{\beta}{\alpha-1}$ and mode[2] $\frac{\beta}{\alpha+1}$. Consider then the following

Definition 43. Assume a certain prior on θ, say $p(\theta)$, is given. A likelihood $p(\mathbf{f}|\theta)$ is said to be *conjugated* with the prior $p(\theta)$ if the posterior $p(\theta|\mathbf{f}) \propto p(\mathbf{f}|\theta)\, p(\theta)$ is in the same family of $p(\theta)$ (i.e. it has the same parametric representation, even if with different values of the parameters).

In case of Gaussian plus average consensus it is possible to use the conjugated prior property of Definition 43 to obtain closed forms for MMSE and MAP estimators. In fact in this case the a-posteriori becomes

$$p\left(\frac{1}{S}\,|f_1,\ldots,f_M,\ \alpha,\ \beta\right) \propto \left(\frac{1}{S}\right)^{-\alpha-\frac{1}{2}} \exp\left(-\left(\beta+\frac{\sum_{m=1}^{M} f_m^2}{2}\right)S\right)\ . \qquad (3.66)$$

The expressions of MAP and MMSE estimators are now easily computable and summarized in Table 3.1.

$\widehat{S}_{\mathrm{ML}}$	$\widehat{S}_{\mathrm{MAP}}$	$\widehat{S}_{\mathrm{MMSE}}$
$\dfrac{M}{\sum_{m=1}^{M} f_m^2}$	$\dfrac{M+2\alpha+2}{\sum_{m=1}^{M} f_m^2+2\beta}$	$\dfrac{M+2\alpha-2}{\sum_{m=1}^{M} f_m^2+2\beta}$

Table 3.1: Summary of the various estimators for Gamma priors on S, Gaussian distribution for the generation of the data y_i^m and F set to average-consensus.

[2]This is equivalent to suppose an approximated Gamma prior on S with shape $k=\alpha$ and scale $\theta=\beta^{-1}$.

Conclusions

In this thesis we focused on network of agents that collaborate in order to achieve a common task, but that do not have strong knowledge about the state of the other cooperating units. Contextualizing this work, we aimed to increase the independence of sensor networks on human intervention.

In the first part we evaluated the performance of fundamental distributed parametric and nonparametric estimation algorithms, with the main questions underneath all the derivations summarizable in:

- is the collaboration among sensors bringing some benefits with respect to behaviors where agents do not share information?
- can the agents distributedly compute the same results that would be obtained collecting all the information in an unique place?
- if not, are the approximations that necessarily have to be introduced leading to excessively unreasonable results?
- can the agents understand by themselves and in a distributed fashion this degree of unreasonability?

In a general framework, the complexity of these questions is formidable, and might lead to no answers. We thus focused on the following significative scenario, where restrictions do not affect the applicability of the results on real-world scenarios, by assuming that:

- the task is to estimate some continuous-valued quantity;
- the information computed at the end of the process has to be the same among agents;
- the cost-functions penalizing the estimation errors are quadratic.

In this scenario, we obtained important answers that can be summarized in few sentences. First of all,

the structure of the distributed estimators is the same of both local and centralized optimal estimators.

Moreover,

in order to distributedly compute the centralized solution, agents need to have some level of knowledge of the topology of the network.

In particular, both in parametric and in non-parametric frameworks,

in some cases, the distributability of the centralized solution is possible once the agents know how many they are.

From the beginning, the knowledge of the number of collaborating sensors assumed an important role. Its importance increased after answering to the first question by paraphrasing Theorem 4 and its subsequent versions:

if the knowledge of the actual number of agents is sufficiently accurate, then distributed estimates are assured to perform better than local ones.

And this knowledge became even more important replying the last question with:

agents can distributedly bound the performance losses incurred using distributed algorithms instead of centralized solutions. Moreover, the accuracy of these bounds depends on the accuracy of the knowledge of the number of agents in the network.

