A DISTRIBUTED CLASSIFICATION/ESTIMATION ALGORITHM FOR SENSOR NETWORKS

FABIO FAGNANI†, SOPHIE M. FOSSON‡, AND CHIARA RAVAZZI§

Abstract. In this paper, we address the problem of simultaneous classification and estimation of hidden parameters in a sensor network with communications constraints. In particular, we consider a network of noisy sensors which measure a common scalar unknown parameter. We assume that a fraction of the nodes represent faulty sensors, whose measurements are poorly reliable. The goal for each node is to simultaneously identify its class (faulty or non-faulty) and estimate the common parameter.

We propose a novel cooperative iterative algorithm which copes with the communication constraints imposed by the network and shows remarkable performance. Our main result is a rigorous proof of the convergence of the algorithm, under a fixed communication graph, and a characterization of the limit behavior as the network size goes to infinity. In particular, we prove that, in the limit when the number of sensors goes to infinity, the common unknown parameter is estimated with arbitrary small error, while the classification error converges to that of the optimal centralized maximum likelihood estimator. We also show numerical results that validate the theoretical analysis and support their possible generalization. We compare our strategy with the Expectation-Maximization algorithm and we discuss trade-offs in terms of robustness, speed of convergence and implementation simplicity.

Key words. Classification, Consensus, Gaussian mixture models, Maximum-likelihood estimation, Sensor networks, Switching systems.

AMS subject classifications. 62F12, 62F15, 62H30, 93E10, 93C30

1. Introduction. Sensor networks are one of the most important technologies introduced in our century. Promoted by the advances in wireless communications and by the pervasive diffusion of smart sensors, wireless sensor networks are largely used nowadays for a variety of purposes, e.g., environmental and habitat surveillance, health and security monitoring, localization, targeting, event detection.

A sensor network basically consists in the deployment of a large numbers of small devices, called sensors, that have the ability to perform measurements and simple computations, to store few amounts of data, and to communicate with other devices. In this paper, we focus on ad hoc networks, in which communication is local: each sensor is connected only with a restricted number of other sensors. This kind of cooperation allows to perform elaborate operations in a self-organized way, with no centralized supervision or data fusion center, with a substantial energy and economic saving on processors and communication links. This allows to construct large sensor networks at contained cost.

A problem that can be addressed through ad hoc sensor networking is the distributed estimation: given an unknown physical parameter (e.g., the temperature in a room, the position of an object), one aims at estimating it using the sensing capabilities of a network. Each sensor performs a (not exact) measurement and shares...
it with the sensors with which it can establish a communication; in turn, it receives
information and consequently updates its own estimate. If the network is connected,
by iterating the sharing procedure, the information propagates and a consensus can
be reached. Neither centralized coordinator nor data fusion center is present. The
mathematical model of this problem must envisage the presence of noise in measure-
ments, which are naturally corrupted by inaccuracies, and possible constraints on the
network in terms of communication, energy or bandwidth limitations, and of necessity
of quantization or data compression.

Distributed estimation in ad hoc sensor networks has been widely studied in the
literature. For the problem of estimating an unknown common parameter, typical
approach is to consider distributed versions of classical maximum likelihood (ML)
or maximum-a-posteriori (MAP) estimators. Decentralization can be obtained, for
instance, through consensus type protocols (see [1], [2], [3], [4], [5]) adapted to the
communication graph of the network, or by belief propagation methods [6] and [7].

A second important issue is sensors’ classification, which we define as follows [8].
Let us imagine that sensors can be divided into different classes according to peculiar
properties, e.g., measurements’ or processing capabilities, and that no sensor knows
to which class it belongs: by classification, we then intend the labeling procedure
that each sensor undertakes to determine its affiliation. This task is addressed to a
variety of clustering purposes, for example, to rebalance the computation load in a
network where sensors can be distinguished according to their processing power. On
most occasions, sensors’ classification is faced through some distributed estimation,
the underlying idea being the following: each sensor performs its measurement of a
parameter, then iteratively modifies it on the basis of information it receives; during
this iterative procedure the sensor learns something about itself which makes it able
to estimate its own configuration.

In this paper, we consider the following model: each sensor \( i \) performs a measure-
ment \( y_i = \theta^* + \omega^*_i \eta_i \), where \( \theta^* \in \mathbb{R} \) is the unknown global parameter, \( \omega^*_i > 0 \)
is the unknown status of the sensor, and \( \eta_i \) is a Gaussian random noise. The larger \( \omega^*_i \) is,
the more the sensor \( i \) is malfunctioning, that is, the quality of its measurement is low.
The \( \omega^*_i \) parameter is supposed to belong to a discrete set, in particular in this paper
we consider the binary case.

The goal of each unit \( i \) is to estimate the parameter \( \theta^* \) and the specific configu-
ration \( \omega^*_i \). The presence of the common unknown parameter \( \theta^* \) imposes a coupling
between the different nodes and makes the problem interesting.

An additive version of the aforementioned model has been studied in [9], where
measurement is given by \( y_i = \theta^* + \omega^*_i + \eta_i \). Another related problem is the so-called
calibration problem [10, 11]: sensor \( i \) performs a noisy linear measurement \( y_i = A_i \theta + \eta_i \),
where the unknown \( \theta \) and \( A_i \) are a vector and a matrix, respectively, while \( \eta_i \) is a
noise; the goal consists in the estimation of \( \theta \) and of \( A_i \), the latter being known as
calibration problem.

All these are particular cases of the problem of the estimation of Gaussian mix-
tures’ parameters [12, 13]. This perspective has been studied for sensor networks
in [14], [15], [16], and [17] where distributed versions of the Expectation-Maximization
(EM) algorithm have been proposed. A network is given where each node independ-
dently performs the E-step through local observations. In particular, in [16] a
consensus filter is used to propagate the local information. The tricky point of such
techniques is the choice of the number of averaging iterations between two consecutive
M-steps, which must be sufficient to reach consensus.
The aim of this paper is the development of a distributed, iterative procedure which copes with the communication constraints imposed by the network, which are supposed to be fixed, and computes an estimation \((\hat{\theta}, \hat{\omega})\) approximating the maximum likelihood optimal solution of the proposed problem. The core of our methodology is an Input Driven Consensus Algorithm (IA for short), inspired by [6] and introduced in [18], which takes care of the estimation of the parameter \(\theta^*\). IA is coupled with a classification step where nodes update the estimation of their own type \(\omega_i^*\) by a simple threshold estimator based on the current estimation of \(\theta^*\). The algorithm is based on a consensus protocol driven by an input signal depending on the original measurements and current estimates of the variances. We however remark that our algorithm is not a classical dynamic consensus protocol [2,4,5,19] driven by an exogenous, time-varying input to be tracked: the innovation we introduce depends in fact on the state of the algorithm and is built on purpose to improve the procedure. Moreover, our input is not a function regularly varying in time (as generally considered in dynamic consensus problems), but a switching signal, which makes the classical dynamic consensus methods unfeasible.

Our main theoretical contribution is a complete analysis of the algorithm in terms of convergence and of behavior with respect to the size of the network. With respect to other approaches like distributed EM for which convergence results are missing, this makes an important difference. We also present a number of numerical simulations showing the remarkable performance of the algorithm which, in many situations, outperforms classical choices like EM.

