Distributed Picard Iteration: Application to Distributed EM and Distributed PCA

Francisco Andrade, Mário A. T. Figueiredo, Fellow, IEEE, and João Xavier

Abstract—In recent work, we proposed a distributed Picard iteration (DPI) that allows a set of agents, linked by a communication network, to find a fixed point of a locally contractive (LC) map that is the average of individual maps held by said agents. In this work, we build upon the DPI and its local linear convergence (LLC) guarantees to make several contributions. We show that Sanger’s algorithm for principal component analysis (PCA) corresponds to the iteration of an LC map that can be written as the average of local maps, each held by at least one node. Consequently, via the DPI, we derive two distributed algorithms whose LLC guarantees follow from those that we proved for the DPI. The verification of the LC condition for EM is challenging, as the underlying parameter estimation from noisy and faulty measurements in a sensor network can be written as the iteration of an LC map that is the average of local maps, each available at just one node. The guarantees of local linear convergence for these distributed algorithms involve verifying condition (1) for the maps inducing them, which allows invoking the results from [1]. Consequently, a great portion of this paper is devoted to proving that (1) holds for these maps, which is far from trivial.

The guarantees of local linear convergence for these distributed algorithms involve verifying condition (1) for the maps inducing them, which allows invoking the results from [1]. Consequently, a great portion of this paper is devoted to proving that (1) holds for these maps, which is far from trivial.

The first part of this work [1] addressed a setup where two or more agents collaborate to collectively find an attractor \( x^* \) of a map \( H \) that can be implicitly represented as an average of local maps, i.e.,

\[
H = \frac{1}{N} \sum_{n=1}^{N} H_n,
\]

where \( H_n \) is the map held by agent \( n \). As defined in [1], an attractor \( x^* \) of \( H \) is a fixed point thereof, \( H(x^*) = x^* \), satisfying

\[
\rho(J_H(x^*)) < 1,
\]

where \( \rho(J_H(x^*)) \) is the spectral radius of the Jacobian of \( H \) at \( x^* \). Moreover, the map \( H \) is not assumed to come from an underlying optimization problem, thus the Jacobian matrix \( J_H(x) \) is not necessarily symmetry, and no global structural assumptions (e.g., Lipschitzianity or coercivity) are made.

The main contributions of [1] are a distributed algorithm to find \( x^* \) – termed the distributed Picard iteration (DPI) – and the proof that the DPI enjoys the local linear convergence of its centralized counterpart, i.e., of the (standard) Picard iteration

\[
x^{k+1} = H(x^k).
\]

This work proposes two instantiations/applications of DPI.

1) A distributed algorithm for principal component analysis (PCA), which results from considering a map that can be implicitly written as an average of local maps and having as a fixed point the solution to the PCA problem.

2) A distributed algorithm that coordinates the \( N \) agents towards collectively estimating a parameter of which each agent has a noisy and possibly faulty measurement. This algorithm stems from the formulation of the problem as a fixed point of a map induced by the stationary equations of the maximum likelihood estimation (MLE) criterion. This map is seen to correspond to the iterations of a slightly modified EM algorithm for a mixture of linear regressions [2].

A. Problem Statement: Distributed PCA

Consider a set of \( N \) agents linked by an undirected and connected communication graph; the nodes are the agents and the edges represent the communication channels between the agents. Each agent \( n \) holds a finite set \( Y_n \subseteq \mathbb{R}^d \) and the agents seek to collectively find the top \( m \) eigenvector of

\[
C = \frac{1}{M} \sum_{n=1}^{N} C_n,
\]

where \( M = \sum_{n=1}^{N} |Y_n| \), i.e., the sum of the cardinalities of each \( Y_n \), and

\[
C_n = \sum_{y \in Y_n} yy^T.
\]

We assume that \( C \) is positive definite, i.e., \( C \succ 0 \).

1This means the \( m \) eigenvectors associated to the largest \( m \) eigenvalues.
**Motivation:** PCA first appeared in the statistics community in the beginning of the 20th century [3] and became one of the workhorses of data analysis, with dimensionality reduction being a notable application.

A technique for dimensionality reduction aims at representing high-dimensional data in a lower dimensional space; this can be crucial to reduce the computational complexity of manipulating data when doing, for example, exploratory analysis. If used as a dimensionality reduction tool, PCA permits to express a high-dimensional data set on the basis formed by its top \( m < d \) eigenvectors. Nowadays, with the ever increasing collection of data by spatially dispersed agents, an algorithm for decentralized PCA constitutes a relevant area of research; see, e.g., [4] for a review of distributed PCA.

### B. Problem Statement: Distributed Parameter Estimation with Noisy and Faulty Measurements

Consider, as before, a set of \( N \) agents linked by an undirected and connected communication graph; each agent holds an independent observation given by

\[
Y_n = Z_n h_n^T \mu^* + W_n, \quad n = 1, \ldots, N, \tag{2}
\]

where: \( \mu^* \in \mathbb{R}^d \) is a fixed and unknown parameter; each \( h_n \in \mathbb{R}^d \) is assumed to be known only at agent \( n \); \( \{W_j\}_{n=1}^N \) are independent and identically distributed (i.i.d.) zero-mean Gaussian random variables with variance \( (\sigma^*)^2 \); and \( \{Z_n\}_{n=1}^N \) are i.i.d. Bernoulli random variables \( (Z_n \in \{0,1\}) \) with

\[
f_{Z_n}(z_n|p^*) = (p^*)^z_n (1-p^*)^{1-z_n}.
\]

The agents seek to collectively estimate \( \mu^* \), treating \( p^* \) and \( \sigma^* \) as nuisance parameters.

**Motivation:** A collection of spatially distributed sensors monitoring the environment for information processing tasks or decision making is a very typical setup nowadays [5]. In common scenarios, these sensors communicate wirelessly and are subject to harsh environments, which may result in faulty communications or sensor malfunctions [6]. The setup described by [2] models a scenario where sensor \( n \) measures the parameter \( \mu^* \) with probability \( p_n \) and, with probability \( 1-p_n \), it senses only noise, indicating a transducer failure [6].

Observe that if the binary variables \( Z_n \) were not random, but fixed and known, the model could be regarded as a (distributed) linear regression problem. However, the randomness introduces an extra layer of difficulty that accounts for the potential sensor malfunctions that typically occur when the sensors are, for example, located outdoors.

A decentralized algorithm, rather than a solution where each sensor first sends its observation to a fusion center that performs a standard estimation algorithm, has the potential for being more robust to faulty wireless communications that may render a sensor useless. Moreover, a decentralized algorithm can result in considerable energy saving [5], certainly a desirable feature.

### C. Contributions and Related Work

The distributed PCA problem described in Section I-A was addressed in [7] (see [5] for a review of distributed PCA algorithms), where an algorithm termed accelerated distributed Sanger’s algorithm (ADSA) was proposed. The authors consider a “mini-batch variant” of Sanger’s algorithm (SA, see [8]) and, inspired by [9], arrive at ADSA. However, no proof of convergence is presented in [7]. Our contributions in this context are: we show that ADSA is recovered by applying DPI to SA and that condition (1) holds for SA, thus, the guarantees of local convergence follow as a consequence of the results in [11].

The noisy and faulty measurements problem presented in section I-B is the setup addressed in [6]; the authors regard [2] as a mixture model [10] and propose a distributed version of the EM algorithm to estimate \( \mu^* \), termed the diffusion-averaging distributed EM (DA-DEM) algorithm. However, DA-DEM, very much in the spirit of [11], uses a diminishing step-size to achieve convergence, which leads to a sublinear convergence rate. Our contribution is a distributed algorithm for this problem that extends to distributed settings a slightly modified version of the centralized EM algorithm. The key challenge is to show that we can “expect” condition (1) to hold and we dedicate a considerable amount of effort to this endeavor. We use the term “expect”, since the operator underlying the Picard iteration depends on the observed samples and, therefore, the existence of an attractor is a probabilistic question. Finally, we compare our algorithm with DA-DEM through Monte Carlo simulations, confirming the linear convergence rate of our algorithm and the sublinear convergence of DA-DEM.

