Distributed consensus-based Bayesian estimation: sufficient conditions for performance characterization

Damiano Varagnolo, Gianluigi Pillonetto, Luca Schenato

***Abstract*— The paper considers the framework of distributed Bayesian linear estimation. We introduce some consensus-based estimation strategies that are equivalent to centralized ones pending knowledge of some parameters, e.g. number of agents in the network. If such parameters are not known, agents can estimate them locally or exploit prior knowledge. We show that in this case the performance depends on parameter uncertainty in such a way that, in case of large errors, the distributed estimator can perform worse than the local one. Then, we find some sufficient conditions on the error magnitude which ensure that the distributed scheme behaves better than the local one.**

***Index Terms*— Bayesian linear model, distributed estimation, consensus, performance characterization, sufficient conditions**

I. Introduction

The continuous growth of large scale networks of devices which are capable of sensing and interacting with the environment, commonly referred as Networked Control Systems (NCSs), is enabling a whole new range of applications ranging from ambient monitoring using wireless sensor networks to surveillance using networks of smart cameras, from multi-robot exploration to energy management using smart grids, just to name a few [1]. However, these new applications come with great challenges since the design of a large scale network of cooperating systems, is still at an empirical level and sound methodological strategies are only recently appearing [2].

Within this context, in this paper we address the problem of distributed estimation, i.e. the problem of estimating a set of unknown system parameters based on the measurements obtained from many sensor nodes. These nodes are provided with computational and communication capabilities and their objective is to obtain an estimate of the unknown parameters possibly through cooperation. We also consider a framework in which there is no central coordinating unit and sensor nodes form a connected network, i.e. they might not be able to communicate directly, but there is path of that allows information to travel from any node to any other node. An example of such a system is given by the next generation power grids [3] where each energy producer or user will be connected through a communication network and can exchange information to estimate some unknown parameters of the network like its efficiency, its capacity, its current utilization, etc. These networks are likely to be dynamic, i.e. new nodes can appear or disappear, and the nodes characteristic might not be known to any node. These features make distributed estimation challenging since it is necessary to design algorithms that do not rely on the a-priori knowledge of the network topology and network parameters, and need to be robust to node failure and dynamic changes.

Distributed estimation and, more generally, distributed computation is a well established research areas [4] [5], in particular in the context of computer networks. However it has been witnessing a renewed interest mainly due to the appearance of new technologies which pose new challenges to these old problems, like lossy communication, bandwidth limitation, energy constraints, unreliable devices [6] [7] [8] (and references therein). Even standard problems in control theory, like Bayesian estimation of linear systems [9], have been shown to be nontrivial in the context of Wireless Sensor Networks mainly due to limited computational and communication resources available to the network nodes [10] [11]. In the framework of Bayesian estimation, several authors focused on distributed or decentralized computations. For example, in [12] 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 [13] 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 [14] authors propose a message-passing scheme for a nonparametric distributed regression algorithm, while in [15] they survey the problems related to the distribution of the learning process in wireless sensor networks, analyzing both parametric and nonparametric scenarios. In [16] 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 performances of decision transmission strategies, seeking for optimality in terms of decision error probability for the central unit [17].

Recently, popular distributed algorithms, known as consensus algorithms [18], have been proposed also for estimation purposes [19] [20] since they require only limited

This work was supported by the Italian CaRiPaRo Foundation under the Wise-Wai funded project and by European Community's Seventh Framework Programme under agreement n. FP7-ICT-223866 - FeedNetBack.

Damiano Varagnolo, Gianluigi Pillonetto and Luca Schenato are with the Department of Information Engineering, University of Padova, Via Gradenigo, 6/b, Padova, Italy. E-mails: {varagnolo | giapi | schenato}@dei.unipd.it

computation and communication resources, minimal node synchronization, and they are robust to link and node failure. The main idea of consensus algorithms is to average measurements or local estimates among all network nodes, based on the intuition that averaging reduces the noise and therefore the parameter estimation error. However, this is not always the case, since local estimates are correlated and a simple average might not be the optimal strategy, in particular when sensors have different accuracy.

In this work we address this problem. In particular we analyze when averaging leads to better performance than a local estimate under different noise conditions and when the network nodes do not know the number of sensors in the network or their noise levels. We show that indeed performance improvement is not always guaranteed by consensus and we provide some sufficient conditions which guarantee it under mild conditions.

II. Problem statement

In this section, first, we introduce three different scenarios of Bayesian estimation. Then, we state the main problem which will be investigated in the subsequent sections.

A. Local Bayesian estimation

Let

$$\mathbf{y}_i = C\mathbf{a} + \nu_i, \qquad i = 1, \dots, S \tag{1}$$

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

Under the assumptions above, the local Minimum Mean Square Error (MMSE) estimator of $\mathbf{a}$ given $\mathbf{y}_i$, is

$$\begin{aligned} \widehat{\mathbf{a}}_{\text{loc},i} &:= \mathbb{E}\left[\mathbf{a} \mid \mathbf{y}_i\right] = \operatorname{cov}(\mathbf{a}, \mathbf{y}_i)\, \operatorname{var}(\mathbf{y}_i)^{-1} \mathbf{y}_i \\ &= \Sigma_{\mathbf{a}} C^T \left(V\left(\sigma_i^2\right)\right)^{-1} \mathbf{y}_i \,. \end{aligned} \tag{2}$$

where, for $\theta \in \mathbb{R}$:

$$V(\theta) := C\Sigma_{\mathbf{a}} C^T + \theta I_M \tag{3}$$

The autocovariance of the local estimation error $\widetilde{\mathbf{a}}_{\text{loc},i} := \mathbf{a} - \widehat{\mathbf{a}}_{\text{loc},i}$ is given by:

$$\operatorname{var}\left(\widetilde{\mathbf{a}}_{\text{loc},i}\right) = \Sigma_{\mathbf{a}} - \Sigma_{\mathbf{a}} C^T \left(V\left(\sigma_i^2\right)\right)^{-1} C\Sigma_{\mathbf{a}} \,. \tag{4}$$