The outline of the paper is the following. In Section 2 we shortly present some graph nomenclature needed in the paper. Section 3 is devoted to a formal description of the problem and to a discussion of the classical centralized maximum likelihood solution. In Section 4, we present the details and the analysis of our IA. Our main results are Theorems 4.1 and 4.2: Theorem 4.1 ensures that, under suitable assumptions on the graph, the algorithm converges to a local maximum of the log-likelihood function; Theorem 4.2 is a concentration result establishing that when the number of nodes \(N \to +\infty\), the estimate \(\hat{\theta}\) converges to the true value \(\theta^*\) (a sort of asymptotic consistency). Finally, we also study the behavior of the relative classification error over the network when \(N \to +\infty\) (see Corollary 4.4). Section 5 contains a set of numerical simulations carried on different graph architectures: complete, circulant, grids, and random geometric graphs. Comparisons are proposed with respect to the optimal centralized ML solution and also with respect to the EM solution. Finally, a long Appendix contains all the proofs.

2. General notation and graph theoretical preliminaries. Throughout this paper, we use the following notational convention. We denote vectors with small letters, and matrices with capital letters. Given a matrix \(M\), \(M^T\) denotes its transpose. Given a vector \(v\), \(||v||\) denotes its Euclidean norm. \(1_A\) is the indicator function of set \(A\). Given a finite set \(V\), \(R^V\) denotes the space of real vectors with components labelled by elements of \(V\). Given two vectors \(x, z \in R^V\), \(d_H(x, z) = |\{i \in V : x_i \neq z_i\}|\). We use the convention that a summation over an empty set of indices is equal to zero, while a product over an empty set gives one.

Given a set \(S \subseteq R^n\), we denote the interior and the closure of \(S\) with \(\text{int}(S)\) and \(\bar{S}\), respectively. A symmetric graph is a pair \(\mathcal{G} = (V, E)\) where \(V\) is a set, called the set of vertices, and \(E \subseteq V \times V\) is the set of edges with the property that \((i, i) \notin E\) for all \(i \in V\) and \((i, j) \in E\) implies \((j, i) \in E\). \(\mathcal{G}\) is strongly connected if, for all \(i, j \in V\), there exist vertices \(i_1, \ldots, i_s\) such that \((i, i_1), (i_1, i_2), \ldots, (i_s, j) \in E\). To any
symmetric matrix $P \in \mathbb{R}^{V \times V}$ with non-negative elements, we can associate a graph $G_P = (V, E_P)$ by putting $(i, j) \in E_P$ if and only if $P_{ij} > 0$. $P$ is said to be adapted to a graph $G$ if $G_P \subseteq G$. A matrix with non-negative elements $P$ is said to be stochastic if $\sum_{j \in V} P_{ij} = 1$ for every $i \in V$. Equivalently, denoting by $1$ the vector of all 1 in $\mathbb{R}^V$, $P$ is stochastic if $P1 = 1$. $P$ is said to be primitive if there exists $n_0 \in \mathbb{N}$ such that $P_{ij}^{n_0} > 0$ for every $i, j \in V$. A sufficient condition ensuring primitivity is that $G_P$ is strongly connected and $P_{ii} > 0$ for some $i \in V$.


3.1. The model. In our model, we consider a network, represented by a symmetric graph $G = (V, E)$. $G$ represents the system communication architecture. We denote the number of nodes by $N = |V|$. We assume that each node $i \in V$ measures the observable

$$y_i = \theta^* + \omega^*_i \eta_i$$

(3.1)

where $\theta^* \in \mathbb{R}$ is an unknown parameter, $\eta_i$’s Gaussian noises $N(0, 1)$, $\omega^*_i$’s Bernoulli random variables taking values in $\{\alpha, \beta\}$ (with $P(\omega^*_i = \beta) = p$). We assume all the random variables $\eta_i$’s and $\omega^*_i$’s to be mutually independent. Notice that each $y_i \in \mathbb{R}$ is a Gaussian mixture distributed according to the probability density function

$$f(y_i) = (1 - p)f(y_i|\theta^*, \alpha) + pf(y_i|\theta^*, \beta)$$

(3.2)

$$f(y_i|\theta^*, x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{(y_i - \theta^* x)^2}{2\sigma^2}} \quad x \in \{\alpha, \beta\}.$$  

(3.3)

The binary model of $\omega^*$ is motivated by different scenarios: as an example, if $0 < \alpha << \beta$, the nodes of type $\beta$ may represent a subset of faulty sensors, whose measurements are poorly reliable; the aim may be the detection of faulty sensors in order to switch them off or neglect their measurements, or for other clustering purposes. It is also realistic to assume that some a-priori information about the quantity of faulty sensors is extracted, e.g., from experimental data on the network, and it is conceivable to represent such information as an a-priori distribution. This is why we assume a Bernoulli distribution on each $\omega^*_i$; on the other hand, we suppose that no a-priori information is available on the unknown parameter $\theta^*$. However, the addition of an a priori probability distribution on $\theta^*$ does not significantly alter our analysis and our results.

3.2. The maximum likelihood solution. The goal is to estimate the parameter $\theta^*$ and the specific configuration $\omega^*_i$ of each unit. Disregarding the network constraints, a natural solution to our problem would be to consider a joint ML in $\theta^*$ and MAP in the $\omega^*_i$’s (see [20, 21]). Let $f(y, \omega|\theta)$ be the joint distribution of $y$ and $\omega$ (density in $y$ and probability in $\omega$) given the parameter $\theta$, and consider the rescaled log-likelihood function

$$L_N(\theta, \omega) := \frac{1}{N} \log f(y, \omega|\theta).$$

(3.4)

The hybrid ML/MAP solution, which for simplicity for now on we will refer to as the ML solution, prescribes to choose $\theta$ and $\omega$ which maximize $L_N(\theta, \omega)$

$$\tilde{\theta}^{ML}, \tilde{\omega}^{ML} := \arg\max_{\theta \in \mathbb{R}, \omega \in \{\alpha, \beta\}^V} L_N(\theta, \omega).$$

(3.5)
Standard calculations lead to the following results.

**Proposition 3.1.** The rescaled log-likelihood function can be expressed as follows

\[
L_N(\theta, \omega) = -\frac{1}{N} \sum_{j \in V} \left( \frac{(y_j - \theta)^2}{2\beta^2} + 1_{\{\omega_j = \alpha\}} \left( \frac{(y_j - \theta)^2}{2\alpha^2} - \log \frac{1 - p\beta}{p\alpha} \right) \right) + c.
\]

(3.6)

where \(c\) is a constant.

**Proof.** Since \(\{y_i\}_{i \in V}\) are i.i.d. random variables, we obtain \(f(y|\omega, \theta) = \prod_{i \in V} f(y_i|\omega_i, \theta)\). Let us define \(V_\alpha = \{v \in V|\omega_v = \alpha\}\). We thus have

\[
f(y|\omega, \theta) \propto \prod_{i \in V_\alpha} \left[ \frac{1 - p}{\alpha} e^{-\frac{(y_i - \theta)^2}{2\alpha^2}} \right] \prod_{i \in V \setminus V_\alpha} \left[ \frac{p}{\beta} e^{-\frac{(y_i - \theta)^2}{2\beta^2}} \right]
\]

\[
\propto e^{-\sum_{i \in V_\alpha} \left[ \frac{(y_i - \theta)^2}{2\alpha^2} - \log \left( \frac{1 - p}{\alpha} \right) \right]} e^{-\sum_{i \in V \setminus V_\alpha} \left[ \frac{(y_i - \theta)^2}{2\beta^2} - \log \left( \frac{p}{\beta} \right) \right]}
\]

\[
\propto e^{-\sum_{i \in V} \left[ \frac{(y_i - \theta)^2}{2\beta^2} + 1_{\omega_i = \alpha} \left( \frac{(y_i - \theta)^2}{2\alpha^2} - \log \left( \frac{1 - p\beta}{p\alpha} \right) \right) \right]}
\]

from which the thesis follows immediately. \(\square\)