After answering these questions, we proposed distributedly compute approximated Regularization Networks with small computational, memory an communication requirements. Then we proposed a random fields regressor, derived from a modification of the previously analyzed nonparametric estimation schemes. In this way we developed an effective distributed algorithm that *do not require the sensors to know where or when measurements have been sampled*, thus allowing non-uniform spatial or temporal sampling grids.

We finally focused on understanding how it is possible to distributedly increase the knowledge of the number of agents in a network. Rather than relying on strategies that count the number of nodes by means of information like IDs, serial numbers, etc., we offered a statistical algorithm that is based on computations of averages or order statistics, and noticed its mathematical properties and effectiveness. But the main nice property is that it can be run in parallel to the regression algorithm without worsening the complexity of the communication scheme.

In conclusion, we offer some questions considered as our future works. As a first main topic, we are seeking answers to the initial questions in case of distributed classification algorithms, in case of different cost functions and in case of different distributed estimation algorithms. Moreover, we are considering strategies that allow the sensors to independently and autonomously tune the parameters of the previous regression algorithms: this because it is important, for example, to let the agents understand by themselves which is the optimal regularization parameter. As a concluding topic, we should analyze the effects of practical issues like quantization or finite numbers of consensus steps in the algorithms for the estimation of the number of sensors.

Bibliography

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control 19*(6), 716 – 723.

Akyildiz, I., W. Su, Y. Sankarasubramaniam, and E. Cayirci (2002, August). A survey on sensor networks. *IEEE Communications Magazine 40*(8), 102 – 114.

Anderson, B. D. O. and J. B. Moore (1979). *Optimal Filtering*. Prentice-Hall.

Aronszajn, N. (1950). Theory of reproducing kernels. *Transactions of the American Mathematical Society 68*, 337 – 404.

Azaria, M. and D. Hertz (1984, April). Time delay estimation by generalized cross correlation methods. *IEEE Transactions on Acoustics, Speech, and Signal Processing 32*(2), 280–285.

Basseville, M. and I. V. Nikiforov (1993, April). *Detection of Abrupt Changes - Theory and Application*. Prentice-Hall.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis* (2nd ed.). Springer - Verlag.

Bertsekas, D. P. and J. N. Tsitsiklis (1997). *Parallel and Distributed Computation: Numerical Methods*. Belmont, MA: Athena Scientific.

Blum, R., S. Kassam, and H. V. Poor (1997, January). Distributed detection with multiple sensors II: Advanced topics. *Proceedings of the IEEE 85*, 64 – 79.

Bolognani, S., R. Carli, and S. Zampieri (2009, September). A PI consensus controller with gossip communication for clock synchronization in wireless sensors networks. In *1st IFAC Workshop on Distributed Estimation and Control in Networked Systems*, Venice, Italy.

Bolognani, S., S. D. Favero, L. Schenato, and D. Varagnolo (2010, January). Consensus-based distributed sensor calibration and least-square parameter estimation in wireless sensor networks. *International Journal of Robust and Nonlinear Control 20*(2), 176 – 193.

Boucher, R. E. and J. C. Hassab (1981, June). Analysis of discrete implementation of generalized cross-correlator. *IEEE Transactions on Acoustics, Speech, and Signal Processing 29*(3), 609–611.

Boyd, S., A. Ghosh, B. Prabhakar, and D. Shah (2006, June). Randomized gossip algorithms. *IEEE Trans. on Information Theory/ACM Trans. on Networking 52*(6), 2508–2530.

Budianu, C., S. Ben-David, and L. Tong (2006, May). Estimation of the number of operating sensors in large-scale sensor networks with mobile access. *IEEE Transactions on Signal Processing 54*(5), 1703 – 1715.

Chamberland, J.-F. and V. Veeravalli (2004, August). Asymptotic results for decentralized detection in power constrained wireless sensor networks. *IEEE Journal on Selected Areas in Communications 22*(6), 1007 – 1015.