B. Centralized Bayesian estimation

If $S \geq 2$ and all measurements $\{\mathbf{y}_i\}$ are collected by a central unit, the MMSE estimate of the parameter vector $\mathbf{a}$ can be computed in a centralized way via the following:

$$\widehat{\mathbf{a}}_{\text{cent}} := \operatorname{cov}\left(\mathbf{a}, \begin{bmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_S \end{bmatrix}\right) \operatorname{var}\left(\begin{bmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_S \end{bmatrix}\right)^{-1} \begin{bmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_S \end{bmatrix} \tag{5}$$

where:

$$\operatorname{var}\left(\begin{bmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_S \end{bmatrix}\right) = \begin{bmatrix} V\left(\sigma_1^2\right) & \dots & V(0) \\ \vdots & & \vdots \\ V(0) & \dots & V\left(\sigma_S^2\right) \end{bmatrix} \tag{6}$$

Using the matrix inversion lemma and simple algebraic manipulations, eqn. (5) can be rewritten as:

$$\widehat{\mathbf{a}}_{\text{cent}} = \Sigma_{\mathbf{a}} C^T \left(V(\alpha^{-1})\right)^{-1} \frac{\sum_{i=1}^S \frac{\mathbf{y}_i}{\sigma_i^2}}{\sum_{i=1}^S \frac{1}{\sigma_i^2}} \,. \tag{7}$$

where above, and in the sequel:

$$\alpha := \sum_{i=1}^S \frac{1}{\sigma_i^2} \,. \tag{8}$$

C. Distributed Bayesian estimation

Let's assume that no central units can collect the whole data set and (in this section) that all sensors know the value of the sum of the precisions α *before* starting the various estimation strategies.

If the number of measurements for each sensor M is smaller than the number of parameters E, then, since:

$$\frac{\sum_{i=1}^S \frac{\mathbf{y}_i}{\sigma_i^2}}{\sum_{i=1}^S \frac{1}{\sigma_i^2}} = \frac{\frac{1}{S}\sum_{i=1}^S \frac{\mathbf{y}_i}{\sigma_i^2}}{\frac{1}{S}\sum_{i=1}^S \frac{1}{\sigma_i^2}} \tag{9}$$

sensors can reach average-consensus on $\frac{\mathbf{y}_i}{\sigma_i^2}$ and $\frac{1}{\sigma_i^2}$, then compute their ratio (9), and then compute the estimate (7), while if $M > E$, it is preferable to use algorithm 1. Since in the practical case it is easier to have $M \gg E$, we will refer through the paper to this situation.

Algorithm 1 distributed estimation with known sum of precisions

1: (requirement) sensors have knowledge of the quantity α before starting step 2
2: sensors achieve average consensus on the quantities $\Sigma_{\mathbf{a}} C^T \left(V\left(\alpha^{-1}\right)\right)^{-1} \frac{\mathbf{y}_i}{\sigma_i^2}$ and $\frac{1}{\sigma_i^2}$;
3: at the end of the consensus process, sensors compute the optimal estimate dividing the two obtained averaged variables.

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

$$h := H\left(\sigma_1^2, \dots, \sigma_S^2\right) := \frac{S}{\sum_{i=1}^S \frac{1}{\sigma_i^2}} \,. \tag{10}$$

It is evident that average consensus on the quantities $\frac{1}{\sigma_i^2}$ corresponds to a distributed estimation of h^{-1}. Since:

$$\alpha := \sum_{i=1}^S \frac{1}{\sigma_i^2} = \frac{S}{h} \,, \tag{11}$$

once the number of sensors in the network S is known, it is possible to satisfy requirement 1 of algorithm 1 using a pre-distributed estimation step for h^{-1}.

D. Problem statement

We are now in a position to formulate the main problem investigated in this paper. Assume that α, S and h are unknown (plausible situation when sensors do not have knowledge on the whole network). Then sensors have two possible strategies: the first is to make a guess $\overline{\alpha}$ of α, use it in step 2 of algorithm 1, run in parallel two average-consensi (one on $\Sigma_{\mathbf{a}} C^T \left(V\left(\overline{\alpha}^{-1}\right)\right)^{-1} \frac{\mathbf{y}_i}{\sigma_i^2}$ and one on $\frac{1}{\sigma_i^2}$) and obtain a (suboptimal) estimate. The second is firstly to distributely estimate h with an average consensus on $\frac{1}{\sigma_i^2}$, then make a guess $\overline{S}$ of S (the same among all sensors), then reach average consensus on $\Sigma_{\mathbf{a}} C^T \left(V\left(\frac{h}{\overline{S}}\right)\right)^{-1} \frac{\mathbf{y}_i}{\sigma_i^2}$ and finally divide the result for h. The first strategy is faster, while the second requires a guess on a quantity that is more easily deducible. Notice also that, as previously discussed, when the number of measurements per sensor is smaller than the number of parameters, it is convenient first to obtain $\frac{1}{S}\sum_i \frac{\mathbf{y}_i}{\sigma_i^2}$ and $\frac{1}{S}\sum_i \frac{1}{\sigma_i^2}$ via consensus and then apply the transformation matrix $\Sigma_{\mathbf{a}} C^T \left(V\left(\frac{h}{\overline{S}}\right)\right)^{-1}$ so that, subsequently, in this case the problem is always reduced to make a guess on S.