It should be noted that partial maximizations of \(L_N(\theta, \omega)\) with respect to just one of the two variables have simple representation. Let

\[
\hat{\theta}(\omega) := \arg\max_{\theta} L_N(\theta, \omega) \
\hat{\omega}(\theta) := \arg\max_{\omega} L_N(\theta, \omega).
\]

(3.7)

Then

\[
\hat{\theta}(\omega) = \frac{\sum_j y_j/\omega_j^2}{\sum_j 1/\omega_j^2} \quad \hat{\omega}(\theta) = \begin{cases} 
\alpha & \text{if } |y_i - \theta| < \delta \\
\beta & \text{otherwise}
\end{cases}
\]

(3.8)

where

\[
\delta = \sqrt{\frac{\ln \left( \frac{1 - p\beta}{p\alpha} \right)}{2\frac{1 - p\beta}{p\alpha} - \frac{1 - p}{\alpha^2}}}
\]

The ML solution can then be obtained, for instance, by considering

\[
\hat{\theta}^{ML} = \arg\max_{\theta} L_N(\theta, \hat{\omega}(\theta)) \quad \hat{\omega}^{ML} = \hat{\omega}(\hat{\theta}^{ML}).
\]

(3.9)

The computation of the \((\hat{\omega}^{ML})\)'s becomes totally decentralized once \(\hat{\theta}^{ML}\) has been computed. For the computation of \(\hat{\theta}^{ML}\) instead one needs to gather information from all units to compute \(L_N(\theta, \hat{\omega}(\theta))\) and it is not at all evident how this can be done in a decentralized way. Moreover, further difficulties are caused by the fact that \(L_N(\theta, \hat{\omega}(\theta))\) may contain many local maxima, as shown in Figure 3.1.

It should be noted that \(L_N(\theta, \hat{\omega}(\theta))\) is differentiable except at a finite number of points, and between two successive non-differentiable points the function is concave. Therefore, the local maxima of the function coincide with its critical points. On the other hand, the derivative, where it exists, is given by

\[
\frac{d}{d\theta} L_N(\theta, \hat{\omega}(\theta)) = \left( \frac{1}{\beta^2} - \frac{1}{\alpha^2} \right) \frac{1}{N} \sum_{i \in V} (\theta - y_i) 1_{\{|y_i - \theta| < \delta\}} - \frac{1}{\beta^2} \left( \theta - \frac{1}{N} \sum_{i \in V} y_i \right).
\]

(3.10)
Stationary points can therefore be represented by the relation

\[
\theta = \frac{1}{N} \sum_i y_i + \left( \frac{1}{N} - \frac{1}{\beta^2} \right) \sum_i y_i \mathbb{1}_{|y_i - \theta| < \delta}.
\]

(3.11)

A moment of thought shows us that (3.11) is equivalent to the relation \( \theta = \hat{\theta}(\hat{\omega}(\theta)) \).

This representation will play a key role in the sequel of this paper.

---

**Figure 3.1.** \( \alpha = 0.3, \beta = 10, p = 0.25 \): Plot of function \( L_N(\theta, \hat{\omega}(\theta)) \) as a function of \( \theta \) and size \( N \in \{50, 100, 400, 500, 1000, 5000\} \).

We remark that our model envisages only one measurement per node, even if it is clear that repeated measurements would increase the available information and improve the estimation accuracy. This paper focuses instead on the role of cooperation in such estimation problems showing its efficiency in achieving desirable performance even if starting from one measurement per node. We also remark that considering only one measurement keeps the mathematical analysis as readable as possible, but without entailing any theoretical loss of generality. Generalizations of such model are object of our current research, as discussed in the conclusive section.

### 3.3. Iterative centralized algorithms.

The computational complexity of the optimization problem (3.5) is practically unfeasible in most situations. However, relations (3.8) suggest a simple way to construct an iterative approximation of the ML solution (which we will denote IML). The formal pattern is the following: fixed \( \hat{\omega}^{(0)} = \alpha \mathbb{1} \), for \( t = 0, 1, \ldots \), we consider the dynamical system

\[
\hat{\theta}^{(t+1)} = \frac{\sum_{j=1}^{N} y_j \left[ \hat{\omega}_j^{(t)} \right]^{-2}}{\sum_{j=1}^{N} \left[ \hat{\omega}_j^{(t)} \right]^{-2}}
\]

\[
\hat{\omega}_i^{(t+1)} = \begin{cases} 
\alpha & \text{if } |y_i - \theta| < \delta \\
\beta & \text{otherwise}
\end{cases}
\]

for any \( i = 1, \ldots, N \).

The algorithm stops whenever \( |\hat{\theta}^{(t+1)} - \hat{\theta}^{(t)}| < \varepsilon \), for some fixed tolerance \( \varepsilon > 0 \).

A more refined iterative solution is given by the so-called Expectation-Maximization (EM) algorithm \[22\]. The main idea is to introduce a hidden (say, unknown and unobserved) random variable in the likelihood; then, at each step, one computes the
mean of the likelihood function with respect to the hidden variable and finds its maximum. Such a method seeks to find the maximum likelihood solution, which in many cases cannot be formulated in a closed form. EM is widely and successfully used in many frameworks and in principle it could also be applied to our problem. In our context, making the variable \( \omega \) to play the part of the hidden variable, equations for EM become (see the tutorial [23] for their derivation)

**Given** \( \hat{\theta}^{(0)} \in \mathbb{R} \), for \( t = 0, 1, \ldots \),

1. **E-step:** for all node \( i \in V \),

\[
q_i^{(t)} = \mathbb{P} \left( \hat{\omega}_i^{(t)} = \alpha | y, \hat{\theta}^{(t)} \right) = \frac{(1 - p) f \left( y | \hat{\omega}_i^{(t)} = \alpha, \hat{\theta}^{(t)} \right)}{(1 - p) f \left( y | \hat{\omega}_i^{(t)} = \alpha, \hat{\theta}^{(t)} \right) + pf \left( y | \hat{\omega}_i^{(t)} = \beta, \hat{\theta}^{(t)} \right)}.
\]

2. **M-step:**

\[
\hat{\theta}^{(t+1)} = \sum_{i \in V} y_i \left( q_i^{(t)} \alpha^{-2} + (1 - q_i^{(t)}) \beta^{-2} \right) / \sum_{i \in V} q_i^{(t)} \alpha^{-2} + (1 - q_i^{(t)}) \beta^{-2}.
\]

The algorithm stops whenever \( |\hat{\theta}^{(t+1)} - \hat{\theta}^{(t)}| < \varepsilon \), for some fixed tolerance \( \varepsilon > 0 \). It is worth to notice that \( q_i^{(t)} \) computed in the E-step actually is the expectation of the binary random variable \( 1 \{ \hat{\omega}_i^{(t)} = \alpha \} \). On the other hand \( \hat{\theta}^{(t+1)} \) computed in the M-step is the maximum of such expectation.

An important feature of EM is that it is possible to prove the convergence of the sequence \( \{\hat{\theta}^{(t)}\}_{t \in \mathbb{N}} \) to a local maximum of the expected value of the log-likelihood with respect to the unknown data \( \omega \), a result which is instead not directly available for IML. Both algorithms however share the drawback of requiring centralization. Distributed versions of the EM have been proposed (see, e.g., [14], [16]) but convergence is not guaranteed for them. In Section 5 we will compare both these algorithms against the distributed IA we are going to present in the next section. While it is true that EM always outperforms IML, algorithm IA outperforms both of them for small size algorithms, while shows comparable performance to EM for large networks.