There is a vast literature concerning the “probabilistic line rate of convergence” of the EM algorithm; see, for example, [12], [13], [14]. However, neither the results in [12], nor those in [13] encapsulate the mixture model underlying [2]. The mixture of regressions presented in [14] bears some similarity with the model underlying [2], but it is not the same: in [14], \( p \) is fixed at 1/2 and \( Z_n \in \{-1,1\} \), thus there are no measurements with just noise. Additionally, the authors are primarily concerned with statistical guarantees for the error with respect to the ground truth, and in our case, we address the somewhat simpler goal of establishing [11], while ignoring the connection with the ground truth.

To summarize, we prove local linear convergence of ADSA [7], thereby filling a gap left by [7], and we propose an algorithm to solve the mixture model underlying [2] with guarantees of local linear convergence, as opposed to the sublinear convergence of the DA-DEM algorithm. Both these algorithms are an application of the DPI algorithm proposed in [11] and their corresponding guarantees of linear convergence are a consequence of verifying the attractor conditions for the map underlying the corresponding centralized algorithms. Moreover, we compare DA-DEM and our proposed algorithm through numerical Monte Carlo simulations and the results obtained confirm the linear rate of convergence of our algorithm and the sublinear rate of convergence of DA-DEM.

### D. Organization of the Paper

Section [11] briefly reviews the distributed Picard iteration proposed in [1] and informally states the main result proved therein. The results characterizing the fixed points of the...
“mini-batch” variant of Sanger’s algorithm, as well as the attractor condition, are presented in section III. Section IV describes the centralized variant of EM that underlies the proposed distributed algorithm for the parameter estimation problem from noisy and faulty measurements, presents the verification of the attractor condition, and reports the results of simulations comparing the distributed Picard iteration with DA-DEM. Appendix A addresses a set of regularity conditions necessary for the proof that establishes the attractor condition. Appendix B contains an instrumental theorem for the proof that establishes the attractor condition, as well as the definitions/notation of convergence to zero in probability and uniform boundedness in probability. Finally, in appendix C we prove the result characterizing the fixed points of Sanger’s algorithm.

Notation
Matrices and vectors are denoted by upper and lower case letters, respectively. The spectral radius of a matrix $A$ is denoted by $\rho(A)$ and its Frobenius norm by $\|A\|_F$. Given a map $H$, $J_H(x)$, and $dH(x)$ denote, respectively, the Jacobian of $H$ at $x$ and the differential of $H$ at $x$. Given a vector $v$, $v_s$ denotes its $s$th component; given a matrix $A$, $A_{st}$ denotes the element on the $s$th line and $t$th column and $A^T$ its transpose. The $d$-dimensional identity matrix is denoted by $I_d$, $1_d$ is the $d$-dimensional vector of ones, and $0_{m,n}$ is the $m \times n$ matrix of zeros. Whenever convenient, we will denote a vector with two stacked blocks, $[v^T, u^T]^T$, simply as $(v,u)$. Given a square matrix $A$, $\mathcal{U}(A)$ is an upper triangular matrix of the same dimension as $A$ and whose upper triangular part coincides with that of $A$. Given a norm $\| \cdot \|$, $\bar{B}_{\delta,\theta}^2$ denotes the closed ball of center $\theta^*$ and radius $\delta$ with respect to $\| \cdot \|$. Random variables and vectors are denoted by upper case letters and, given a random variable $Y$, the probability density (or mass) function of $Y$ is denoted by $f_Y$. The probability density of a Gaussian of mean $\mu$ and variance $\sigma^2$ is denoted by $\mathcal{N}(\cdot | \mu, \sigma^2)$.

II. REVIEW OF DISTRIBUTED PICARD ITERATION

In [1], we considered a network of $N$ agents, where the interconnection structure is represented by an undirected connected graph: the nodes correspond to the agents and an edge between two agents indicates they can communicate (are neighbours). Each agent $n \in \{1, ..., N\}$ holds an operator $H_n : \mathbb{R}^d \rightarrow \mathbb{R}^d$, and the goal is to compute a fixed point of the average operator

$$H = \frac{1}{N} \sum_{n=1}^{N} H_n. \quad (3)$$

Crucially, each agent $n$ is restricted to performing computations involving $H_n$ and communicating with its neighbours.

Our only assumption about $H$ is the existence of a locally attractive fixed point $x^*$, i.e., satisfying [1].

Let $R$ be the map on $\mathbb{R}^{dN}$ defined, for $z = [z^T_1, ..., z^T_N]^T$ with $z_j \in \mathbb{R}^d$, by

$$R(z) = \left[(H_1(z_1) - z_1)^T, ..., (H_N(z_N) - z_N)^T\right]^T, \quad (4)$$

and let $W = \bar{W} \otimes I_d$, where $\bar{W}$ is the Metropolis weight matrix associated to the agents’ communication graph [15]. The distributed algorithm proposed in [1] is presented in Algorithm 1, where $\alpha \in \mathbb{R}_{>0}$.

**Algorithm 1** Distributed Picard Iteration (DPI)

1: Initialization:

$$z^0 \in \mathbb{R}^{dN}, \quad z^1 = Wz^0 + \alpha R(z^0), \quad (5)$$

2: Update:

$$z^{k+2} = (I + W)z^{k+1} - \frac{I + W}{2} z^k + \alpha (R(z^{k+1}) - R(z^k)). \quad (6)$$

Informally, Theorem 2 in [1] shows that $\alpha$ can be chosen such that if $z^k$ gets sufficiently close to $1 \otimes x^*$, then it converges to $1 \otimes x^*$ at least linearly (the precise statement and proof can be found in [1]).

III. PCA

To arrive at a distributed algorithm for solving the PCA problem described in section I-A as an application of the DPI algorithm, we introduce a map $H$ having the desired solution as a fixed point. Moreover, the guarantees of local linear convergence follow as a result of verifying [1].

The “mini batch variant” of Sanger’s algorithm (SA) proposed in [7] and inspired by [8] is the Picard iteration

$$X^{k+1} = H(X^k),$$

where $H : \mathbb{R}^{d \times m} \rightarrow \mathbb{R}^{d \times m}$ is defined by

$$H(X) = X + \eta \left( CX - XU(X^T CX) \right). \quad (5)$$

Observe that $H$ can be written as an average of local maps, i.e.,

$$H = \frac{1}{N} \sum_{n=1}^{N} H_n, \quad (6)$$

where $H_n : \mathbb{R}^{d \times m} \rightarrow \mathbb{R}^{d \times m}$ is defined by

$$H_n(X) = X + \eta \left( \frac{N}{M} C_n X - XU(X^T \frac{N}{M} C_n X) \right), \quad (6)$$

with $U(\cdot)$ being the map (see the notation section) that maps a square matrix to an upper triangular matrix with the same dimension and upper triangular part.

Let $R$ be the map on $\mathbb{R}^{dN}$ defined, for $z = [z^T_1, ..., z^T_N]^T$, as in [1] with $H_n$ as in [6]. The distributed algorithm for solving the PCA problem (section I-A) presented in [7] is exactly the distributed Picard iteration, i.e., Algorithm 1 with this choice of $R$.

This section verifies the attractor condition [1] for $H$, thus establishing, as a consequence of the results in [1], the local linear convergence of Algorithm 1 with each $H_n$ in [4] defined as in [6].

**Lemma 1**: Let $C > 0$. If $X^* \in \mathbb{R}^{d \times m}$ satisfies

$$CX^* = X^*U((X^*)^T CX^*) \quad (7)$$


then, each column of $X^*$ is either 0 or a unit-norm eigenvector of $C$. Moreover, the columns are orthogonal, i.e., $(X^*)^T X^*$ is diagonal with the diagonal elements being either one or zero.

We defer the proof of Lemma 1 to appendix C as we regard it as an auxiliary result to Theorem 1.

To see that the ADSA, i.e., Algorithm 1 with $R$ as in (4) and each $H_n$ as in (6), solves the PCA problem in a distributed fashion at a local linear rate, we couple Lemma 1 with Theorem 1. In the following theorem, the $i$th column of $X^*$ is denoted by $x_i^*$.

**Theorem 1:** Let $\lambda_1 > \ldots > \lambda_m > \lambda_m + 1 > \ldots > \lambda_d > 0$ be the eigenvalues of $C$. Suppose that $X^*$ is a $d \times m$ matrix such that $C x_i^* = \lambda_i x_i^*$, for $i = 1, \ldots, m$, and $(X^*)^T X^* = I_m$. Then, there exists an eigenvector $\eta^*$ such that, for $0 < \eta < \eta^*$,

$$\rho(J_H(X^*)) < 1,$$

where the dependence of $H$ on $\eta$ is given in (5).