Choi, J., S. Oh, and R. Horowitz (2009). Distributed learning and cooperative control for multi-agent systems. *Automatica 45*(12), 2802 – 2814.

Cohen, E. (1997, December). Size-estimation framework with applications to transitive closure and reachability. *Journal of Computer and System Sciences 55*(3), 441 – 453.

Cortés, J. (2008, March). Distributed algorithms for reaching consensus on general functions. *Automatica 44*(3), 726 – 737.

Cortés, J. (2009, December). Distributed Kriged Kalman filter for spatial estimation. *IEEE Transactions on Automatic Control 54*(12), 2816 –2827.

Craven, P. and G. Wahba (1979). Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik 31*, 377 – 403.

Cucker, F. and S. Smale (2002). On the mathematical foundations of learning. *Bulletin of the American Mathematical Society 39*, 1 – 49.

David, H. A. and H. N. Nagaraja (2003). *Order Statistics*. Wiley series in Probability and Statistics.

D'Costa, A. and A. M. Sayeed (2003). Collaborative signal processing for distributed classification in sensor networks. *Lecture Notes in Computer Science - Information Processing in Sensor Networks 2634*, 558 – 575.

De Nicolao, G. and G. Ferrari-Trecate (1999, November). Consistent identification of NARX models via Regularization Networks. *IEEE Transactions on Automatic Control 44*(11), 2045 – 2049.

De Nicolao, G. and G. Ferrari-Trecate (2001, March). Regularization networks: Fast weight calculation via kalman filtering. *IEEE Transactions on Neural Networks 12*(2), 228 – 235.

Delouille, V., R. Neelamani, and R. Baraniuk (2004, April). Robust distributed estimation in sensor networks using the embedded polygons algorithm. In *Proceedings of the 3rd International Symposium on Information Processing in Sensor Networks*, pp. 405 – 413.

Dogandžić, A. and B. Zhang (2006, August). Distributed estimation and detection for sensor networks using hidden Markov random field models. *IEEE Transactions on Signal Processing 54*(8), 3200 – 3215.

Çetin, M., L. Chen, J. W. Fisher III, A. T. Ihler, R. L. Moses, M. J. Wainwright, and A. Willsky (2006). Distributed fusion in sensor networks - a graphical models perspective. *IEEE Signal Processing Magazine 23*(4), 42 – 55.

Evgeniou, T., M. Pontil, and T. Poggio (2000). Regularization networks and support vector machines. *Advances in Computational Mathematics 13*, 1 – 50.

Fagnani, F. and S. Zampieri (2008a). Asymmetric randomized gossip algorithms for consensus. In *Proceedings of the 17th IFAC world congress*, Seoul, Korea.

Fagnani, F. and S. Zampieri (2008b, May). Randomized consensus algorithms over large scale networks. *IEEE Journal on Selected Areas in Communications 26*(4), 634 – 649.

Feller, W. (1971). *An introduction to probability theory and its application.* Wiley series in Probability and Mathematical Statistics.

Fiacco, A. V. and G. P. Cormick (1968). *Nonlinear Programming: Sequential Unconstrained Minimization Techniques.* John Wiley & Sons.

Garin, F. and L. Schenato (2011). *Networked Control Systems*, Chapter A Survey on distributed estimation and control applications using linear consensus algorithms, pp. 75–107. Springer Lecture Notes in Control and Information Sciences. Springer.

Gelb, A. (1974). *Applied optimal estimation.* Cambridge, MA: MIT Press.

Gilks, W. R., S. Richardson, and D. J. Spiegelhalter (1996). *Markov chain Monte Carlo in Practice.* London: Chapman and Hall.

Girosi, F., M. Jones, and T. Poggio (1995, March). Regularization theory and neural networks architectures. *Neural computation 7*(2), 219 – 269.