We would like now to derive conditions that guarantee that the process of sharing and combining the information described by these suboptimal versions of algorithm 1 improves the estimation of $\mathbf{a}$ with respect to the local estimation strategy of eqn. (2). In other words, we want to 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.

III. An uniform sufficient condition

We start the analysis considering the first suboptimal strategy, where sensors make a (common) guess $\overline{\alpha}$ of α and then run 2 parallel average-consensi. The second strategy will be considered in sec. III-C.

Using a guess $\overline{\alpha}$, at the end of the consensus process the distributed estimate result is:

$$\widehat{\mathbf{a}}_{\text{dist}}\left(\overline{\alpha}\right) := \Sigma_{\mathbf{a}} C^T \left(V\left(\overline{\alpha}^{-1}\right)\right)^{-1} \frac{\frac{1}{S}\sum_{i=1}^{S} \frac{\mathbf{y}_i}{\sigma_i^2}}{\frac{1}{S}\sum_{i=1}^{S} \frac{1}{\sigma_i^2}} . \tag{12}$$

Obviously the variance of the estimation error of $\widehat{\mathbf{a}}_{\text{cent}}$ is smaller than the one of $\widehat{\mathbf{a}}_{\text{dist}}\left(\overline{\alpha}\right)$. An interesting question is: can we find values of $\overline{\alpha}$ s.t. at the end of the estimation process the variance of the error of the distributed strategy is smaller than the error of the local strategy, independently of $\Sigma_{\mathbf{a}}$ and on C? The answer is in the following:

Theorem 1. If

$$\overline{\alpha} \in \left[\alpha - \sqrt{\alpha^2 - \frac{\alpha}{\sigma_i^2}},\ \alpha + \sqrt{\alpha^2 - \frac{\alpha}{\sigma_i^2}}\right] \tag{13}$$

then the variance of the estimation error of the distributed estimator $\widehat{\mathbf{a}}_{\text{dist}}\left(\overline{\alpha}\right)$ is smaller than the one of the local estimator $\widehat{\mathbf{a}}_{\text{loc},i}$, for every prior $\Sigma_{\mathbf{a}}$, number of parameters E, sum of precisions α and matrix C.

Notice that even if $\overline{\alpha}$ is assumed to be the same among all the sensors, the bound (13) is different for each sensor i.

A. Asymptotic analysis of bound (13)

Before deriving other results it is interesting to analyze the asymptotic behavior of bound (13). For ease of notation we define $b_-\left(i\right) := \alpha - \sqrt{\alpha^2 - \frac{\alpha}{\sigma_i^2}}$ and $b_+\left(i\right) := \alpha + \sqrt{\alpha^2 - \frac{\alpha}{\sigma_i^2}}$:

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

$$\exists j \text{ s.t. } \sigma_j^2 \to 0 \quad \Rightarrow \quad b_-\left(i\right) \to \frac{1}{2\sigma_i^2}, \quad b_+\left(i\right) \to +\infty \tag{14}$$

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

$$\forall j\ \sigma_j^2 \to +\infty \quad \Rightarrow \quad b_-\left(i\right) \to \frac{1}{\sigma_i^2}, \quad b_+\left(i\right) \to \frac{1}{\sigma_i^2}, \tag{15}$$

i.e. if all the sensors have unreliable measures then sensor i should use the local estimator (2);

• if the noisiness of all the sensors are the same but we vary the number of sensors S in the network, we have that:

$$S \to +\infty \quad \Rightarrow \quad b_-\left(i\right) \to 0 \quad b_+\left(i\right) \to +\infty \tag{16}$$

but we send back the reader to Sec. V for a more detailed discussion of this case;

• 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:

$$\sigma_i^2 \to 0 \quad \Rightarrow \quad b_-\left(i\right) \to +\infty, \quad b_+\left(i\right) \to +\infty \tag{17}$$

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 $\overline{\alpha}$. Unfortunately from bound (13) we obtain only the following:

$$\sigma_i^2 \to +\infty \quad \Rightarrow \quad b_-\left(i\right) \to 0, \quad b_+\left(i\right) \to 2\alpha \tag{18}$$

i.e. a subset of the interval we were expecting. This is due to the fact that thm. 1 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 (13) is thight for the sensor i (the accurate one), so it is more probable that the guessed $\overline{\alpha}$ falls outside of its bound. Since (13) is a sufficient condition, it could be that, if $\overline{\alpha}$ 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.

B. Conditions referred to the network as a whole

The following condition assures that each sensor in the network has an advantage from the distributed algorithm:

Corollary 2. Define $\sigma_{\min}^2 := \min_i \left\{\sigma_i^2\right\}$. Then if

$$\overline{\alpha} \in \left[\alpha - \sqrt{\alpha^2 - \frac{\alpha}{\sigma_{\min}^2}},\ \alpha + \sqrt{\alpha^2 - \frac{\alpha}{\sigma_{\min}^2}}\right] \tag{19}$$

then the variance of the estimation error of the distributed estimator $\widehat{\mathbf{a}}_{\text{dist}}(\overline{\alpha})$ is smaller than the one of the local estimator $\widehat{\mathbf{a}}_{\text{loc},i}$ for each sensor i.