Note that Lemma 1 shows that the solution sought in the PCA problem (see section I-A) is a fixed point of $H$ and Theorem 1 shows that, at this solution, the Picard iteration of $H$ has local linear convergence. From the results in (11), we obtain the local linear convergence of Algorithm 1 with $R$ as in (4) and each $H_n$ as in (6).

**Proof of Theorem 1**

The proof amounts to checking that the eigenvalues of $J_S(X^*)$, where

$$S(X) = CX - \mathcal{U}(X^T CX)$$

are real and negative. In fact, note that the eigenvalues of

$$J_H(X^*) = I + \eta J_S(X^*)$$

are of the form $1 + \eta \beta$, where $\beta$ is an eigenvalue of $J_S(X^*)$; hence, if all the eigenvalues of $J_S(X^*)$ are real and negative, then, for sufficiently small $\eta$, the result of the theorem holds.

From the rules of matrix differential calculus (see (16) and (17)), the differential of $S$ at $X$, denoted by $dS(X)$, is the linear map

$$dX \rightarrow CDX - (dx) \mathcal{U}(X^T CX) - XD(\mathcal{U}(X^T CX))(X).$$

Observe that $\mathcal{U}(\cdot)$ is a linear map, thus the composition rule for differentials yields

$$d((\mathcal{U}(X^T CX))(X)) = \mathcal{U}((dx)^T CX + X^T CDx).$$

By assumption, $CX^* = X^* D$ with $D = \text{diag}(\lambda_1, \ldots, \lambda_m)$. Let $\tilde{X}$ be an extension of $X^*$ to an orthonormal basis of eigenvectors of $C$, i.e., $(\tilde{X})^T \tilde{X} = I_d$ and $C \tilde{X} = \tilde{X}^* \bar{D}$, where $\bar{D} = \text{diag}(\lambda_1, \ldots, \lambda_m, \lambda_{m+1}, \ldots, \lambda_d)$.

The eigenvalues of $J_S(X^*)$ coincide with those of $dS(X^*)$, under the identification between Jacobians and differentials (see (16)). To understand the eigenvalues of $dS(X^*)$, consider the invertible linear map given by

$$dX \rightarrow (\tilde{X})^T dX \tilde{X}^*.$$

Eigenvalues are invariant under a similarity transformation, hence, the eigenvalues of $J_S(X^*)$ coincide with does of $dS(X^*)$ under the change of variables given by (5), i.e., those of the linear map

$$W \rightarrow \tilde{D} W - WD - ADU(\tilde{D} A^T W + \tilde{W}^T A D),$$

where $A = \begin{bmatrix} I_m & \mathbf{0}_{d-m,m} \end{bmatrix}$.

Let $Z$ be an eigenvector (note that $Z$ is in fact a matrix) of (9) associated to the eigenvalue $\beta$, i.e.,

$$\tilde{D}Z - ZD - U(\tilde{D} A^T Z + Z^T \tilde{A} D) = \beta Z,$$

and we show that $\beta < 0$.

Consider a block partition of $Z$ of the form

$$Z = \begin{bmatrix} \tilde{Z} \\ \bar{Z} \end{bmatrix},$$

where $\tilde{Z}$ is an $m \times m$ matrix and $\bar{Z}$ is a $(d-m) \times m$ matrix. The eigenvalue matrix equation (10) induces the following system of matrix equations

$$D\tilde{Z} - \tilde{Z} D - U(D\tilde{Z} + \tilde{Z}^T D) = \beta \tilde{Z},$$

$$D\bar{Z} - \bar{Z} D = \beta \bar{Z},$$

where $\tilde{D} = \text{diag}(\lambda_{m+1}, \ldots, \lambda_d)$.

There are two non-mutually-exclusive cases to consider: $\tilde{Z} \neq 0$ or $\bar{Z} \neq 0$ ($Z \neq 0$, by virtue of being an eigenvector).

**Case 1:** Suppose that $Z_{st} \neq 0$. Then, (12) implies that $\lambda_{m+s} Z_{st} - \lambda_t Z_{st} = \beta Z_{st}$, and, hence, $\beta = \lambda_{m+s} - \lambda_t < 0$.

**Case 2:** Suppose that $Z_{st} \neq 0$. This case splits in two: either $s > t$ or $s \leq t$. If $s > t$, then (11) and the “upper triangularization” operation yields

$$\lambda_s \tilde{Z}_{st} - \lambda_t \tilde{Z}_{st} = \beta \tilde{Z}_{st},$$

which, after division by $\tilde{Z}_{st}$, yields $\lambda_s - \lambda_t = \beta < 0$. If $s \leq t$, then,

$$\beta \tilde{Z}_{st} = \lambda_s \tilde{Z}_{st} - \lambda_t \tilde{Z}_{st} - U(\tilde{D} \tilde{Z} + \tilde{Z}^T D)_{st}$$

$$= \lambda_s \tilde{Z}_{st} - \lambda_t \tilde{Z}_{st} - \lambda_s \tilde{Z}_{st} - \lambda_t \tilde{Z}_{ts}$$

$$= -\lambda_t (Z_{ts} + Z_{ts}).$$

Now the trick is to notice that if $s < t$, then $Z_{ts}$ can be assumed to be 0, since, otherwise, we could deal with it as in (13) with the roles of $s$ and $t$ reversed to conclude $\beta < 0$. Hence, assuming $Z_{ts} = 0$, we obtain, after division by $\tilde{Z}_{st}$, that $\beta = -\lambda_t < 0$.

Finally, if $s = t$, then $\beta = -2\lambda_t < 0$.

This concludes the proof.
IV. PARAMETER ESTIMATION WITH NOISY MEASUREMENTS

A. Roadmap

This is a rather long section, hence the need for a road map. The analysis of (2) is simplified if the measurements are identically distributed besides being just independent and, therefore, we start by introducing a probability distribution on the vectors $h_n$ and a joint model on $(Y, H)$.

To estimate $\mu^*$ ($p^*$ and $\sigma^*$ are treated as nuisance parameters), we consider the stationary equations imposed by equating to zero the gradient of the log-likelihood function, a necessary condition satisfied by the maximum likelihood estimator (MLE). Once the particular form of the stationary equations is realized, we reformulate them as a fixed point equation of the form $g_1 \circ g_2(\theta^*) = \theta^*$ that naturally suggests the Picard iteration

$$\theta^{k+1} = g_1 \circ g_2(\theta^k).$$

Observing that the map $g_1 \circ g_2$ cannot be written as an average of local maps, we switch to the map $H = g_2 \circ g_1$, which can be implicitly written as an average of local maps. With this map, we arrive at a distributed algorithm by considering the map $R$ (see section II) arising from (1). The full part of the section is thus devoted to verifying (1) for the map $g_1 \circ g_2$, and to a numerical simulation comparing Algorithm 1 and DA-DEM from [5].

B. Identically distributed observations

Let $\theta^* = (\mu^*, p^*, \sigma^*) \in \Theta = \mathbb{R}^d \times (0, 1) \times (0, +\infty)$ be an unknown and fixed vector which we term the ground truth.

The agents’ measurements are assumed to be independent (see [2]); however, they are not identically distributed, given the presence of the vectors $h_n$ in (2). To address this issue, let $Z \in \{0, 1\}, H \in \mathbb{R}^d$, and $Y \in \mathbb{R}$ be, respectively, a binary random variable, a random vector, and a real random variable. Suppose the joint density on $(Y, H, Z)$ factors as

$$f_{Y,H,Z}(y, h, z|\theta^*) = f_H(h)f_Z(z|p^*)f_{Y|H,Z}(y|h, z, \mu^*, (\sigma^*)^2),$$

where

$$f_H(h) = \mathcal{N}(h|0, I_d),$$
$$f_Z(z|p^*) = (p^*)^z(1 - p^*)^{1-z},$$

and

$$f_{Y|H,Z}(y|h, z, \theta^*) = \mathcal{N}(y|h^T \mu^*, (\sigma^*)^2)^2 \mathcal{N}(y|0, (\sigma^*)^2)^{1-z}.$$  

Instead of assuming that the $h_n$ are fixed as in [6], we assume that each sensor $n$ has a measurement $(y_n, h_n)$, where $(y_n, h_n, z_n)$ was drawn from (14), but agent $n$ has no knowledge of $z_n$. After marginalization, the joint density of $Y, H$ is given by

$$f_{Y,H}(y, h|\theta^*) = f_H(h)f_{Y|H}(y|h, \theta^*) = f_H(h)\left(p^* \mathcal{N}(y|h^T \mu^*, (\sigma^*)^2) + (1 - p^*) \mathcal{N}(y|0, (\sigma^*)^2)\right),$$

which is a mixture model [10].