Glanzmann, G., R. Negenborn, G. Andersson, B. D. Schutter, and J. Hellendoorn (2007, July). Multi-area control of overlapping areas in power systems for FACTS control. In *Proceedings of Power Tech 2007 (PT 2007)*.

Glaser, S. D. (2004, March). Some real-world applications of wireless sensor nodes. In *SPIE Symposium on Smart Structures and Materials*, San Diego, California.

Good, I. J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika 40*, 237 – 264.

Guestrin, C., P. Bodik, R. Thibaux, M. Paskin, and S. Madden (2004). Distributed regression: an efficient framework for modeling sensor network data. In *Proceedings of the third International Symposium on Information Processing in Sensor Networks (IPSN)*, pp. 1–10.

Hastie, T., R. Tibshirani, and J. Friedman (2001). *The Elements of Statistical Learning*. New York: Springer.

He, T., S. Ben-David, and L. Tong (2006, April). Nonparametric change detection and estimation in large-scale sensor networks. *IEEE Transactions on Signal Processing 54*, 1204–1217.

Hendrickx, J. M., A. Olshevsky, and J. N. Tsitsiklis (2010, April). Distributed anonymous discrete function computation.

Hoerl, A. E. and R. W. Kennard (2000, February). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics 42*(1), 80–86. Special 40th Anniversary Issue.

Honeine, P., M. Essoloh, C. Richard, and H. Snoussi (2008, November - December). Distributed regression in sensor networks with a reduced-order kernel model. In *IEEE Global Telecommunications Conference*, pp. 1 – 5.

Honeine, P., C. Richard, J. Bermudez, H. Snoussi, M. Essoloh, and F. Vincent (2009, April). Functional estimation in Hilbert space for distributed learning in wireless sensor networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 2861–2864.

Howlader, M. S. A., M. R. Frater, and M. J. Ryan (2008). Estimating the number and distribution of the neighbors in an underwater communication network. In *Second International Conference on Sensor Technologies and Applications*, pp. 693–698. IEEE Computer Society.

Huang, J.-L., S.-C. Chid, and X.-M. Huang (2009). GPE: A grid-based population estimation algorithm for resource inventory applications over sensor networks. *Journal of Information Science and Engineering 25*, 201 – 218.

Huang, Y.-D. and M. Barket (1991). On estimating the number of sources with a frequency-hopped signaling sensor array. *IEEE Transactions on Antennas and Propagation 39*, 1384 – 1390.

Huber, P. J. (1964). Robust estimation of a location parameter. *Annals of Statistics 53*, 73 – 101.

Ihler, A. (2005, June). *Inference in sensor networks: Graphical models and particle methods*. Ph. D. thesis, MIT.

Jacovitti, G. and G. Scarano (1993, February). Discrete time techniques for time delay estimation. *IEEE Transactions on Signal Processing 41*(2), 525–533.

Jelasity, M. and A. Montresor (2004). Epidemic-style proactive aggregation in large overlay networks. In *24th International Conference on Distributed Computing Systems*, pp. 102 – 109.

Jenkins, G. M. and D. G. Watts (1969). *Spectral analysis and its applications*. Holden-Day Series in Time Series Analysis. London: Holden-Day.

Kay, S. M. (1993). *Fundamentals of statistical signal processing: estimation theory*. Prentice-Hall.

Kearns, M. and H. S. Seung (1995, February). Learning from a population of hypotheses. *Machine Learning 18*(2-3), 255 – 276.

Kimeldorf, G. and G. Wahba (1971, January). Some results on Tchebycheffian spline functions. *Journal of Mathematical Analysis and Applications 33*(1), 82–95.

Kimeldorf, G. S. and G. Wahba (1970, April). A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *The Annals of Mathematical Statistics 41*(2), 495–502.

König, H. (1986). *Eigenvalue distribution of compact operators*, Volume 9 of *Operator theory: advances and applications*. Basel-Boston-Stuttgart: Birkhauser Verlag.