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 $\overline{\alpha}$ 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 $\Sigma_{\mathbf{a}}$ and of the matrix C? The answer is given in the following:

Theorem 3. Considering the harmonic mean h defined in Eqn. (10), if

$$\overline{\alpha} \in \left[\alpha - \sqrt{\alpha^2 - \frac{\alpha}{h}},\ \alpha + \sqrt{\alpha^2 - \frac{\alpha}{h}}\right] \tag{20}$$

then the variance of the estimation error of the distributed estimator $\widehat{\mathbf{a}}_{\text{dist}}(\overline{\alpha})$ is smaller than the average variance of the estimation errors of the local estimators $\widehat{\mathbf{a}}_{\text{loc},i}$.

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 (19) is always included in the interval described in bound (20), implying that condition (19) is sufficient for condition (20).

C. Uniform conditions when knowing the harmonic mean of the measurements noises

If sensors use the second suboptimal strategy of sec. II-D (composed by a distributed estimation of h, a guess $\overline{S}$ of S and subsequently an average consensus), the previous results can be immediately reformulated as follows:

Corollary 4. If

$$\overline{S} \in \left[S - \sqrt{S^2 - \frac{Sh}{\sigma_i^2}},\ S + \sqrt{S^2 - \frac{Sh}{\sigma_i^2}}\right] \tag{21}$$

then the variance of the estimation error of the distributed estimator $\widehat{\mathbf{a}}_{\text{dist}}(\overline{\alpha})$ is smaller than the one of the local estimator $\widehat{\mathbf{a}}_{\text{loc},i}$, for every prior $\Sigma_{\mathbf{a}}$, number of parameters E, sum of precisions α and matrix C.

Corollary 5. If

$$\overline{S} \in \left[S - \sqrt{S^2 - \frac{Sh}{\sigma_{\min}^2}},\ S + \sqrt{S^2 - \frac{Sh}{\sigma_{\min}^2}}\right] \tag{22}$$

then the variance of the estimation error of the distributed estimator $\widehat{\mathbf{a}}_{\text{dist}}(\overline{\alpha})$ is smaller than the one of the local estimator $\widehat{\mathbf{a}}_{\text{loc},i}$ for each sensor i.

Corollary 6. If

$$\overline{S} \in \left[S - \sqrt{S^2 - S},\ S + \sqrt{S^2 - S}\right] \tag{23}$$

then the variance of the estimation error of the distributed estimator $\widehat{\mathbf{a}}_{\text{dist}}(\overline{\alpha})$ is smaller than the average variance of the estimation errors of the local estimators $\widehat{\mathbf{a}}_{\text{loc},i}$.

Notice that corollary 6 is not independent of the various noises variances σ_i^2 since it implicitly requires the knowledge on their harmonic mean h.

IV. Equal measurements noises variances

The special case $\sigma_i^2 = \sigma^2$ for all i's is interesting because it corresponds to networks composed by the same type of sensors. In this case we have that:

$$\alpha := \sum_{i=1}^{S} \frac{1}{\sigma_i^2} = \frac{S}{\sigma^2} \tag{24}$$

and the centralized estimator can be rewritten as:

$$\widehat{\mathbf{a}}_{\text{cent}} := \Sigma_{\mathbf{a}} C^T \left(V\left(\frac{\sigma^2}{S}\right)\right)^{-1} \left(\frac{1}{S}\sum_{i=1}^{S} \mathbf{y}_i\right) . \tag{25}$$

The estimation error variance in this case is optimal and given by:

$$\text{var}\left(\widetilde{\mathbf{a}}_{\text{cent}}\right) = \Sigma_{\mathbf{a}} - \Sigma_{\mathbf{a}} C^T \left(V\left(\frac{\sigma^2}{S}\right)\right)^{-1} C\Sigma_{\mathbf{a}} . \tag{26}$$

Eqn. (25) can be distributely computed using algorithm 2. The main differences with algorithm 1 and the different noises scenario of sec. II-C are:

- sensors must know the exact number of active sensors S and not the sum of precisions α;
- sensors need to achieve consensus only to one vectorial quantity (consensus on the precisions is not needed);
- even if the steady state has not been reached, the quantities involved in the consensus algorithm are estimations of the parameters $\mathbf{a}$.

Algorithm 2 distributed estimation with equal noises and known number of sensors

1: each sensor i locally computes an initial estimate of the parameters $\Sigma_{\mathbf{a}} C^T \left(V\left(\frac{\sigma^2}{S}\right)\right)^{-1} \mathbf{y}_i$;
2: sensors achieve average consensus on the previous quantities.

A. Estimation without perfect knowledge on S

In order to achieve the optimal performances, the various sensors *must* use the correct parameter $\frac{\sigma^2}{S}$ in step 1 of algorithm 2, thus all sensors must have perfect knowledge on S. But what happens when this is infeasible? Assume all sensors use in step 1 of this algorith a certain guess $\overline{S}$ (the same among all the sensors), instead of the correct parameter S. The resulting distributed estimator is now dependent on this new parameter $\overline{S}$:

$$\begin{aligned}\widehat{\mathbf{a}}_{\text{dist}}\left(\overline{S}\right) &:= \frac{1}{S}\sum_{i=1}^{S}\left(\Sigma_{\mathbf{a}} C^T \left(V\left(\frac{\sigma^2}{\overline{S}}\right)\right)^{-1} \mathbf{y}_i\right) \\ &= \Sigma_{\mathbf{a}} C^T \left(V\left(\frac{\sigma^2}{\overline{S}}\right)\right)^{-1} \left(\frac{1}{S}\sum_{i=1}^{S}\mathbf{y}_i\right) .\end{aligned} \tag{27}$$