To estimate $\mu^*$, the agents seek $\theta \in \Theta$ such that

$$\frac{1}{N} \sum_{n=1}^{N} \nabla_\theta \phi(y_n, h_n, \theta) = 0,$$

where

$$\phi(y, h, \theta) = \log(f_H(h)) + \log(f_{Y|H}(y|h, \theta)).$$

Since $f_H(h)$ does not depend on $\theta$, we observe that the existence of a fixed point of $g_1 \circ g_2$ satisfying (1) follows from the existence of a fixed point of $g_2 \circ g_1$ satisfying (1). The full part of the section is thus devoted to verifying (1) for the map $g_1 \circ g_2$, and to a numerical simulation comparing Algorithm 1 and DA-DEM from [5].

C. Gradient of $\phi$ and the centralized algorithm

Before explicitly writing the stationary equations corresponding to (17), we introduce the responsibility functions (10).

$$r(y, h, \theta) = \frac{p \mathcal{N}(y|h^T \mu, \sigma^2)}{p \mathcal{N}(y|h^T \mu, \sigma^2) + (1 - p) \mathcal{N}(y|0, \sigma^2)}.$$  

Notice that $r(y, h, \theta) = \mathbb{P}(z = 1|y, h, \theta)$, the posterior probability that the observation $y$ was not a result of measuring only noise.

A simple calculation reveals that the gradient of $\phi$ satisfies the identities

$$\sigma^2 \nabla_\mu \phi(y, h, \theta) = r(y, h, \theta)(y - h^T \mu)h,$$
$$p + p(1 - p) \frac{\partial \phi}{\partial p}(y, h, \theta) = r(y, h, \theta),$$
$$\sigma^2 + 2(\sigma^2)^2 \frac{\partial \phi}{\partial \sigma^2}(y, h, \theta) = r(y, h, \theta)(y - h^T \mu)^2 + (1 - r(y, h, \theta))y^2,$$

thus, (17) can be explicitly written as

$$\left(\frac{1}{N} \sum_{n=1}^{N} \Gamma(y_n, h_n, \theta)\right) \mu = \frac{1}{N} \sum_{n=1}^{N} \psi(y_n, h_n, \theta),$$
$$p = \frac{1}{N} \sum_{n=1}^{N} r(y_n, h_n, \theta),$$
$$\sigma^2 = \frac{1}{N} \sum_{n=1}^{N} \gamma(y_n, h_n, \theta),$$

where

$$\Gamma(y, h, \theta) = r(y, h, \theta)hh^T,$$
$$\psi(y, h, \theta) = r(y, h, \theta)yh,$$
$$\gamma(y, h, \theta) = r(y, h, \theta)(y - h^T \mu)^2 + (1 - r(y, h, \theta))y^2.$$
If the matrix $\frac{1}{N} \sum_{n=1}^{N} \Gamma(y_n, h_n, \theta)$ is invertible, then (21)-(23) can be written as a fixed point equation. This constitutes the motivation for the centralized algorithm that we suggest next (see Algorithm 2) and from which we will derive the distributed version; observe that it is the Picard iteration motivated by (21)-(23).

Another way to write (27)-(29) is $\theta^{k+1} = g_1 \circ g_2(\theta^k)$, where

$$g_2(\theta) = \frac{1}{N} \left( \sum_{n=1}^{N} \Gamma(y_n, h_n, \theta), \sum_{n=1}^{N} \psi(y_n, h_n, \theta), \sum_{n=1}^{N} r(y_n, h_n, \theta), \sum_{n=1}^{N} \gamma(y_n, h_n, \theta) \right)$$

and

$$g_1(\Gamma, \psi, p, \sigma^2) = (\Gamma^{-1}, \psi, p, \sigma^2).$$

Algorithm 2 Centralized variant of EM

$$\mu^{k+1} = \frac{1}{N} \sum_{n=1}^{N} \Gamma(y_n, h_n, \theta^k)^{-1} \frac{1}{N} \sum_{n=1}^{N} \psi(y_n, h_n, \theta^k)$$

$$\lambda^{k+1} = \frac{1}{N} \sum_{n=1}^{N} r(y_n, h_n, \theta^k)$$

$$(\sigma^{k+1})^2 = \frac{1}{N} \sum_{n=1}^{N} \gamma(y_n, h_n, \theta^k).$$

D. Distributed Algorithm

The map $g_1 \circ g_2$ is not an average of local maps, given the matrix inversion in (27). As a consequence, we cannot directly use Algorithm 1 to extend (27)-29 to a distributed algorithm. However, switching the order of $g_1$ and $g_2$ results in a map that can be implicitly written as an average of local maps. In fact, instead of the iteration $\theta^{k+1} = g_1 \circ g_2(\theta^k)$, consider the iteration

$$z^{k+1} = H(z^k),$$

where $H = g_2 \circ g_1$ and $z = (\Gamma, \psi, p, \sigma^2)$. Let

$$G_n(\theta) = \left( \Gamma(y_n, h_n, \theta), \psi(y_n, h_n, \theta), r(y_n, h_n, \theta), \gamma(y_n, h_n, \theta) \right),$$

and it follows that $H = \frac{1}{N} \sum_{n=1}^{N} H_n$, where

$$H_n(z) = G_n \circ g_1(z).$$

To conclude, the distributed algorithm we suggest amounts to Algorithm 1, with $R : \mathbb{R}^{(d^2 + d + 2)N} \rightarrow \mathbb{R}^{(d^2 + d + 2)N}$ defined, for $z = (z^1, \ldots, z^N)^T$, as in (30), and $H_n$ as in (30). Additionally, following [6], we suggest the initialization

$$z^0 = \frac{1}{N} \sum_{m=1}^{N} W_{nm} G_m \left( y_n h_n, \frac{1}{2} \nu_n^2 \right).$$

Some remarks are due:

a) The existence of a fixed point of $g_1 \circ g_2$ satisfying (1) is addressed in section IV-E.

b) The existence of a fixed point of $g_2 \circ g_1$ satisfying (1) follows from the existence of a fixed point of $g_1 \circ g_2$ satisfying (1), by the chain rule.

c) Expanding (29) yields

$$(\sigma^{k+1})^2 = \frac{1}{N} \sum_{n=1}^{N} r(y_n, h_n, \theta^k)(y_n - h_n^T \mu_k)^2 + (1 - r(y_n, h_n, \theta_k)) y_n^2,$$

and, if the update rule is modified according to

$$(\sigma^{k+1})^2 = \frac{1}{N} \sum_{n=1}^{N} r(y_n, h_n, \theta^k)(y_n - h_n^T \mu_{k+1})^2 + (1 - r(y_n, h_n, \theta_k)) y_n^2,$$

then, a straightforward manipulation recovers the EM algorithm presented in [6]. Moreover, the EM algorithm derived in [6] is still amenable to a distributed implementation using Algorithm 1. However, we found it easier to prove (1) for this variant of EM, than for the standard EM.

E. Convergence Analysis

The proof of local linear convergence of the centralized variant of EM, i.e., Algorithm 2, is not trivial. In fact, this question is probabilistic in nature, given the fact that updates (27)-29 are dependent on observations that are, in turn, samples from a probability distribution. Before presenting the main convergence result (Theorem 2), we need to introduce some definitions and only one mild assumption: the Fisher information at $\theta^*$, given by

$$I(\theta^*) = E_{\theta^*} \left[ \nabla_{\theta} \phi(y, h, \theta^*)(\nabla_{\theta} \phi(y, h, \theta^*))^T \right].$$

is non-singular.