Kostoulas, D., D. Psaltoulis, I. Gupta, K. Birman, and A. Demers (2005, July). Decentralized schemes for size estimation in large and dynamic groups. In *Fourth IEEE International Symposium on Network Computing and Applications*, pp. 41 – 48.

Kumar, S., F. Zhao, and D. Shephard (2002). Collaborative signal and information processing in microsensor networks. *IEEE Signal Processing Magazine 19*(2), 13 – 14.

Le Merrer, E., A.-M. Kermarrec, and L. Massoulié (2006). Peer to peer size estimation in large and dynamic networks: A comparative study. In *15th IEEE International Symposium on High Performance Distributed Computing*, pp. 7 – 17.

Leshem, A. and L. Tong (2005). Estimating sensor population via probabilistic sequential polling. *IEEE Signal Processing Letters 12*, 395 – 398.

Li, L., J. A. Chambers, C. G. Lopes, and A. H. Sayed (2010, January). Distributed estimation over an adaptive incremental network based on the affine projection algorithm. *IEEE Transactions on Signal Processing 58*(1), 151 – 164.

Ljung, L. (1999). *System identification: theory for the user*. Prentice Hall PTR.

MacKay, D. J. (2003). *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press.

Marple, S. L. (1999). Estimating group delay and phase delay via discrete-time analytic cross-correlation. *IEEE Transactions on Signal Processing 47*(9), 2604–2607.

Martínez, S. (2010, March). Distributed interpolation schemes for field estimation by mobile sensor networks. *IEEE Transactions on Control Systems Technology 18*(2), 491 –500.

Massoulié, L., E. Le Merrer, A.-M. Kermarrec, and A. Ganesh (2006). Peer counting and sampling in overlay networks: random walk methods. In *Proceedings of the twenty-fifth annual ACM symposium on Principles of distributed computing*.

Mateos, G., J. A. Bazerque, and G. B. Giannakis (2010, October). Distributed sparse linear regression. *IEEE Transactions on Signal Processing 58*(10), 5262 – 5276.

Micchelli, C., Y. Xu, and H. Zhang (2006). Universal kernels. *Journal of Machine Learning Research 7*, 2651–2667.

Mosk-Aoyama, D. and D. Shah (2008, July). Fast distributed algorithms for computing separable functions. *IEEE Transactions on Information Theory 54*(7), 2997 – 3007.

Müller, K., A. Smola, G. Rätsch, B. Schölkopf, J. Kohlmorgen, and V. Vapnik (1997). Predicting time series with support vector machines. In *Artificial Neural Networks - ICANN'97*, Volume 1327 of *Lecture Notes in Computer Science*, pp. 999–1004. GMD FIRST Rudower Chaussee 5 12489 Berlin Germany Rudower Chaussee 5 12489 Berlin Germany: Springer Berlin / Heidelberg.

Nakamura, E. F., A. A. F. Loureiro, and A. C. Frery (2007, August). Information fusion for wireless sensor networks: methods, models, and classifications. *ACM Computing Surveys 39*(3), 9 / 1 – 9 / 55.

Nasipuri, A. and S. Tantaratana (1997, March). Nonparametric distributed detection using wilcoxon statistics. *Signal Processing 57*(2), 139–146.

Nef, W. (1967). *Linear Algebra*. McGraw-Hill.

Nguyen, X., M. J. Wainwright, and M. I. Jordan (2005, November). Nonparametric decentralized detection using kernel methods. *IEEE Transactions on Signal Processing 53*(11), 4053 – 4066.

Olfati-Saber, R., J. A. Fax, and R. M. Murray (2007). Consensus and cooperation in multi-agent networked systems. *Proceedings of the IEEE 95*, 215 – 233.

Olfati-Saber, R. and R. M. Murray (2004). Consensus problems in networks of agents with switching topology and time-delays. *IEEE Transactions on Automatic Control 49*(9), 1520–1533.

Oliveira, R. I. (2010, June). Sums of random Hermitian matrices and an inequality by Rudelson. *Electronic Communications in Probability 15*, 203–212.