**4. Input Driven Consensus Algorithm.**

**4.1. Description.** In this section we propose a distributed iterative algorithm approximating the centralized ML estimator. The algorithm is suggested by the expressions in (3.8) and consists of the iteration of two steps: an averaging step where all units aim at computing \( \hat{\theta} \) through a sort of input driven consensus algorithm followed by an update of the classification estimation performed autonomously by all units.

Formally, IA is parametrized by a symmetric stochastic matrix \( P \), adapted to the communication graph \( G \) (\( P_{ij} > 0 \) if and only if, \((i, j) \in E\)), and by a real sequence \( \gamma^{(t)} \rightarrow 0 \). Every node \( i \) has three messages stored in its memory at time \( t \), denoted with \( \mu_i^{(t)}, \nu_i^{(t)}, \) and \( \hat{\omega}_i^{(t)} \). Given the initial conditions \( \mu_i^{(0)} = 0, \nu_i^{(0)} = 0 \) and the initial estimate \( \hat{\omega}_i^{(0)} = \alpha \), the dynamics consists of the following steps.
1. **Average step:**

\[
\begin{align*}
\mu_i^{(t+1)} &= (1 - \gamma^{(t)}) \sum_j P_{ij} \mu_j^{(t)} + \gamma^{(t)} y_i \left( \hat{\omega}_i^{(t)} \right)^{-2} \\
\nu_i^{(t+1)} &= (1 - \gamma^{(t)}) \sum_j P_{ij} \nu_j^{(t)} + \gamma^{(t)} \left( \hat{\omega}_i^{(t)} \right)^{-2} \\
\hat{\theta}_i^{(t+1)} &= \mu_i^{(t+1)} / \nu_i^{(t+1)}. 
\end{align*}
\]

(4.1a) (4.1b) (4.1c)

2. **Classification step:**

\[
\hat{\omega}_i^{(t+1)} = \hat{\omega}_i(\hat{\theta}^{(t+1)}) = \begin{cases} 
\alpha & \text{if } |y_i - \hat{\theta}_i^{(t+1)}| < \delta \\
\beta & \text{otherwise.}
\end{cases} 
\]

(4.2)

It should be noted that the algorithm provides a distributed protocol: each node only needs to be aware of its neighbors and no further information about the network topology is required.

**4.2. Convergence.** The following theorem ensures the convergence of IA. The proof is rather technical and therefore deferred to Appendix A.

**Theorem 4.1.** Let

(a) \(\gamma^{(t)} \in (0, 1)\), \(\gamma^{(t)} \to 0\), \(\gamma^{(t)} \geq 1/t\), and \(\gamma^{(t)} = \gamma^{(t+1)} + o(\gamma^{(t+1)})\) for \(t \to \infty\);

(b) \(P \in \mathbb{R}^{V \times V}\) be a stochastic, symmetric, and primitive matrix with positive eigenvalues.

Then, there exist \(\hat{\gamma}^{IA} \in \{\alpha, \beta\}^V\) and \(\hat{\theta}^{IA} \in \mathbb{R}\) such that

1. \[
\lim_{t \to +\infty} \hat{\omega}_i^{(t)} \overset{a.s.}{=} \hat{\omega}^{IA}, \quad \lim_{t \to +\infty} \hat{\theta}_i^{(t)} \overset{a.s.}{=} \hat{\theta}^{IA}
\]

for all \(i \in V\);

2. they satisfy the relations

\[
\hat{\theta}^{IA} = \hat{\theta}(\hat{\omega}^{IA}), \quad \hat{\omega}^{IA} = \hat{\omega}(\hat{\theta}^{IA}).
\]

A number of remarks are in order.

- The assumption on the eigenvalues of \(P\) is essentially a technical one: in simulations it does not seem to have a crucial role, but we need it in our proof of convergence. On the other hand, given any symmetric stochastic primitive \(P\), we can consider a ‘lazy’ version of it \(P_\tau = (1 - \tau)I + \tau P\) and notice that for \(\tau \in (0, 1)\) sufficiently small, indeed \(P_\tau\) will satisfy the assumption on the eigenvalues.

- The requirement \(\gamma^{(t)} \geq 1/t\) is not new in decentralized algorithms (see for instance the Robbins-Monro algorithm, introduced in [24]) and serves the need of maintaining ‘active’ the system input for sufficiently long time. Less classical is the assumption \(\gamma^{(t)} \sim \gamma^{(t+1)}\) which is essentially a request of regularity in the decay of \(\gamma^{(t)}\) to 0. Possible choices of \(\gamma^{(t)}\) satisfying the above conditions are \(\gamma^{(t)} = t^{-\zeta}\) for \(\zeta \in (0, 1)\), or \(\gamma^{(t)} = t^{-1}(\ln t)^{\alpha}\) for any \(\alpha > 0\).

- The proof (see Appendix A) will also give an estimation on the speed of convergence: indeed it will be shown that \(||\hat{\theta}^{(t)} - \hat{\theta}^{IA}|| = O(\gamma^{(t)})\) for \(t \to \infty\).

- Relations in item 2. implies that \(\hat{\theta}^{IA}\) is a local maximum of the function \(L_N(\theta, \hat{\omega}(\theta))\) (see (3.11)).
4.3. Limit behavior. In this section we present results on the behavior of our algorithm for $N \to +\infty$. All quantities derived so far are indeed function of network size $N$. In order to emphasize the role of $N$, we will add an index $N$ when dealing with quantities like $\theta^*$ (e.g. $\hat{\theta}^{\text{ML}}_N$ is the scalar estimate, which is computed using $N$ measurements $\{y_i\}_{i=1}^N$). Instead we will not add anything to expressions where there are vectors $\omega$ involved since their dimension is itself $N$.

Figure 3.1 shows a sort of concentration of the local maxima of $L_N(\theta, \tilde{\omega}(\theta))$ to a global maximum for large $N$. Considering that IA converges to a local maximum, this observation would lead to the conclusion that, for large $N$, the IA resembles the optimal ML solution. This section provides some results which make rigorous these considerations.

Notice first that, applying the uniform law of large numbers \[\text{(25)}\] to the expression \((3.6)\), we obtain that, for any compact $K \subseteq \mathbb{R}$, almost surely
\[
\lim_{N \to +\infty} \max_{\theta \in K} \left| \int_{\mathbb{R}} J(s, \theta) f(s) ds \right| = 0
\]
where
\[
J(s, \theta) = -\left\{ \frac{(s-\theta)^2}{2\beta^2} + 1_{|s-\theta|<\delta} \left( \frac{(s-\theta)^2}{2 \left( \frac{1}{\alpha^2} - \frac{1}{\beta^2} \right)} - \log \frac{1-p}{p} \right) \right\} + c
\]
(4.4)
where $c$ is the same constant as in \((3.6)\). The limit function $\int_{\mathbb{R}} J(s, \theta) f(s) ds$ turns out to be differentiable for every value of $\theta$ and to have a unique stationary point for $\theta = \theta^*$ which turns out to be the global maximum. Unfortunately, this fact by itself does not guarantee that global and local maxima will indeed converge to $\theta^*$. In our derivations the properties of the function $\int_{\mathbb{R}} J(s, \theta) f(s) ds$ will not play any direct role and therefore they will not be proven here. The main technical result which will be proven in Appendix B is the following:

**Theorem 4.2.** Denote by $S_N$ the set of local maxima of $L_N(\theta, \tilde{\omega}(\theta))$. Then,
\[
\lim_{N \to +\infty} \max_{\xi \in S_N} \left| \xi - \theta^* \right| = 0
\]
(4.5)
almost surely and in mean square sense.