Let $T_N$ denote the map underlying the Picard iteration (27)-29, i.e., the map $g_1 \circ g_2$. A straightforward manipulation, using (20), shows that

$$T_N(\theta) = \theta + A_N(\theta)^{-1} \frac{1}{N} \sum_{n=1}^{N} \nabla_{\theta} \phi(y_n, h_n, \theta),$$

where

$$A_N(\theta) = \left[ \frac{1}{N} \sum_{n=1}^{N} \frac{1}{p} \Gamma(y_n, h_n, \theta) \begin{bmatrix} 0 & 0 \\ 0 & \frac{1}{p(1-p)} \end{bmatrix} 0 \right] \frac{1}{2(\sigma^2)^2}.$$

Before stating the main result, we introduce the “infinite sample” map, i.e.,

$$T(\theta) = \theta + (A(\theta))^{-1} \mathcal{L}(\theta),$$

The subscript $N$ is to recall that $T_N$ depends on $N$ observations.
where
\[
A(\theta) = \begin{bmatrix}
\mathbb{E}_{\theta^{*}} \left[ \frac{1}{\sigma^{2}} \Gamma(y, h, \theta) \right] & 0 & 0 \\
0 & \frac{1}{p(1-p)} & 0 \\
0 & 0 & \frac{1}{2(\sigma^{2})^2}
\end{bmatrix},
\]
and
\[
\mathcal{L}(\theta) = \mathbb{E}_{\theta^{*}} \left[ \nabla \phi(y, h, \theta) \right].
\]

Observe that \( T \) is a “natural” map to consider:

a) For fixed \( \theta \), \( T_{N}(\theta) \) converges in probability to \( T(\theta) \), from the weak law of large numbers. Said differently, \( T_{N} \) converges pointwise in probability to \( T \);

b) The map \( \mathcal{L} \) satisfies \( \mathcal{L}(\theta^{*}) = 0 \), hence,
\[
T(\theta^{*}) = \theta^{*},
\]
\( i.e., \theta^{*} \) is a fixed point of \( T \);

c) The map \( T \) is differentiable (this requires differentiability under the integral sign - see appendix \([A]\)).

Moreover,
\[
J_{T}(\theta^{*}) = I + (A(\theta^{*}))^{-1} \mathbb{E}_{\theta^{*}} \left[ \nabla^{2} \phi(y, h, \theta^{*}) \right].
\]

Under suitable regularity conditions (see appendix \([A]\) that hold for the model \([10]\),
\[
\mathbb{E}_{\theta^{*}} \left[ \nabla^{2} \phi(y, h, \theta^{*}) \right] = -I(\theta^{*}).
\]

Additionally, a simple calculation reveals that
\[
A(\theta^{*}) = I_{c}(\theta^{*}),
\]
where \( I_{c}(\theta^{*}) \) is the Fisher information of the complete model \([13]\). The non-singularity assumption on \( I(\theta^{*}) \), together with the principle of missing information (see \([13]\), page 101), implies that
\[
0 < I(\theta^{*}) \leq I_{c}(\theta^{*}).
\]

All these observations show that
\[
J_{T}(\theta^{*}) = I - (I_{c}(\theta^{*}))^{-1} I(\theta^{*}),
\]
and, Theorem 7.7.3 of \([19]\), together with \( 0 < I(\theta^{*}) \leq I_{c}(\theta^{*}) \), implies that
\[
\rho(J_{T}(\theta^{*})) < 1,
\]
concluding that \( \theta^{*} \) is an attractor of \( T \).

We show that the probabilistic that \( T_{N} \) has an attractor approaches 1 as \( N \) tends to infinity; the strategy is to derive this from the existence of an attractor of \( T \). Pointwise convergence in probability is not enough to arrive at this result. In fact, the proof is built on a stronger notion that is a probabilistic version of uniform convergence of maps.

Observe that if \( \theta_{N} \) is a fixed point of \( T_{N} \), then
\[
J_{T_{N}}(\theta_{N}) = I + (A_{N}(\theta_{N}))^{-1} \frac{1}{N} \sum_{n=1}^{N} \nabla^{2} \phi(y_{n}, h_{n}, \theta_{N}),
\]
so let
\[
T_{N}^{*}(\theta) = I + (A_{N}(\theta))^{-1} \frac{1}{N} \sum_{n=1}^{N} \nabla^{2} \phi(y_{n}, h_{n}, \theta),
\]
\[
T^{*}(\theta) = I + (A(\theta))^{-1} \mathbb{E}_{\theta^{*}} \left[ \nabla^{2} \phi(y, h, \theta) \right].
\]

We now state the main convergence result; recall (see the notation section) that \( \bar{B}_{\delta,\theta^{*}}^{\parallel} \) is the closed ball of center \( \theta^{*} \) and radius \( \delta \), with respect to the metric induced by the norm \( \parallel \cdot \parallel \).

**Theorem 2:** There exists \( \delta > 0 \) and a norm \( \parallel \cdot \parallel \) such that
\[
\mathbb{P}_{\theta^{*}} \left( \sup_{\theta \in \bar{B}_{\delta,\theta^{*}}^{\parallel}} \parallel T_{N}(\theta) - \theta^{*} \parallel \leq \delta \right) \rightarrow 1
\]
and
\[
\mathbb{P}_{\theta^{*}} \left( \sup_{\theta \in \bar{B}_{\delta,\theta^{*}}^{\parallel}} \parallel T_{N}'(\theta) \parallel < 1 \right) \rightarrow 1,
\]
where \( \parallel T_{N}'(\theta) \parallel \) is the induced matrix norm.

Before presenting the proof, we explain why Theorem 2 encapsulates the notion that, with probability approaching 1, the map \( T_{N} \) has an attractor. Let
\[
\mathcal{A}_{N} = \{ (y, h) \in \mathbb{R}^{N} \times \mathbb{R}^{dN} : \sup_{\theta \in \bar{B}_{\delta,\theta^{*}}^{\parallel}} \parallel T_{N}(\theta) - \theta^{*} \parallel \leq \delta \}
\]
\[
\mathcal{B}_{N} = \{ (y, h) \in \mathbb{R}^{N} \times \mathbb{R}^{dN} : \sup_{\theta \in \bar{B}_{\delta,\theta^{*}}^{\parallel}} \parallel T_{N}'(\theta) \parallel < 1 \}.
\]
The statement of Theorem 2 is that the (non-random sequences) \( \mathbb{P}_{\theta^{*}}(\mathcal{A}_{N}) \) and \( \mathbb{P}_{\theta^{*}}(\mathcal{B}_{N}) \) both tend to 1. The inequalities
\[
\mathbb{P}_{\theta^{*}}(\mathcal{A}_{N}) + \mathbb{P}_{\theta^{*}}(\mathcal{B}_{N}) - 1 \leq \mathbb{P}_{\theta^{*}}(\mathcal{A}_{N} \cap \mathcal{B}_{N}) \leq \mathbb{P}_{\theta^{*}}(\mathcal{A}_{N})
\]
imply that
\[
\mathbb{P}_{\theta^{*}}(\mathcal{A}_{N} \cap \mathcal{B}_{N}) \rightarrow 1.
\]

Now note that, if both inequalities hold, namely
\[
\sup_{\theta \in \bar{B}_{\delta,\theta^{*}}^{\parallel}} \parallel T_{N}(\theta) - \theta^{*} \parallel \leq \delta \quad (33)
\]
\[
\sup_{\theta \in \bar{B}_{\delta,\theta^{*}}^{\parallel}} \parallel T_{N}'(\theta) \parallel < 1, \quad (34)
\]
then \( 33 \), together with Brouwer’s fixed point theorem (see \([21]\), page 180) implies that \( T_{N} \) has a fixed point \( \theta_{N} \) in \( \bar{B}_{\delta,\theta^{*}}^{\parallel} \) (this idea is loosely inspired by \([22]\), page 69). Moreover, being a fixed point, at a \( \theta_{N} \) it holds that \( T_{N}(\theta_{N}) = J_{T_{N}}(\theta_{N}) \), so, \( 34 \) implies that
\[
\rho(J_{T_{N}}(\theta_{N})) \leq \parallel J_{T_{N}}(\theta_{N}) \parallel \leq \sup_{\theta \in \bar{B}_{\delta,\theta^{*}}^{\parallel}} \parallel T_{N}'(\theta) \parallel < 1.
\]

This explains why Theorem 2 expresses the notion that we can “expect” \( 1 \) to hold.