The new estimation error, defined as $\widetilde{\mathbf{a}}_{\text{dist}}\left(\overline{S}\right) := \mathbf{a} - \widehat{\mathbf{a}}_{\text{dist}}\left(\overline{S}\right)$, has a variance equal to:

$$\begin{aligned}\text{var}\left(\widetilde{\mathbf{a}}_{\text{dist}}\left(\overline{S}\right)\right) = \Sigma_{\mathbf{a}} - 2\Sigma_{\mathbf{a}} C^T \left(V\left(\frac{\sigma^2}{\overline{S}}\right)\right)^{-1} C\Sigma_{\mathbf{a}} + \\ + \Sigma_{\mathbf{a}} C^T \left(V\left(\frac{\sigma^2}{\overline{S}}\right)\right)^{-1} V\left(\frac{\sigma^2}{S}\right) \left(V\left(\frac{\sigma^2}{\overline{S}}\right)\right)^{-1} C\Sigma_{\mathbf{a}}\end{aligned} \tag{28}$$

that obviously is equal to expression (26) whenever $\overline{S} = S$.

V. An uniform sufficient condition for equal measurements noises variances

As before, we are interested to understand what happens when the various sensors use in step 1 of algorithm 2 not the exact number of sensors S but a guess $\overline{S}$ (the same among the network). Using eqn. (24) we can reformulate bound (13) for the current case, and obtain the following bound:

$$\overline{\alpha} \in \left[\frac{1}{\sigma^2}\left(S - \sqrt{S^2 - S}\right),\ \frac{1}{\sigma^2}\left(S + \sqrt{S^2 - S}\right)\right] \quad (29)$$

that is dependent on the measurement noise variance σ^2. Using a different proof (not suitable for thm. 1) we can remove this dependence and obtain a more elegant result:

Theorem 7. If

$$\overline{S} \in [1, 2(S-1)] \quad (30)$$

then the variance of the estimation error of the distributed estimator $\widehat{\mathbf{a}}_{\text{dist}}\left(\overline{S}\right)$ is smaller than the one of the local estimators $\widehat{\mathbf{a}}_{\text{loc},i}$, for every prior $\Sigma_{\mathbf{a}}$, number of parameters E, measurement noise variance σ^2, matrix C and sensor i.

Notice that if $\overline{S} = 1$ then equation (12) reduces to the local strategy (2). For this reason thm. 7 assures that local estimation (2) plus average consensus is, at the end of the consensus process, always better than local MMSE estimate, independently of S, σ^2, $\Sigma_{\mathbf{a}}$ and C.

VI. Some non uniform sufficient conditions for equal measurements noises variances

Considering still the case $\sigma_i^2 = \sigma^2$, assuming the knowledge of $C\Sigma_{\mathbf{a}}C^T$ (or equivalently on its eigenvalues d_m), it is possible to enlarge bound (30) and find some other interesting properties.

First of all, there could be networks (i.e. S and σ^2) where, no matter how the guess $\overline{S}$ is chosen, distributed estimation leads to a smaller error variance than the local one:

Proposition 8. If $d_{\min}$ is the smallest eigenvalue of $C\Sigma_{\mathbf{a}}C^T$ and if

$$d_{\min} > \frac{\sigma^2}{S-1} \quad (31)$$

then the variance of the estimation error of the distributed estimator $\widehat{\mathbf{a}}_{\text{dist}}\left(\overline{S}\right)$ is smaller than the one of the local estimators $\widehat{\mathbf{a}}_{\text{loc},i}$, for every sensor i and guess $\overline{S} \in [1, +\infty)$.

In this case, the distributed estimator behave better than the local one *also* assuming $\overline{S} = +\infty$, 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 (31). The statistical requirement of proposition 8 is that the smallest eigenvalue of $C\Sigma_{\mathbf{a}}C^T$ has to dominate the resulting noise of the *averaged* measurements.

If S and σ^2 are s.t. proposition 8 is not satisfied, then we can state (as an intermediate consequence of the proof of Thm. 7) the following:

Corollary 9. Define:

$$\widehat{d}(S) := \min_{m \in \{1,\ldots,M\}} \left\{d_m \text{ s.t. } \sigma^2 + (1-S)\, d_m > 0\right\} \quad (32)$$

and:

$$\mathcal{S}_{\min}(S) := \frac{\sigma^2 S + \sqrt{\sigma^2 S(S-1)\left(\sigma^2 + \widehat{d}(S)\right)}}{\sigma^2 + (1-S)\,\widehat{d}(S)}. \quad (33)$$

If

$$\overline{S} \in [1, 2(\mathcal{S}_{\min}(S) - 1)] \quad (34)$$

then the variance of the estimation error of the distributed estimator $\widehat{\mathbf{a}}_{\text{dist}}\left(\overline{S}\right)$ is smaller than the one of the local estimators $\widehat{\mathbf{a}}_{\text{loc},i}$, for every prior $\Sigma_{\mathbf{a}}$, number of parameters E, measurement noise variance σ^2, matrix C and sensor i.

VII. Conclusions

In this work we studied general consensus-based distributed Bayesian estimation scheme useful for networks of smart sensors whose number and noise characteristics are only partially known. In particular, we have derived mild sufficient conditions on the system parameters ensuring that the error variances affecting the estimates obtained by each node after consensus are smaller than those affecting the local estimates, i.e. the ones obtained by a sensor using only its sensor data.

This is a preliminary work to analyze consensus-based distributed estimation performance in terms of robustness to system parameters and characteristics. Future work also includes the extension to sensors with different observation models, i.e. $C = C_i$, to measurements with correlated noise, i.e. $\mathbb{E}\left[\nu_{i,m}{\nu_{j,n}}^T\right] \neq 0$, and to non-parametric function estimation where the parameter vector a is replaced by an unknown infinite-dimensional function.