This has an immediate consequence,

**Corollary 4.3.**
\[
\lim_{N \to +\infty} \hat{\theta}^{\text{IA}}_N = \lim_{N \to +\infty} \hat{\theta}^{\text{ML}}_N = \theta^*
\]
(4.6)
almost surely and in mean square sense.

Regarding the classification error, we have instead the following result:

**Proposition 4.4.**
\[
\lim_{N \to +\infty} \frac{1}{N} \mathbb{E}d_H(\tilde{\omega}(\theta^*), \omega^*) = \lim_{N \to +\infty} \frac{1}{N} \mathbb{E}d_H(\tilde{\omega}(\theta^*), \omega^*)
\]
\[
= q(p, \alpha, \beta)
\]
(4.7)
where
\[
q(p, \alpha, \beta) = (1-p) \text{erfc} \left( \frac{\delta}{\alpha \sqrt{2}} \right) + p \left[ 1 - \text{erfc} \left( \frac{\delta}{\beta \sqrt{2}} \right) \right]
\]
and \( \text{erfc}(x) := \frac{2}{\sqrt{\pi}} \int_{x}^{+\infty} e^{-t^2} dt \) is the complementary error function.

These results ensure that the IA performs, in the limit of large number of units \( N \), as the centralized optimal ML estimator. Moreover, they also show, consistency in the estimation of the parameter \( \theta^\star \). As expected, for \( N \to +\infty \) the classification error does not go to 0 since the increase of measurements is exactly matched by the same increase of variables to be estimated. Consistency however is obtained when \( p \) goes to zero since we have that \( \lim_{p \to 0} q(p, \alpha, \beta) = 0 \). Moreover, notice that the dependence of function \( q \) on the parameters \( \alpha \) and \( \beta \) is exclusively through their ratio \( \beta/\alpha \). In particular, we have

\[
\lim_{\beta/\alpha \to +\infty} q(p, \alpha, \beta) = 0 \quad \lim_{\beta/\alpha \to 1} q(p, \alpha, \beta) = 1.
\]

5. Simulations. In this section, we propose some numerical simulations to show the properties of IA in terms of convergence, velocity, and performance. We test it for different graph architectures and dimensions, and we compare its performance with the IML and EM algorithms described in Section 3.3. Our goal is to give evidence of the theoretical results’ validity and also to evaluate cases that are not included in our analysis: the good numerical outcomes we obtain suggest that convergence should hold in broader frameworks.

In the next, we will refer to stabilization as the step at which the classification stabilizes, and to convergence time as the time when the estimates converge and consensus is achieved.

The numerical setting for our simulations is now presented.

The sensors perform measurements according to the model (3.1) with \( \theta^\star = 0 \), \( \alpha = 0.3 \), \( \beta = 10 \); the prior probability \( P(\omega^\star_i = \beta) \) is equal to \( p = 0.25 \).

Given a strongly connected symmetric graph \( G = (V, E) \), we use the so-called Metropolis random walk construction for \( P \) (see [26]) which amounts to the following: if \( i \neq j \),

\[
P_{ij} = \begin{cases} 
0 & \text{if } (i, j) \notin E \\
\left( \max\{\deg(i) + 1, \deg(j) + 1\} \right)^{-1} & \text{if } (i, j) \in E 
\end{cases}
\]

where \( \deg(i) \) denotes the degree (the number of neighbors) of unit \( i \) in the graph \( G \). \( P \) constructed in this way is automatically irreducible and aperiodic.

We consider the following topologies:

1. Complete graph: \( P_{ij} = \frac{1}{N} \) for every \( i, j = 1, \ldots, N \); it actually corresponds to the centralized case.

2. Ring graph: \( N \) agents are disposed on a circle, and each agent communicates with its first neighbor on each side (left and right). The corresponding circulant symmetric matrix \( P \) is given by \( P_{ij} = \frac{1}{3} \) for every \( i = 2, \ldots, N - 1 \) and \( j \in \{i - 1, i, i + 1\} \); \( P_{11} = P_{12} = P_{1N} = \frac{5}{9} \); \( P_{N1} = P_{NN-1} = P_{NN} = \frac{3}{9} \); \( P_{ij} = 0 \) elsewhere.

From Theorem 4.1, \( P \) is required to possess positive eigenvalues: our intuition is that these hypotheses, that are useful to prove the convergence of the IA, are not really necessary. We test this conjecture on the ring graph, whose eigenvalues are known [27] to be \( \lambda_m = \frac{1}{4} \left( 1 + 2 \cos \left( \frac{2\pi m}{N} \right) \right) \), \( m = 0, \ldots, N - 1 \) and which are not necessarily positive.

3. Torus-grid graph: sensors are deployed on a two dimensional grid and are each connected with their four neighbors; the last node of each row of the grid is connected with the first node of the same row, and analogously on columns, so that a torus is obtained. The so-obtained graph is regular.
**Table 5.1**

<table>
<thead>
<tr>
<th>Step</th>
<th>Estimation/Classification</th>
</tr>
</thead>
</table>
| \( t = 0 \) | \( \theta^{IA}(0) = (8.28, 0.19, -0.66, -11.04, 0.38, -0.43, -5.54, -0.51, 5.87, 0.49) \)  
| \( \hat{\omega}^{IA}(0) = (\alpha, \alpha, \alpha, \alpha, \alpha, \alpha, \alpha, \alpha, \alpha, \alpha) \) |
| \( t = 2 \) | \( \theta^{IA}(2) = (6.73, 0.90, -1.59, -8.91, -0.81, -0.85, -4.54, -0.38, 4.72, 1.78) \)  
| \( \hat{\omega}^{IA}(2) = (\beta, \alpha, \beta, \beta, \alpha, \alpha, \beta, \alpha, \beta, \beta) \) |
| \( t = 4 \) | \( \theta^{IA}(4) = (2.26, 0.31, -1.07, -3.52, -2.28, -0.97, -0.94, -0.37, 1.30, 3.19) \)  
| \( \hat{\omega}^{IA}(4) = (\beta, \alpha, \alpha, \beta, \alpha, \alpha, \beta, \alpha, \beta, \beta) \) |
| \( t = 6 \) | \( \theta^{IA}(6) = (0.76, 0.12, -0.53, -1.10, -1.14, -0.62, -0.60, -0.43, 0.17, 1.20) \)  
| \( \hat{\omega}^{IA}(6) = (\beta, \alpha, \alpha, \beta, \alpha, \beta, \alpha, \beta, \alpha, \alpha) \) |
| \( t = 10 \) | \( \theta^{IA}(10) = (0.20, 0.01, -0.44, -0.51, -0.51, -0.46, -0.48, -0.44, -0.02, 0.37) \)  
| \( \hat{\omega}^{IA}(10) = (\beta, \alpha, \alpha, \beta, \alpha, \beta, \alpha, \beta, \alpha, \alpha) \) |

**Example 1:** Classification and estimation obtained by IA during 10 iterations.

4. *Random Geometric Graph* with radius \( r = 0.3 \): sensors are deployed in the square \([0, 1] \times [0, 1]\), their positions being randomly generated with a uniform distribution; a link is switched on between two sensors whenever their reciprocal distance is less than \( r \). We envisage only connected realizations.