**Proof of Theorem 2.** Let \( \parallel \cdot \parallel \) be any norm. Then
\[
\parallel T_{N}(\theta) - \theta^{*} \parallel \leq \parallel T_{N}(\theta) - T(\theta) \parallel + \parallel T(\theta) - \theta^{*} \parallel, \quad (35)
\]
Under regularity conditions (see Appendix \([A]\), \( T \) is differentiable, \( \mathbb{E}_{\theta^{*}} \left[ \nabla^{2} \phi(y, h, \theta^{*}) \right] = -I(\theta^{*}) \), and
\[
\rho(J_{T}(\theta^{*})) = \rho(I - (I_{c}(\theta^{*}))^{-1}I(\theta^{*})) < 1,
\]
as shown above in \( c \).

\(^{4}\)The measurability of the maps in this Theorem are a consequence of Proposition 7.32 in \([20]\).
From the proof of Ostrowski’s Theorem (see [23], page 300), there exists a norm \( \| \cdot \| \) on \( \mathbb{R}^{d+2} \), an open neighborhood \( \mathcal{V} \) of \( \theta^* \), and \( \lambda < 1 \), such that

1. \( \| T(\theta) - \theta^* \| \leq \lambda \| \theta - \theta^* \| \), for \( \theta \in \mathcal{V} \);
2. \( \| J_T(\theta^*) \| < 1 \), where here the norm is the induced matrix norm.

Choose \( \delta \) sufficiently small such that

i) \( \bar{B}_{\delta,\theta} \subseteq \mathcal{V} \);
ii) \( \| T'(\theta) \| = \| J_T(\theta) \| \leq \beta < 1 \), for \( \theta \in \bar{B}_{\delta,\theta} \),

where the validity of ii) follows from the compactness of \( \bar{B}_{\delta,\theta} \), and the continuity of \( T' \).

Now, for any \( \theta \in \bar{B}_{\delta,\theta} \), (35) implies that

\[
\| T_N(\theta) - \theta^* \| \leq \| T_N(\theta) - T(\theta) \| + \lambda \delta
\]

and, hence,

\[
\sup_{\theta \in \bar{B}_{\delta,\theta}} \| T_N(\theta) - \theta^* \| \leq \sup_{\theta \in \bar{B}_{\delta,\theta}} \| T_N(\theta) - T(\theta) \| + \lambda \delta. \tag{36}
\]

A similar reasoning shows that

\[
\sup_{\theta \in \bar{B}_{\delta,\theta}} \| T'_N(\theta) \| \leq \sup_{\theta \in \bar{B}_{\delta,\theta}} \| T'_N(\theta) - T'(\theta) \| + \beta. \tag{37}
\]

We now invoke an instrumental result (proof below).

**Theorem 3:** With \( T_N \), \( T'_N \), and \( T \) as defined above,

\[
\sup_{\theta \in \bar{B}_{\delta,\theta}} \| T_N(\theta) - T(\theta) \| \rightarrow 0 \tag{38}
\]

\[
\sup_{\theta \in \bar{B}_{\delta,\theta}} \| T'_N(\theta) - T'(\theta) \| \rightarrow 0 \tag{39}
\]

both in probability, as \( N \rightarrow \infty \).

To conclude, we show why Theorem 3 implies the result. Let \( \epsilon_1 = (1 - \lambda) \delta \) and \( \epsilon_2 = \frac{1 - \beta}{\delta} \). From the properties of \( \lambda \) and \( \beta \), it holds that \( \epsilon_1 > 0 \) and \( 0 < \epsilon_2 < 1 \). By the definition of convergence in probability, it holds that

\[
P_{\theta^*} \left( \sup_{\theta \in \bar{B}_{\delta,\theta}} \| T_N(\theta) - T(\theta) \| \leq \epsilon_1 \right) \rightarrow 1
\]

\[
P_{\theta^*} \left( \sup_{\theta \in \bar{B}_{\delta,\theta}} \| T'_N(\theta) - T'(\theta) \| \leq \epsilon_2 \right) \rightarrow 1.
\]

From (36), (37), and the forms of \( \epsilon_1 \) and \( \epsilon_2 \), we conclude the result.

**Proof of Theorem 3**. We give only a sketch of the proof of (39) (the proof of (38) is analogous). Observe that

\[
T'_N(\theta) - T'(\theta) =
\]

\[
(A(\theta))^{-1} \left( A(\theta) - A_N(\theta) \right) \left( A_N(\theta) \right)^{-1} \frac{1}{N} \sum_{n=1}^{N} \nabla_{\theta}^2 \phi(y_n, h_n, \theta) + (A(\theta))^{-1} \left( \frac{1}{N} \sum_{n=1}^{N} \nabla_{\theta}^2 \phi(y_n, h_n, \theta) - \mathbb{E}_{\theta^*} \left[ \nabla_{\theta}^2 \phi(y, h, \theta) \right] \right).
\]

which implies that

\[
\| T'_N(\theta) - T'(\theta) \| \leq \frac{1}{N} \sum_{n=1}^{N} \nabla_{\theta}^2 \phi(y_n, h_n, \theta) \]

\[
\left( \left\| A(\theta) - A_N(\theta) \right\| \left\| A_N(\theta) \right\|^{-1} \frac{1}{N} \sum_{n=1}^{N} \nabla_{\theta}^2 \phi(y_n, h_n, \theta) \right) \]

\[
+ \left( \frac{1}{N} \sum_{n=1}^{N} \nabla_{\theta}^2 \phi(y_n, h_n, \theta) - \mathbb{E}_{\theta^*} \left[ \nabla_{\theta}^2 \phi(y, h, \theta) \right] \right).
\]

(40)

From Theorem 6 (see appendix 1),

\[
\sup_{\theta \in \bar{B}_{\delta,\theta}} \| A_N(\theta) - A(\theta) \| \rightarrow 0 \tag{41}
\]

\[
\sup_{\theta \in \bar{B}_{\delta,\theta}} \left( \frac{1}{N} \sum_{n=1}^{N} \nabla_{\theta}^2 \phi(y_n, h_n, \theta) - \mathbb{E}_{\theta^*} \left[ \nabla_{\theta}^2 \phi(y, h, \theta) \right] \right) \rightarrow 0 \tag{42}
\]

in probability; these are consequences of Theorem 6, by noting that:

a) \( \| \frac{1}{\sigma^2} \Gamma(y, h, \theta) \| \leq M \| hh^T \| \), where \( M \) is the maximum of \( \frac{1}{\sigma^2} \) on \( \bar{B}_{\delta,\theta} \), and where we used the fact that \( |r(y, h, \theta)| \leq 1 \);

b) \( \| \nabla_{\theta}^2 \phi(y, h, \theta) \| \leq g(y, h) \) on \( \bar{B}_{\delta,\theta} \), for some map \( g \) not depending on \( \theta \) for which \( \mathbb{E}_{\theta^*}[g(y, h)] \) exists and is finite (see appendix A).

Since all norms are equivalent, (41)-(42) also holds if the Frobenius norm is replaced by any other norm.

Taking the supremum over on \( \bar{B}_{\delta,\theta} \), on both sides of (40), we obtain, from (41)-(42), that

\[
\sup_{\theta \in \bar{B}_{\delta,\theta}} \| T'_N(\theta) - T'(\theta) \|
\]

\[
\leq o_P(1) \sup_{\theta \in \bar{B}_{\delta,\theta}} \| A_N(\theta) \|^{-1} \sup_{\theta \in \bar{B}_{\delta,\theta}} \left( \frac{1}{N} \sum_{n=1}^{N} \nabla_{\theta}^2 \phi(y_n, h_n, \theta) \right)
\]

\[
+ o_P(1),
\]

where the definitions of \( O_P(1) \) and \( o_P(1) \) can be found in appendix 1.

From (41) and (42), together with the compactness of \( \bar{B}_{\delta,\theta} \), we can deduce (proof omitted) that

\[
\sup_{\theta \in \bar{B}_{\delta,\theta}} \| A_N(\theta) \|^{-1} = O_P(1)
\]

\[
\sup_{\theta \in \bar{B}_{\delta,\theta}} \left( \frac{1}{N} \sum_{n=1}^{N} \nabla_{\theta}^2 \phi(y_n, h_n, \theta) \right) = O_P(1).
\]

Putting everything together, we conclude that

\[
\sup_{\theta \in \bar{B}_{\delta,\theta}} \| T'_N(\theta) - T'(\theta) \| \leq \]

\[
o_P(1)O_P(1) + o_P(1) = o_P(1),
\]

where the equality follows from the calculus rules with \( O_P(1) \) and \( o_P(1) \) (22).