Appendix

Lemma 10. If $a_i \geq 0$, $i = 1, \ldots, S$ and $b \geq 0$, then:

$$H(b + a_1, \ldots, b + a_S) \geq b + H(a_1, \ldots, a_S) \quad (35)$$

Proof. Defining:

$$f(b) := H(b + a_1, \ldots, b + a_S) - H(a_1, \ldots, a_S) - b \quad (36)$$

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

$$S\sum_{i=1}^{S}\left(\frac{1}{b+a_i}\right)^2 \geq \left(\sum_{i=1}^{S}\frac{1}{b+a_i}\right)^2. \quad (37)$$

Considering the two vectors $x = \left[\frac{1}{b+a_1}, \ldots, \frac{1}{b+a_S}\right]^T$ and $y = [1, \ldots, 1]^T$, condition (37) corresponds to $\langle x, x\rangle \langle y, y\rangle \geq |\langle x, y\rangle|^2$ that is the well-known Cauchy-Schwarz inequality. □

Proof. (of thm. 1) Let us introduce the orthogonal matrix U that diagonalizes the first addendum of matrix (3):

$$C\Sigma_{\mathbf{a}}C^T = UDU^T \ . \tag{38}$$

Since $UU^T = I_M$, we can also write:

$$\begin{aligned} V(\theta) &:= C\Sigma_{\mathbf{a}}C^T + \theta I_M \\ &= UDU^T + \theta UU^T \\ &= U(D + \theta I_M)U^T \ . \end{aligned} \tag{39}$$

Now we are looking for the set of $\overline{\alpha}$ s.t.:

$$\text{var}(\mathbf{a} - \widehat{\mathbf{a}}_{\text{dist}}(\overline{\alpha})) \leq \ \text{var}(\mathbf{a} - \widehat{\mathbf{a}}_{\text{loc},i}) \tag{40}$$

where:

$$\begin{aligned} \text{var}(\mathbf{a} - \widehat{\mathbf{a}}_{\text{dist}}(\overline{\alpha})) = \Sigma_{\mathbf{a}} - 2\Sigma_{\mathbf{a}}C^T\left(V\left(\tfrac{1}{\overline{\alpha}}\right)\right)^{-1}C\Sigma_{\mathbf{a}} \\ +\Sigma_{\mathbf{a}}C^T\left(V\left(\tfrac{1}{\overline{\alpha}}\right)\right)^{-1}V\left(\tfrac{1}{\alpha}\right)\left(V\left(\tfrac{1}{\overline{\alpha}}\right)\right)^{-1}C\Sigma_{\mathbf{a}} \end{aligned} \tag{41}$$

and:

$$\text{var}(\mathbf{a} - \widehat{\mathbf{a}}_{\text{loc},i}) = \Sigma_{\mathbf{a}} - \Sigma_{\mathbf{a}}C^T\left(V\left(\sigma_i^2\right)\right)^{-1}C\Sigma_{\mathbf{a}} \ . \tag{42}$$

Using the fact that for a generic matrix B then $A \leq 0 \Rightarrow BAB^T \leq 0$, we can derive a sufficient condition assuring (40):

$$\begin{aligned} -2\left(V\left(\tfrac{1}{\overline{\alpha}}\right)\right)^{-1} + \left(V\left(\tfrac{1}{\overline{\alpha}}\right)\right)^{-1}V\left(\tfrac{1}{\alpha}\right)\left(V\left(\tfrac{1}{\overline{\alpha}}\right)\right)^{-1} \leq \\ \leq -\left(V\left(\sigma_i^2\right)\right)^{-1} \ . \end{aligned} \tag{43}$$

Diagonalizing the various $V(\theta)$'s in eqn. (43) we obtain:

$$\begin{aligned} &-2U\left(D + \tfrac{1}{\overline{\alpha}}I_M\right)^{-1}U^T + \\ &+U\left(D + \tfrac{1}{\overline{\alpha}}I_M\right)^{-2}\left(D + \tfrac{1}{\alpha}I_M\right)U^T \leq \\ &\leq -U\left(D + \sigma_i^2 I_M\right)^{-1}U^T \end{aligned} \tag{44}$$

where we also used the fact that diagonal matrices commute. It is easy to show that for orthogonal matrices U we have that $A \leq 0 \Leftrightarrow UAU^T \leq 0$, so if we remove all the U's from equation (44) we still have a sufficient condition for inequality (40). Now all the remaining matrices are diagonal, so the condition is satisfied as soon as it is satisfied component by component, so once we define $d_m := D_{mm}$, with $m = 1, \ldots, M$, the sufficient condition is:

$$\frac{-2}{d_m + \frac{1}{\overline{\alpha}}} + \frac{d_m + \frac{1}{\alpha}}{\left(d_m + \frac{1}{\overline{\alpha}}\right)^2} \leq \frac{-1}{d_m + \sigma_i^2} \quad \forall m. \tag{45}$$

Note that each d_m is an eigenvalue of $C\Sigma_{\mathbf{a}}C^T$, thus it is $d_m \geq 0$ for all m's since $\Sigma_{\mathbf{a}}$ is at least semi-positive definite. Now condition (45) can be rewritten as:

$$p_{i,m}(\overline{\alpha}) := \left(\sigma_i^2 + \left(1 - \alpha\sigma_i^2\right)d_m\right)\overline{\alpha}^2 - \left(2\alpha\sigma_i^2\right)\overline{\alpha} + \alpha \leq 0 \tag{46}$$

for all m. Notice that:

$$\alpha\sigma_i^2 = \sum_{j=1}^{S}\frac{\sigma_i^2}{\sigma_j^2} = 1 + \sum_{j\neq i}\frac{\sigma_i^2}{\sigma_j^2} \geq 1 \tag{47}$$