In the next, we will analyze our simulations’ results in terms of convergence time and estimation performance, extracting information from 500 Monte Carlo runs for each different setting. Before this, let us present a single example to give an idea of the algorithm’s behavior.

**Example 1.** Consider a ring with \( N = 10 \) nodes; let \( \theta^* = 0 \) and \( \omega^* = (\beta, \alpha, \alpha, \beta, \alpha, \beta, \alpha, \beta, \beta)^T \). Suppose sensors perform the following measures:

\[
y = (8.284, 0.198, -0.661, -11.046, 0.385, -0.435, -5.537 - 0.517, 5.870, 0.495)^T.
\]

![Log-likelihood function (black curve) and estimation paths of three different sensors](image)

**Figure 5.1.** Example 1: Log-likelihood function (black curve) and estimation paths of three different sensors (green, blue and pink curves).

*Table 5.1* collects the estimates and the classification obtained by IA with \( \gamma^{(t)} = 1/\sqrt{T} \) and initial conditions \( \hat{\theta}^{(1)} = y, \hat{\omega}^{(1)} = \alpha \mathbf{I} \) during 10 iterations. We report the results at those steps at which the classification changes. We notice that the classification stabilizes rapidly (after 10 iterations) and is exact. The estimates’ consensus time is instead longer and comparable to the decrease of the parameter \( \gamma^{(t)} \) (see Section A.1).
In Figure 5.1 the likelihood function (black curve) and the estimates provided by three different sensors are plotted: the green, blue, and pink curves respectively represent the estimation paths of sensors whose initial measurements are −11.04, −0.43, and 5.87 (see Figure 5.1). The markers individuate the iteration steps. We can see that the estimates converge to the global maximum of the log-likelihood, which corresponds to $\hat{\theta}^{ML}$ and is the most desirable result.

We remark that in this example, convergence is numerically shown for the ring topology, which is not envisaged by our theoretical analysis. Hence, our guess is that convergence should be proved even under weaker hypotheses on matrix $P$.

![Figure 5.2. Ring: average stabilization times as a function of $N$.](image)

5.1. Stabilization and convergence times. As pointed out in the Example 1, stabilization is very quick with respect to convergence. We analyze this fact by collecting results from 500 Monte Carlo Runs. We consider again the ring topology, which is expected to be the slower one, due to its low connectivity. In Figure 5.2, the average stabilization times are collected as a function of the network size, for different values of the parameter $\zeta$: we see that the increase is almost linear with $N$. As soon as the classification is competed, the convergence time to reach consensus is comparable to the decrease of the parameter $\gamma(t)$ for fixed network size $N$. On the other hand, an inspection of the convergence proofs (see Section A.1) shows that the more the graph is connected the better convergence rate.

5.2. Performance analysis. We implement and compare the following algorithms: IA with $\gamma(t) = 1/t^\zeta$ for different choices of $\zeta \in \{0.5, 0.7, 0.9\}$, IML, and EM (see Section 3.3).

We show the performance of the aforementioned algorithms in terms of classification error and of mean square error on the global parameter, in function of the number of sensors $N$. All the outcomes are obtained averaging over 500 Monte Carlo runs.

We observe that the classification error (Figure 5.3) converges to the value $q(p, \alpha, \beta)$ defined in (4.7) (see the red line) when $N \to \infty$ for all the considered algorithms. On the other hand, when $N$ is small, IA performs better than IML and EM for almost all the considered topologies: this suggests that decentralization is then not a drawback for IA. In particular, we notice that performance improve when connectivity is lower.

Furthermore, the role of $\gamma(t)$ is considerable. In Figure 5.3, we can see that if the exponent $\zeta$ of $\gamma(t) = t^{-\zeta}$ decreases (i.e., if $\gamma(t)$ vanishes more slowly) the performance
of IA improve. This can be intuitively explained as follows: a slower vanishing of $\gamma(t)$ slows down the IA procedure; this implies a deeper exploration of the space, that is, a larger number of detection regions (see Figure A.1) is visited and a larger number of paths is possible. Such a sweeping exploration increases the probability of achieving the best possible performance. On the other hand, if $\gamma(t)$ vanishes too rapidly, the estimates are driven in few steps to consensus and bound to move along a very few different directions.

Analogous considerations can be done for the mean square error on $\theta$: when $N$ increases, the mean square error decays to zero (see Figure 5.4), and for low connected graphs (random geometric and ring), IA performs better than IML and IM.

We remind the interested reader that a graphical user interface of our algorithm is available at http://calvino.polito.it/~fosson/software.html, which can be used to study the IA behavior in time, for a variety of different parameters’ settings.

6. Concluding remarks. In this paper, we have presented a fully distributed algorithm for the simultaneous estimation and classification in a sensor network, given from noisy measurements. The algorithm only requires the local cooperation between the network’s nodes, and numerical simulations show remarkable performance. The role of the cooperation has been our main focus, and our theoretical contribution in this sense is twofold: we have proved the convergence of the algorithm to a local
maximum of the centralized ML estimator and the concentration around the ML solution when the network size is sufficiently large.

The model considered in this paper envisages only one measurement per node, but adaptations to multiple measurements are straightforward, and the theoretical arguments used to prove convergence can be applied to this case. Regarding the performance, we expect that increasing the number of measurements would allow to lower the variance by averaging over themselves, up to the limit case of infinite measurements, where each node could exactly recover the common mean and its own standard deviation by itself. The study of a trade-off between the number of individual measurements and extent of cooperation, when communication costs enter into the picture, is a very interesting point that will be taken into account as future development.

Another relevant extension is the introduction of time-varying parameters: in such case, repeated measurements are necessary in time to track the objective [28]. From the theoretical viewpoint, the time-varying setting poses new challenges: convergence will no longer be an issue, and new metrics should be considered to measure the algorithm’s performance, based also on the characteristics of the time-varying functions. Given the relevance of such problems in real applications [2, 5], much attention will be devoted to this subject in our future work.
7. Acknowledgment. The authors wish to thank Sandro Zampieri for bringing the problem to our attention and Luca Schenato for useful discussions. F. Fagnani and C. Ravazzi further acknowledge the financial support provided by MIUR under the PRIN project no. 20087W5P2K.

REFERENCES


Appendix A. Proof of convergence. Consider the discrete-time dynamical system defined by the update equations (4.1) and (4.2): the proof of its convergence is obtained through intermediate steps.

1. First, we show that, for sufficiently large \( t \), vectors \( \mu^{(t)}, \mu^{(t)}, \) and \( \hat{\omega}^{(t)} \) are close to consensus vectors and we prove their convergence, assuming \( \hat{\omega}^{(t)} \) has already stabilized.

2. Second, we prove the stabilization of \( \hat{\omega}^{(t)} \) in finite time, by modelling the system in (4.1) and (4.2) as a switching dynamical system.

3. Finally, combining these facts together we conclude the proof.

A.1. Towards consensus. We start with some notation: let \( \Omega := I - N^{-1} \mathbb{1} \mathbb{T} ; \) for any fixed \( x \in \mathbb{R}^V \) and \( \bar{\pi} := N^{-1} \mathbb{1} \mathbb{T} x \) so that \( x = \bar{\pi} \mathbb{1} + \Omega x \).