As mentioned, the proof of (35) is entirely analogous; just note that, in order to use Theorem 6, we need to check that \( \| \nabla_{\theta} \phi(y, h, \theta) \| \leq \tilde{g}(y, h) \) on \( \bar{B}_{\delta,\theta} \), for some map \( \tilde{g} \) not depending on \( \theta \), for which \( \mathbb{E}[\tilde{g}(y, z)] \) exists and is finite; this is again a consequence of the results proved in appendix A.
F. Simulations

In this section, we compare our algorithm with the one from [6] (DA-DEM) through Monte Carlo simulations. The parameters generated once and fixed throughout all Monte Carlo runs were: $d=3$, $N=100$, a unit-norm vector $\mu^* \in \mathbb{R}^d$, $p^* = 0.7$, and an undirected connected graph on $N$ nodes with connectivity radius $r_c = 0.18$.

Each Monte Carlo run consisted in

1) Generating a data set: each $h_n$ was independently sampled from a Gaussian with zero mean and covariance $I_3$; the variance of the noise $(\sigma^*)^2$ was generated according to

$$(\sigma^*)^2 = \frac{\|H\|^2_F}{N \times \text{SNR}},$$

with $H^T = [h_1 \ldots h_N]$ and where SNR is the signal to noise ratio (we experimented with SNR $\in \{10\text{db}, 20\text{db}\}$). Finally, each $y_n$ was sampled according to $f_{y|H}$ (see (16)), with $h_n, \mu^*, p^*$, and $(\sigma^*)^2$.

2) Computing 10000 iterations of the algorithm proposed in [6], with $\rho \in \{2, 3, 4\}$, and of Algorithm 1, with $\alpha \in \{0.001, 0.005, 0.01\}$. Both algorithms were initialized according to (31).

3) The performance metrics consisted in finding a fixed point using the centralized operators as follows. We first computed

$$\theta^0(\alpha) = \frac{1}{N} \sum_{n=1}^{N} g_1(z_n^{10000}(\alpha))$$

$$\theta^0(\rho) = \frac{1}{N} \sum_{n=1}^{N} \hat{g}_1(z_n^{10000}(\rho)),
$$

where: $\alpha \in \{0.001, 0.005, 0.01\}$; $\rho \in \{2, 3, 4\}$; $\hat{g}_1$ corresponds to the map arising from the standard EM algorithm derived in [6]. In fact, as seen in [6], the EM algorithm can be written as

$$\theta^{k+1} = \hat{g}_1 \circ \hat{g}_2(\theta^k),$$

where

$$\hat{g}_2(\theta) = \frac{1}{N} \left( \sum_{n=1}^{N} \Gamma(y_n, h_n, \theta) \sum_{n=1}^{N} \psi(y_n, h_n, \theta), \sum_{n=1}^{N} r(y_n, h_n, \theta), \sum_{n=1}^{N} y_n^2 \right)$$

and

$$\hat{g}_1(\Gamma, \psi, p, a) = (\Gamma^{-1} \psi, \hat{p}, a - \psi^T \Gamma^{-1} \psi).$$

We ran the algorithms, with initialization as in (43) and (44), given by

$$\theta^{k+1}(\alpha) = g_1 \circ \hat{g}_2(\theta^k(\alpha))$$

$$\theta^{k+1}(\rho) = \hat{g}_1 \circ \hat{g}_2(\theta^k(\rho)),$$

until we found $\theta^*(\alpha)$ and $\theta^*(\rho)$ satisfying

$$\left\| \theta^*(\alpha) - g_1 \circ \hat{g}_2(\theta^*(\alpha)) \right\| \leq 10^{-10}$$

$$\left\| \theta^*(\rho) - \hat{g}_1 \circ \hat{g}_2(\theta^*(\rho)) \right\| \leq 10^{-10}.$$ The error at iteration $k$ of the distributed algorithms was then computed as

$$\frac{1}{N} \sum_{n=1}^{N} \left\| \pi_1 \circ g_1((z_n^k(\alpha)) - \theta^*(\alpha)) \right\|$$

$$\frac{1}{N} \sum_{n=1}^{N} \left\| \pi_1 \circ \hat{g}_1((z_n^k(\rho)) - \theta^*(\rho)) \right\|,$$

where $\pi_1$ is the projection onto the average, i.e., $\pi_1(\mu, p, \sigma^2) = \mu$ (as mentioned before, $p$ and $\sigma^2$ were treated as nuisance parameters).

The number of Monte Carlo tests was 100 and the errors at iteration $k$ were averaged out of 100 for each $\alpha$ and $\rho$. The results for two different SNR values are shown in logarithmic scale in Figures 1 and 2.

![Fig. 1.](image1.png)

**Fig. 1.** The figure shows the result of the Monte Carlo simulation of the error with respect to each optimum for an SNR = 10db and a connectivity radius of 0.18. The dashed curves correspond to the algorithm from [6] with parameter $\rho \in \{2, 3, 4\}$ and the non-dashed curves correspond to the DPI algorithm with parameter $\alpha \in \{0.001, 0.005, 0.01\}$. The number of Monte Carlo tests was 100 and the errors at iteration $k$ were averaged out of 100 for each $\alpha$ and $\rho$. The results for two different SNR values are shown in logarithmic scale in Figures 1 and 2.

![Fig. 2.](image2.png)

**Fig. 2.** The figure shows the result of the Monte Carlo simulation of the error with respect to each optimum for an SNR = 20db and a connectivity radius of 0.18. The dashed curves correspond to the algorithm from [6] with parameter $\rho \in \{2, 3, 4\}$ and the non-dashed curves correspond to the DPI algorithm with parameter $\alpha \in \{0.001, 0.005, 0.01\}$. 

\[5\] $N$ points were randomly deployed on the unit square; two points were then connected by an edge if their distance was less than $r_c$. 


The simulations show, as expected from the theory, that Algorithm 1 converges linearly and clearly outperforms the algorithm from [6], which, given its diminishing step-size, is bound to converge only sub-linearly. Moreover, both algorithms require just one round of communications per iteration.

V. CONCLUSION

This article built upon the distributed Picard algorithm and its convergence properties provided in [11] to make two main contributions: we provided a proof of local linear convergence for the distributed PCA algorithm suggested in [7], thereby filling a gap left by that work; starting from the distributed Picard iteration, we proposed a distributed algorithm for solving the parameter estimation problem from noisy and faulty measurements that had been addressed in [6]. Unlike the algorithm in [6], which uses diminishing step sizes, thus exhibiting sublinear convergence rate, the proposed instance of the distributed Picard iteration is guaranteed to have local linear convergence. Numerical experiments confirm the theoretical advantage of the proposed method with respect to that from [6].

APPENDIX A

REGULARITY CONDITIONS

Theorem 4: Let $K \subset \Theta$ be a compact set containing $\theta^*$. Then, for all $\theta \in K$,
\begin{align}
    \frac{\partial \phi}{\partial x_i}(y, h, \theta) & \leq \mathcal{P}^\phi_{\theta^* x_1 \ldots \theta^* x_k}(y, |h_1|, \ldots, |h_d|) \tag{45}
\end{align}
\begin{align}
    \frac{\partial \theta}{\partial y_i}(y, h, \theta) & \leq \mathcal{P}^\theta_{\theta^* x_1 \ldots \theta^* x_k}(y, |h_1|, \ldots, |h_d|), \tag{46}
\end{align}
where $\sum_{j=1}^K i_j = i \geq 1$, $x_1, \ldots, x_k$ are dummy variables in $\{\mu_1, \ldots, \mu_d, p, \sigma^2\}$ (i.e. consider partial differentiation with respect to the components of $\theta$), and where $\mathcal{P}^r_{\theta^* x_1 \ldots \theta^* x_k}$ and $\mathcal{P}^\phi_{\theta^* x_1 \ldots \theta^* x_k}$ are polynomials.