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

$$\begin{aligned} r_{\pm}(i,m) &:= \frac{\alpha\sigma_i^2 \pm \sqrt{\left(\alpha\sigma_i^2 - 1\right)\left(\alpha d_m + \alpha\sigma_i^2\right)}}{\sigma_i^2 + \left(1 - \alpha\sigma_i^2\right)d_m} \\ &= \frac{\alpha}{\alpha\sigma_i^2 \mp \sqrt{\left(\alpha\sigma_i^2 - 1\right)\left(\alpha d_m + \alpha\sigma_i^2\right)}}. \end{aligned} \tag{48}$$

Recalling that we have to find the $\overline{\alpha}$'s that assure condition (46) 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:

$$r_-(i,m) < \frac{\alpha}{\alpha\sigma_i^2 + \sqrt{\left(\alpha\sigma_i^2 - 1\right)\alpha\sigma_i^2}} =: b_-(i) \tag{49}$$

$$r_+(i,m) > \frac{\alpha\sigma_i^2 + \sqrt{\left(\alpha\sigma_i^2 - 1\right)\alpha\sigma_i^2}}{\sigma_i^2} =: b_+(i) \tag{50}$$

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 $p_{i,m}(\overline{\alpha})$:

$$\overline{\alpha} \in [b_-(i),\ b_+(i)] \quad \Rightarrow \quad p_{i,m}(\overline{\alpha}) \leq 0 \ . \tag{51}$$

Concave parabolas: (i.e. $\sigma_i^2 + \left(1 - \alpha\sigma_i^2\right)d_m < 0$): we check that implication (51) 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:

$$\text{sign}\left(\left.\frac{\partial p_{i,m}(\overline{\alpha})}{\partial\overline{\alpha}}\right|_{b_-(i)}\right) = \text{sign}\left(\left.\frac{\partial p_{i,m}(\overline{\alpha})}{\partial\overline{\alpha}}\right|_{b_+(i)}\right) \tag{52}$$

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 $p_{i,m}(\overline{\alpha}) = -\left(2\alpha\sigma_i^2\right)\overline{\alpha} + \alpha$ 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 (51) is always satisfied, for all m. Now, by simple algebraic manipulations, it can be shown that $\overline{\alpha} \in [b_-(i), b_+(i)]$ is equivalent to condition (13). □

Proof. (of thm. 3) We are seeking the guesses $\overline{\alpha}$ such that:

$$\frac{1}{S}\sum_{i=1}^{S}\text{var}(\mathbf{a} - \widehat{\mathbf{a}}_{\text{dist}}(\overline{\alpha})) \leq \frac{1}{S}\sum_{i=1}^{S}\text{var}(\mathbf{a} - \widehat{\mathbf{a}}_{\text{loc},i}) \tag{53}$$

and, repeating the initial steps of the proof of thm. (1), we obtain the following sufficient condition:

$$\frac{-2}{d_m + \frac{1}{\overline{\alpha}}} + \frac{d_m + \frac{1}{\alpha}}{\left(d_m + \frac{1}{\overline{\alpha}}\right)^2} \leq \frac{1}{S}\sum_{i=1}^{S}\frac{-1}{d_m + \sigma_i^2} \quad \forall m. \tag{54}$$

Now if the following inequality is true:

$$\frac{-1}{d_m + h} \overset{?}{\leq} \frac{1}{S}\sum_{i=1}^{S}\frac{-1}{d_m + \sigma_i^2} \quad \forall m \tag{55}$$

then we can repeat the other steps of proof of thm. 1 to obtain the bound (20). Now condition (55) can be rewritten as:

$$d_m + h \overset{?}{\leq} H\left(d_m + \sigma_1^2, \ldots, d_m + \sigma_S^2\right) \tag{56}$$

but, since $h = H\left(\sigma_1^2, \ldots, \sigma_S^2\right)$, this is true for lemma 10. □

Proof. (of thm. 7) As in thm. 3 we are seeking the guesses $\overline{S}$ such that:

$$\text{var}\left(\mathbf{a} - \widehat{\mathbf{a}}_{\text{dist}}\left(\overline{S}\right)\right) \leq \text{var}\left(\mathbf{a} - \widehat{\mathbf{a}}_{\text{loc},i}\right) \tag{57}$$

and, repeating once more the initial steps of the proof of thm. (1), we obtain the sufficient condition (recall that $\sigma_i^2 = \sigma^2$ for all i's):

$$p_m\left(\overline{S}\right) := \left(\sigma^2 + (1-S)\, d_m\right)\overline{S}^2 + \left(-2\sigma^2 S\right)\overline{S} + \left(\sigma^2 S\right) \leq 0 \tag{58}$$

for all m's. Now for all m's and d_m's we have that $p_m(0) = \sigma^2 S > 0$, that $p_m(1) = (1-S)\left(d_m + \sigma^2\right) < (1-S)\,\sigma^2 < 0$ (we are assuming there are at least two sensors), and if $\dot{p}_m\left(\overline{S}\right) := \partial p_m\left(\overline{S}\right)/\partial\overline{S}$, then we have also that $\dot{p}_m(0) = -2\sigma^2 S < 0$ and that $\dot{p}_m(1) = p_m(1) < 0$. This imply that each $p_m(\cdot)$ has exactly one root in $(0,1)$ (say $r_1(m)$), while the other root, say $r_2(m)$, can be before 0 or after 1 depending on the sign of $\sigma^2 + (1-S)d_m$.