Given a bounded sequence \( w(t) \in \mathbb{R}^N \), consider the dynamics

\[
\begin{align*}
x(t+1) & = \left( 1 - \gamma(t)^{\Omega} \right) P x(t) + \gamma(t)^{\Omega} u(t) \quad t \in \mathbb{N} \\
\end{align*}
\]  

(A.1)

where \( x^{(0)} \) is any fixed vector, and where, we recall the standing assumptions,

\begin{enumerate}[label=(A.\arabic*)]
\item \( \gamma(t) \in (0,1), \gamma(t) \geq 1/t, \gamma(t) \downarrow 0 \) and \( \gamma(t) = \gamma(t+1) + o(\gamma(t+1)) \) for \( t \to +\infty \);
\item \( P \in \mathbb{R}^{V \times V} \) is a stochastic, symmetric, primitive matrix with positive eigenvalues.
\end{enumerate}

A useful fact consequence of the assumptions on \( \gamma(t) \), is the following:

\[
1 - \frac{\sum_{s=t_0}^{t-1} 1/s}{t} \leq \frac{t_0}{t} \leq \frac{t_0}{t} \gamma(t)
\]  

(A.2)

for any choice of \( t \geq t_0 > 0 \).

On the other hand, as a consequence of the assumptions of \( P \) (see [29]) we have that \( P^t \to N^{-1} \mathbb{1} \mathbb{T} \), or equivalently that \( P^t \Omega \to 0 \) for \( t \to +\infty \). More precisely, we can order the eigenvalues of \( P \) as \( 1 = \mu_1 \geq \mu_2 \geq \cdots \geq \mu_N \geq 0 \), and we have that \( \|P^t \Omega\| \leq \mu_2^t \).

**Lemma A.1.** It holds

\[
\Omega x(t) = O(\gamma(t)), \quad \text{for } t \to +\infty.
\]

**Proof.** From (A.1) and the fact that \( \Omega P = P \Omega \) we get, for any fixed \( t_0 \) and \( t \geq t_0 \),

\[
\Omega x(t+1) = \prod_{s=t_0}^{t} \left( 1 - \gamma(s) \right) P^s \Omega x(t_0) + \sum_{s=t_0}^{t} \prod_{k=s+1}^{t} \left( 1 - \gamma(k) \right) \gamma(s) P^{t-s} \Omega u(s). \quad \text{(A.3)}
\]
This yields
\[
\| \Omega x^{(t+1)} \|_2 \leq \prod_{s=t_0}^t (1 - \gamma^{(s)}) \| \Omega x^{(t_0)} \|_2 + \sum_{s=t_0}^t \prod_{k=s+1}^t (1 - \gamma^{(k)}) \gamma^{(s)} \mu_2^{1-s} \| u^{(s)} \|_2 \\
\leq \prod_{s=t_0}^t (1 - \gamma^{(s)}) \| \Omega x^{(t_0)} \|_2 + K \sum_{s=t_0}^t \prod_{k=s+1}^t (1 - \gamma^{(k)}) \gamma^{(s)} \mu_2^{1-s} \quad \text{(A.4)}
\]

with \( K := \max_s \| u^{(s)} \|_2. \)

Fix now \( 0 < \mu_2 < 1 - \varepsilon \) (as a consequence of the assumptions of \( P, \mu_2 < 1 \)) and let \( t_0 \) be such that, for all \( t \geq t_0, \frac{\gamma^{(t)}}{1 - \gamma^{(t)}} > 1 - \varepsilon. \) Hence, there exists \( t_0 \in \mathbb{N} \) such that, for \( t \geq s \geq t_0, \) we have that \( \gamma^{(s)} < \frac{1\varepsilon}{(1-\varepsilon)} \) (the assumptions on sequence \( \gamma^{(t)} \)). Consider now the estimation (A.4) with this choice of \( t_0 \) and notice that \( \prod_{k=s+1}^t (1 - \gamma^{(k)}) < 1, \) being \( \gamma^{(t)} \in (0, 1) \) for all \( t \in \mathbb{N}. \) We get
\[
\| \Omega x^{(t+1)} \|_2 \leq \prod_{s=t_0}^t (1 - \gamma^{(s)}) \| \Omega x^{(t_0)} \|_2 + K \gamma^{(t)} \sum_{s=t_0}^t \frac{(\mu_2)(s)}{1 - \varepsilon} \| u^{(s)} \|_2 \\
\leq \prod_{s=t_0}^t (1 - \gamma^{(s)}) \| \Omega x^{(t_0)} \|_2 + \frac{K \gamma^{(t)}}{1 - \frac{\varepsilon}{1 - \varepsilon}}.
\]

Using now (A.2) the proof is completed. \( \square \)

We now apply these result to the analysis of \( \tilde{\mu}^{(t)}. \) We start with a representation result.

**Lemma A.2.** It holds, for \( t \to +\infty, \)
\[
\tilde{\mu}^{(t)} = \frac{\bar{\mu}^{(t)}(t)}{\nu^{(t)}} + \frac{1}{\nu^{(t)}} \Omega \left( \mu^{(t)} - \frac{\bar{\mu}^{(t)}}{\nu^{(t)}} \nu^{(t)} \right) + o \left( \gamma^{(t)} \right). \quad \text{(A.5)}
\]

**Proof.** For any \( i \in V, \)
\[
\frac{\mu_i^{(t)}}{\nu_i^{(t)}} - \frac{\tilde{\mu}^{(t)}}{\nu_i^{(t)}} = \frac{\mu_i^{(t)}}{\nu_i^{(t)}} - \frac{\bar{\mu}^{(t)}}{\nu_i^{(t)}} + \frac{\bar{\mu}^{(t)}}{\nu_i^{(t)}} - \frac{\tilde{\mu}^{(t)}}{\nu_i^{(t)}} \\
= \frac{\mu_i^{(t)}}{\nu_i^{(t)}} - \frac{\tilde{\mu}^{(t)}}{\nu_i^{(t)}} + \frac{1}{\nu_i^{(t)}} \left( \frac{\bar{\mu}^{(t)}}{\nu_i^{(t)}} - \frac{1}{\nu_i^{(t)}} \right) \\
= \frac{1}{\nu_i^{(t)}} \left( \Omega \mu_i^{(t)} \right) - \frac{\mu_i^{(t)}}{\nu_i^{(t)}} \left( \Omega \nu_i^{(t)} \right).
\]

It follows from Lemma A.1 that \( \mu^{(t)} = \tilde{\mu}^{(t)} + O(\gamma^{(t)}) \) and \( \nu^{(t)} = \nu^{(t)} + O(\gamma^{(t)}) \) for \( t \to +\infty. \) This and the fact that \( \bar{\nu}^{(t)} \) is bounded away from 0 (indeed \( \bar{\nu}^{(t)} \geq \alpha^{-2} \) for all \( t > 0 \)), yields
\[
\frac{\mu_i^{(t)}}{\nu_i^{(t)}} \left( \Omega \nu_i^{(t)} \right) = \frac{\mu_i^{(t)}}{\nu_i^{(t)}} \left[ \frac{\Omega \nu_i^{(t)}}{\nu_i^{(t)}} \right] \left( 1 + O(\gamma^{(t)}) \right)
\]
from which thesis follows. □

Corollary A.3. It holds, for $t \to +\infty$,

$$\theta(t) = \frac{\mu(t)}{\rho(t)} + o(\gamma(t)), \quad \Omega(0) = O(\gamma(t)).$$

Proof. Both relations are obtained from (A.5). The first one is immediate. The second one follows from Lemma A.1 and the fact that $\rho(t)$ stays bounded away from 0. □

Corollary A.3 says that the estimate $\theta(t)$ is close to a consensus for sufficiently large $t$. Something more precise can be stated if we suppose that if $\tilde{\omega}(i)$ stabilizes at finite time as explained in the next assertions. We can now present our first convergence result.