We start with a proof of (46). Note that $r$ satisfies the following differential equations
\begin{align}
    \frac{\partial r}{\partial \mu_i}(y, h, \theta) &= (1 - r(\theta)) r(\theta) \left( \frac{y - h^T \mu_i}{\sigma^2} \right), \quad i = 1, \ldots, d, \\
    \frac{\partial r}{\partial p}(y, h, \theta) &= \frac{1}{p(1 - p)} r(\theta)(1 - r(\theta)), \\
    \frac{\partial r}{\partial \sigma^2}(y, h, \theta) &= \left( \frac{(y - h^T \mu)^2}{2(\sigma^2)^2} - \frac{y^2}{2(\sigma^2)^2} \right) r(\theta)(1 - r(\theta)).
\end{align}
We deduce that
\begin{align}
    \frac{\partial r}{\partial \lambda}(y, h, \theta) = \frac{\mathcal{P}^r_{\theta^* x_1 \ldots \theta^* x_k}(y, h, \theta)}{\mathcal{Q}^r_{\theta^* x_1 \ldots \theta^* x_k}(p, \sigma^2)}, \tag{47}
\end{align}
where $\mathcal{P}^r_{\theta^* x_1 \ldots \theta^* x_k}$ and $\mathcal{Q}^r_{\theta^* x_1 \ldots \theta^* x_k}$ are polynomials and $\lambda$ is a dummy variable in $\{\mu_1, \ldots, \mu_d, p, \sigma^2\}$.

The chain rule of differentiation, the quotient rule of differentiation and the form (47) imply that
\begin{align}
    \frac{\partial r}{\partial x_i}(y, h, \theta) &= \frac{\mathcal{P}^r_{\theta^* x_1 \ldots \theta^* x_k}(y, h, \theta)}{\mathcal{Q}^r_{\theta^* x_1 \ldots \theta^* x_k}(p, \sigma^2)} \tag{48}
\end{align}
The result now follows easily from $0 < r(\theta) \leq 1$, the compactness of $K$ and the identity (48).

The inequality (45) follows from (48) and the form of the gradient of $\phi$ in (26). This concludes the proof of Theorem 4.

An immediate corollary of Theorem 4 is that the modulus of the partial derivatives of both $\phi$ and $r$ are dominated by functions whose expectation exists and is finite; this is the content of the following result.

Theorem 5: Let $\mathcal{P}$ be a polynomial in $d + 1$ variables. Then $E_\theta \mathcal{P}(y, |h_1|, \ldots, |h_d|)$ exists and is finite.

To prove this theorem, observe that $\mathcal{P}(y, |h_1|, \ldots, |h_d|)$ is a sum of elements of the form
\begin{align}
    b|y|^n |h_1|^{n_1} \ldots |h_d|^{n_d},
\end{align}
and, hence, it is enough to show that
\begin{align}
    E_\theta [b|y|^n |h_1|^{n_1} \ldots |h_d|^{n_d}]
\end{align}
exists and is finite. This last fact is an easy consequence of the existence of absolute non-central and central moments of Gaussians and, therefore, we will skip the proof.

APPENDIX B

AUXILIARY RESULTS AND DEFINITIONS

Theorem 6 (22, page 2129): Let $a(z, \theta)$ be a matrix of functions of an observation $z$ and the parameter $\theta$. If the $z_1, \ldots, z_N$ are i.i.d., $\Omega$ is compact, $a(z_i, \theta)$ is continuous at each $\theta$ and there is $d(z)$ with $\|a(z, \theta)\|_F \leq d(z)$ for all $\theta \in \Omega$, where $E[d(z)]$ exists and is finite, then $E[a(z, \theta)]$ is continuous and
\begin{align}
    \sup_{\theta \in \Omega} \left\| \frac{1}{N} \sum_{j=1}^N a(z_j, \theta) - E[a(z, \theta)] \right\|_F \to 0
\end{align}
in probability.

Let $X_n$ be a sequence of random vectors. We use the notation $X_n = o_P(1)$, to denote that $X_n$ converges to 0 in probability, i.e., if, for every $\epsilon > 0$, the non-random sequence $\mathbb{P}(\|X_n\| \leq \epsilon)$ converges to 1.

If $X_n$ is uniformly bounded in probability, i.e., if, for every $\epsilon > 0$, there exists $M(\epsilon) > 0$, such that
\begin{align}
    \mathbb{P}(\|X_n\| > M(\epsilon)) < \epsilon, \quad \forall n,
\end{align}
we denote this by $X_n = O_P(1)$ (see [22] for more details and also for the calculus with the $O_P(1)$ and $o_P(1)$).
APPENDIX C

PROOF OF LEMMA 1

Suppose $X^*$ satisfies (7). Throughout this proof, $x_i^*$ denotes the $i$th column of $X^*$. Consider the equation imposed by the first column, $x_1^*$, i.e.,

$$Cx_1^* = ((x_1^*)^TCx_1^*)x_1^*,$$

and multiply both sides by $(x_1^*)^T$, which yields

$$((x_1^*)^TCx_1^*)(1 - ||x_1^*||^2) = 0.$$  

From the two equalities

$$((x_1^*)^TCx_1^*)x_1^* = Cx_1^*,$$

$$((x_1^*)^TCx_1^*)(1 - ||x_1^*||^2) = 0,$$

we conclude that either $x_1^* = 0$ or $x_1^*$ is a unit-norm eigenvector of $C$.

Considering the second column, we prove that $x_2^*$ is either zero or a unit-norm eigenvector of $C$ that is orthogonal to $x_1^*$. Observe that

$$Cx_2^* = ((x_2^*)^TCx_2^*)x_1^* + ((x_2^*)^TCx_2^*)x_2^*.$$  

(49)

Now recall that $x_1^* = 0$ or $x_1^*$ is a unit-norm eigenvector of $C$. If $x_1^* = 0$, then (49) reduces to

$$Cx_2^* = ((x_2^*)^TCx_2^*)x_2^*$$

and the result follows as in the case of $x_1^*$. If $x_1^* \neq 0$, then it is a unit-norm eigenvector of $C$ and, hence, there exists $\beta$ such that $(x_1^*)^TC = \beta(x_1^*)^T$ and (49) reduces to

$$Cx_2^* = \beta((x_1^*)^TCx_2^*)x_1^* + ((x_2^*)^TCx_2^*)x_2^*.$$  

(50)

Multiply on the left by $(x_1^*)^T$ and use $||x_1^*||^2 = 1$ to obtain

$$((x_2^*)^TCx_2^*)(x_1^*)^Tx_2^* = 0.$$  

If $x_2^* = 0$, we are done. If not, then $0 = (x_1^*)^Tx_2^*$ and, returning to (50), it holds that

$$Cx_2^* = ((x_2^*)^TCx_2^*)x_2^*.$$  

This establishes the claim for $x_1^*$ and $x_2^*$. Proceeding as we did for the second column, it is possible to construct a proof by induction establishing the result.

ACKNOWLEDGMENT

This work was partially funded by the Portuguese Fundação para a Ciência e Tecnologia (FCT), under grants PD/BD/135185/2017 and UIDB/50008/2020. The work of João Xavier was supported in part by the Fundação para a Ciência e Tecnologia, Portugal, through the Project LARSyS, under Project FCT Project UIDB/50099/2020 and Project HARMONY PTDC/EEI-AUT/31411/2017 (funded by Portugal 2020 through FCT, Portugal, under Contract AAC n 2/SAICT/2017–031411. IST-ID funded by POR Lisboa under Grant LISBOA-01-0145-FEDER-031411).
Mário Figueiredo Mário Figueiredo received his PhD (1994) in Electrical and Computer Engineering from Instituto Superior Técnico, University of Lisbon, where he is an IST Distinguished Professor and holds the Feedzai Chair in Machine Learning. He is also a senior researcher and area coordinator at Instituto de Telecomunicações. His research areas include machine learning, signal processing, and optimization. He received several honors and awards, namely: Fellow of the IEEE, Fellow of the International Association for Pattern Recognition (IAPR), Fellow of the European Association for Signal Processing (EURASIP), W. R. G. Baker Award (IEEE), EURASIP Technical Achievement Award, member of the Portuguese Academy of Engineering and of the Lisbon Academy of Science.

João Xavier João Xavier received the Ph.D. degree in electrical and computer engineering from the Instituto Superior Técnico (IST), Lisbon, Portugal, in 2002. He is currently an Associate Professor with the Department of Electrical and Computer Engineering, IST. He is also a Researcher with the Institute of Systems and Robotics (ISR Lisboa), Lisbon. His current research interest includes optimization and statistical inference for distributed systems.