Now consider a fixed m. Condition (46) is assured for $\overline{S} \in [1, \mathcal{S}_m)$, where:

$$\mathcal{S}_m := \begin{cases} +\infty & \text{if } r_2(m) < 0 \\ r_2(m) & \text{otherwise.} \end{cases} \tag{59}$$

If we define $\mathcal{S}_{\min} := \min_m(\mathcal{S}_m)$, condition (43) is now assured for $\overline{S} \in [1, \mathcal{S}_{\min})$. Note that this condition still depends on m (i.e. depends on $C\Sigma_{\mathbf{a}}C^T$). Consider now the parabola with the smallest $\mathcal{S}_m$, say the $\widehat{m}$-th. If $m_{\widehat{m}}$ is its point of minimum, then $2(m_{\widehat{m}} - 1) < \mathcal{S}_{\min}$, so if $\overline{S} \in [1, 2(m_{\widehat{m}} - 1)]$ then condition (43) is again satisfied. Since $(1-S)d_{\widehat{m}} < 0$ we have:

$$m_{\widehat{m}} = \frac{\sigma^2 S}{\sigma^2 + (1-S)d_{\widehat{m}}} > \frac{\sigma^2 S}{\sigma^2} = S \tag{60}$$

and thus $[1, 2(S-1)] \subset [1, 2(\mathcal{S}_{\min} - 1)]$. Now we can conclude that if $\overline{S} \in [1, 2(S-1)]$ then inequality (43) is satisfied, and this proves the proposition. □

Proof. (of prop. 8) Condition (31) assures parabolas of equation (46) to be all concave, thus $\mathcal{S}_{\min} = +\infty$, and this is sufficient for the thesis. □

REFERENCES

[1] D. Puccinelli and M. Haenggi, "Wireless sensor networks: applications and challenges of ubiquitous sensing," *Circuits and Systems Magazine, IEEE*, vol. 5, no. 3, 2005.

[2] A. Papachristodoulou, L. Li, and J. C. Doyle, "Methodological frameworks for large-scale network analysis and design," *SIGCOMM Comput. Commun. Rev.*, vol. 34, no. 3, 2004.

[3] G. Glanzmann, R. Negenborn, G. Andersson, B. D. Schutter, and J. Hellendoorn, "Multi-area control of overlapping areas in power systems for FACTS control," in *Proceedings of Power Tech 2007 (PT 2007)*, July 2007.

[4] D. Bertsekas and J. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods*. Belmont, MA: Athena Scientific, 1997.

[5] P. K. Varshney, *Distributed Detection and Data Fusion*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 1996.

[6] I. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, "A survey on sensor networks," *Communications Magazine, IEEE*, vol. 40, no. 8, pp. 102–114, Aug 2002.

[7] J.-J. Xiao, A. Ribeiro, Z.-Q. Luo, and G. Giannakis, "Distributed compression-estimation using wireless sensor networks," *Signal Processing Magazine, IEEE*, vol. 23, no. 4, pp. 27–41, July 2006.

[8] E. F. Nakamura, A. A. F. Loureiro, , and A. C. Frery, "Information fusion for wireless sensor networks: methods, models, and classifications," *ACM Computing Surveys*, vol. 39, no. 3, August 2007.

[9] S. M. Kay, *Fundamentals of statistical signal processing: estimation theory*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1993.

[10] P. Honeine, C. Richard, J. Bermudez, H. Snoussi, M. Essoloh, and F. Vincent, "Functional estimation in hilbert space for distributed learning in wireless sensor networks," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, April 2009, pp. 2861–2864.

[11] D. Varagnolo, G. Pillonetto, and L. Schenato, "Distributed function and time delay estimation using nonparametric techniques," *Conference on Decision and Control, 2009*, December 2009.

[12] M. Kearns and H. S. Seung, "Learning from a population of hypotheses," *Machine Learning*, vol. 18, no. 2-3, pp. 255–276, February 1995.

[13] K. Yamanishi, "Distributed cooperative bayesian learning strategies," in *COLT '97: Proceedings of the tenth annual conference on Computational learning theory*. New York, NY, USA: ACM, 1997, pp. 250–262.

[14] J. B. Predd, S. R. Kulkarni, and H. V. Poor, "Regression in sensor networks: training distributively with alternating projections," vol. 5910, no. 1, 2005.

[15] ——, "Distributed learning in wireless sensor networks," *Signal Processing Magazine, IEEE*, vol. 23, no. 4, pp. 56–69, July 2006.

[16] ——, "Consistency in models for distributed learning under communication constraints," *Information Theory, IEEE Transactions on*, vol. 52, no. 1, pp. 52–63, January 2006.

[17] J.-F. Chamberland and V. Veeravalli, "Asymptotic results for decentralized detection in power constrained wireless sensor networks," *Selected Areas in Communications, IEEE Journal on*, vol. 22, no. 6, pp. 1007–1015, August 2004.

[18] R. Olfati-Saber, J. Fax, and R. Murray, "Consensus and cooperation in multi-agent networked systems," *Proceedings of IEEE*, vol. 95, no. 1, pp. 215–233, January 2007.

[19] S. Bolognani, S. D. Favero, L. Schenato, and D. Varagnolo, "Consensus-based distributed sensor calibration and least-square parameter estimation in WSNs„" *International Journal of Robust and Nonlinear Control, to appear*, 2009.

[20] J. Cortés, "Distributed algorithms for reaching consensus on general functions," *Automatica*, vol. 44, no. 3, pp. 726–737, March 2008.