Papachristodoulou, A., L. Li, and J. C. Doyle (2004, July). Methodological frameworks for large-scale network analysis and design. *ACM SIGCOMM Computer Communication Review 34*(3), 7 – 20.

Pérez-Cruz, F. and S. R. Kulkarni (2010, April). Robust and low complexity distributed kernel least squares learning in sensor networks. *IEEE Signal Processing Letters 17*(4), 355 – 358.

Pillonetto, G. and B. M. Bell (2007, October). Bayes and empirical Bayes semi-blind deconvolution using eigenfunctions of a prior covariance. *Automatica 43*(10), 1698–1712.

Pillonetto, G., A. Chiuso, and G. De Nicolao (2011, February). Prediction error identification of linear systems: A nonparametric gaussian regression approach. *Automatica 47*(2), 291 – 305.

Pillonetto, G. and G. De Nicolao (2010, January). A new kernel-based approach for linear system identification. *Automatica 46*(1), 81 – 93.

Poggio, T. and F. Girosi (1990, September). Networks for approximation and learning. *Proceedings of the IEEE 78*(9), 1481 – 1497.

Poor, H. V. (2009). Competition and collaboration in wireless sensor networks. In *Sensor Networks*, Signals and Communication Technology, pp. 3 – 15. Springer.

Predd, J. B., S. R. Kulkarni, and H. V. Poor (2005). Regression in sensor networks: training distributively with alternating projections. In *Advanced Signal Processing Algorithms, Architectures, and Implementations XV*, Volume SPIE 5910 - 1, San Diego, California.

Predd, J. B., S. R. Kulkarni, and H. V. Poor (2006a, January). Consistency in models for distributed learning under communication constraints. *IEEE Transactions on Information Theory 52*(1), 52 – 63.

Predd, J. B., S. R. Kulkarni, and H. V. Poor (2006b, March). Distributed kernel regression: An algorithm for training collaboratively. In *Proceedings of the IEEE Information Theory Workshop*, pp. 332 – 336.

Predd, J. B., S. R. Kulkarni, and H. V. Poor (2006c, July). Distributed learning in wireless sensor networks. *IEEE Signal Processing Magazine 23*(4), 56 – 69.

Predd, J. B., S. R. Kulkarni, and H. V. Poor (2009, April). A collaborative training algorithm for distributed learning. *IEEE Transactions on Information Theory 55*(4), 1856 – 1871.

Psaltoulis, D., D. Kostoulas, I. Gupta, K. Birman, and A. Demers (2007). Practical algorithms for size estimation in large and dynamic groups.

Puccinelli, D. and M. Haenggi (2005). Wireless sensor networks: applications and challenges of ubiquitous sensing. *IEEE Circuits and Systems Magazine 5*(3), 19 – 31.

Rasmussen, C. E. and C. K. I. Williams (2006). *Gaussian Processes for Machine Learning*. The MIT Press.

Schizas, I. D. and G. B. Giannakis (2006, October - November). Consensus-based distributed estimation of random signals with wireless sensor networks. In *40th Asilomar Conference on Signals, Systems and Computers*, pp. 530 – 534.

Schizas, I. D., A. Ribeiro, and G. B. Giannakis (2008, January). Consensus in ad hoc WSNs with noisy links - part I: Distributed estimation of deterministic signals. *IEEE Transactions on Signal Processing 56*(1), 350 – 364.

Schölkopf, B. and A. Smola (2001). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge, MA, USA: MIT Press.

Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics 6*(2), 461 – 464.

Simić, S. N. (2003). A learning theory approach to sensor networks. *IEEE Pervasive Computing 2*(4), 44 – 49.

Smale, S. and D.-X. Zhou (2005). Shannon sampling II: Connections to learning theory. *Applied and Computational Harmonic Analysis 19*, 285–302.

Smale, S. and D.-X. Zhou (2007). Learning theory estimates via integral operators and their approximations. *Constructive approximation 26*, 153–172.

Snoussi, H. and C. Richard (2006, November). Distributed Bayesian fault diagnosis in collaborative wireless sensor networks. In *IEEE Global Telecommunications Conference*.

Söderström, T. and P. Stoica (1989). *System Identification*. Prentice Hall.

Stein, M. L. (1999). *Interpolation of spatial data: some theory for Kriging*. Springer.

Sudderth, E. B., A. T. Ihler, W. T. Freeman, and A. S. Willsky (2003, June). Nonparametric belief propagation. *IEEE Conference on Computer Vision and Pattern Recognition 1*, 605 – 612.

Tikhonov, A. N. and V. Y. Arsenin (1977). *Solution of Ill-posed Problems*. Wiston.

Tropp, J. A. (2004, October). Greed is good: algorithmic results for sparse approximation. *IEEE Transactions on Information Theory 50*, 2231 – 2242.

Vapnik, V. N. (1995). *The nature of statistical learning theory*. New York: Springer.

Varagnolo, D., G. Pillonetto, and L. Schenato (2009, December). Distributed function and time delay estimation using nonparametric techniques. In *IEEE Conference on Decision and Control*, pp. 7608 – 7613.

Varagnolo, D., G. Pillonetto, and L. Schenato (2010a, June - July). Distributed consensus-based bayesian estimation: sufficient conditions for performance characterization. In *American Control Conference*, pp. 3986 – 3991.

Varagnolo, D., G. Pillonetto, and L. Schenato (2010b, December). Distributed statistical estimation of the number of nodes in sensor networks. In *IEEE Conference on Decision and Control*.

Varshney, P. K. (1996). *Distributed Detection and Data Fusion*. Secaucus, NJ, USA: Springer-Verlag New York, Inc.

Viola, F. and W. F. Walker (2005, January). A spline-based algorithm for continuous time-delay estimation using sampled data. *IEEE Transactions on Ultrasonics, Ferroelectrict, and Frequency Control 52*(1), 80–93.

Viswanathan, R. and P. K. Varshney (1997, January). Distributed detection with multiple sensors i: Fundamentals. *Proceedings of the IEEE 85*, 64 – 63.

Wahba, G. (1990). *Spline models for observational data*. SIAM.

Wang, P., H. Li, and J. Fang (2008, March). Distributed non-parametric estimation in a bandwidth-constrained sensor network. In *42nd Annual Conference on Information Sciences and Systems*, pp. 1031 – 1036.

Weinert, H. L. (1982). *Reproducing Kernel Hilbert Spaces: Applications in Statistical Signal Processing*. Stroudsburg, Pennsylvania: Hutchinson Ross.

Xiao, J.-J., A. Ribeiro, Z.-Q. Luo, and G. Giannakis (2006, July). Distributed compression-estimation using wireless sensor networks. *IEEE Signal Processing Magazine 23*(4), 27 – 41.

Yaglom, A. M. (1987). *Correlation theory of stationary and related random functions*, Volume 1. New York: Springer.

Yamanishi, K. (1997). Distributed cooperative Bayesian learning strategies. In *COLT '97: Proceedings of the tenth annual conference on Computational learning theory*, New York, NY, USA, pp. 250 – 262. ACM.

Yosida, K. (1965). *Functional Analysis*, Volume 123. Springer-Verlag.

Zheng, H., S. R. Kulkarni, and H. V. Poor (2008, June - July). Dimensionally distributed learning models and algorithm. In *11th International Conference on Information Fusion*, pp. 1 – 8.

Zhu, H. and R. Rohwer (1996, September). Bayesian regression filters and the issue of priors. *Journalof Neural Computing and Applications 4*(3), 130–142.

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

Zhu, K. (2007). *Operator theory in function spaces*. Number 138 in Mathematical Surveys and Monographs. American Mathematical Society.