Proposition A.4. Let us consider the dynamics in (A.1). If $\exists t_0 \in \mathbb{N}$ s.t. $u(t) = u \forall t \geq t_0$ then

$$\lim_{t \to +\infty} x(t) = \overline{x}.$$  

Proof. Write $x(t) = \tilde{x}(t) + \Omega x(t)$ and notice that from Lemma A.1 it is sufficient to prove that $\lim_{t \to +\infty} \tilde{x}(t) = \overline{x}$. From (A.1) and the fact that $\mathbb{1}^T P = \mathbb{1}^T$, we obtain

$$\tilde{x}(t) = \mathbb{1} = \prod_{s=t_0}^{t-1} (1 - \gamma(s)) (\tilde{x}(s) - \overline{x})$$

which goes to zero from the non-summability of $\gamma(t)$. □

Corollary A.5. If $\exists t_0 \in \mathbb{N}$ s.t. $\tilde{\omega}(t) = \tilde{\omega}^{IA} \forall t \geq t_0$ then

$$\lim_{t \to +\infty} \tilde{\omega}(t) = \tilde{\omega}(IA) = \frac{\sum_{i \in V} y_i [\tilde{\omega}^{IA}]^{-2}}{\sum_{i \in V} [\tilde{\omega}^{IA}]^{-2}}.$$

Proof. Proposition A.4 guarantees that $\mu(t)$ and $\rho(t)$ converge to $\frac{1}{N} \sum_{i \in V} y_i [\tilde{\omega}^{IA}]^{-2} \mathbb{1}$ and $\frac{1}{N} \sum_{i \in V} [\tilde{\omega}^{IA}]^{-2} \mathbb{1}$, respectively. This yields the thesis. □

A.2. Stabilization of $\tilde{\omega}(t)$. We are going to prove that vector $\tilde{\omega}(t)$ almost surely stabilizes in finite time: this, by virtue of previous considerations will complete our proof. To prove this fact will take lots of effort and will be achieved through several intermediate steps.

We start observing that, since $\tilde{\omega}(t)$ can only assume values in a finite set, equations in (4.1) and (4.2) can be conveniently modeled by a switching system as shown below.

For reasons which will be clear below, in this subsection we will replace the configuration space $\{\alpha, \beta\}^V$ with the augmented state space $\{\alpha, \beta+, \beta-\}^V$. If $\omega \in \{\alpha, \beta+, \beta-\}^V$, define

$$\Theta_\omega = \{ x \in \mathbb{R}^V : |x_i - y_i| < \delta, if \omega_i = \alpha, x_i \geq y_i + \delta, if \omega_i = \beta+, x_i \leq y_i - \delta, if \omega_i = \beta- \}.$$

We clearly have $\mathbb{R}^V = \bigcup_{\omega \in \{\alpha, \beta+, \beta-\}^V} \Theta_\omega$ (see the case with $N = 2$ in Fig. A.1).
On each \( \Theta_\omega \), the dynamical system is linear. Indeed, define the maps \( f_\omega : \mathbb{R} \times \mathbb{R}^V \to \mathbb{R}^V \) and \( g_\omega : \mathbb{R} \times \mathbb{R}^V \to \mathbb{R}^V \) by

\[
[f_\omega(t,x)]_i = (1 - \gamma(t))[P_x(t)]_i + \gamma(t) \frac{y_i}{\omega_i^2}
\]

\[
[g_\omega(t,x)]_i = (1 - \gamma(t))[P_x(t)]_i + \gamma(t) \frac{1}{\omega_i^2}
\]

where, conventionally, \( \omega_i^2 = \beta^2 \) if \( \omega_i = \beta^+, \beta^- \). Then, if \( \hat{\theta}(t) \in \Theta_\omega \), (4.1a), (4.1b), and (4.1c) can be written as

\[
\mu^{(t+1)} = f_\omega(t, \mu^{(t)}) \quad \nu^{(t+1)} = g_\omega(t, \nu^{(t)}) \quad \hat{\theta}_i^{(t+1)} = \mu_i^{(t+1)} / \nu_i^{(t+1)}.
\]

Notice that this is a closed-loop switching system, since the switching policy is determined by \( \hat{\theta}(t) \). It is clear that the stabilization of \( \hat{\omega}(t) \) is equivalent to the fact that there exist an \( \omega \in \{\alpha, \beta, +, -\}^N \) and a time \( \bar{t} \) such that \( \hat{\theta}(t) \in \Theta_\omega \) for all \( t \geq \bar{t} \).

From Corollary A.5 candidate limit points for \( \hat{\theta}(t) \) are

\[
\hat{\theta}(\omega) = \frac{\sum_{i \in V} y_i \omega_i^{-2}}{\sum_{i \in V} \omega_i^{-2}} \quad \omega \in \{\alpha, \beta\}^N.
\]

Also, from Corollary A.3, the dynamics can be conveniently analyzed by studying it in a neighborhood of the line \( \Lambda = \{\lambda \| \lambda \in \mathbb{R}\} \) (see the example in Fig. A.2).

We now make an assumption which holds almost everywhere with respect to the choice of \( y_i \)'s and, consequently, does not entail any loss of generality in our proof.

**Assumption:**

- \( y_i - y_j \notin \{0, \pm \delta, \pm 2\delta\} \) for all \( i \neq j \);
- \( \hat{\theta}(\omega) - y_i \notin \{\pm \delta\} \) for all \( \omega \in \{\alpha, \beta, +, -\}^V \) and for all \( i \).

This assumption has a number of consequences which will be used later on:
(C1) $\hat{\theta}(\omega)_i \in \bigcup_{\omega \in \{\alpha, \beta +, \beta -\}} \text{int}(\Theta_{\omega})$ for all $\omega \in \{\alpha, \beta +, \beta -\}$ and for all $i \in V$.

(C2) $\Lambda \cap \bar{\Theta}_{\omega} \cap \bar{\Theta}_{\omega'} \cap \bar{\Theta}_{\omega''} = \emptyset$ for any triple of distinguished $\omega, \omega', \omega''$. In other terms, $\Lambda$ always crosses boundaries among regions $\Theta_{\omega}$ at internal point of faces.

We now introduce some further notation, which will be useful in the rest of the paper.

$\Theta^\epsilon := \{x \in \mathbb{R}^V : ||\Omega x||_2 < \epsilon\}, \quad \Theta^\epsilon_{\omega} := \Theta^\epsilon \cap \Theta_{\omega}$

$\Gamma := \{\omega \in \{\alpha, \beta +, \beta -\}^V : \Theta_{\omega} \cap \Lambda \neq \emptyset\}$.

Consider the case with $N = 2$ in Figure A.2: we have that

$\Gamma = \{ (\beta -, \beta -), (\alpha, \beta -), (\beta +, \beta -), (\beta +, \alpha), (\beta +, \beta +) \}$

For any $\omega \in \Gamma$ consider

$\Pi_{\omega} = \{ \pi = \bar{\Theta}_{\omega} \cap \Theta_{\omega} : d_H(\omega, \omega') = 1, \pi \cap \Lambda = \emptyset \}$

and define $\sigma_{\omega} := \min_{\pi \in \Pi_{\omega}} d(\Theta_{\omega} \cap \Lambda, \pi) > 0$.

In the sequel, we will use the natural ordering on $\Lambda$: given the sets $X, Y \subseteq \Lambda$, $X < Y$ means that $x < y$ for all $x \in X$ and $y \in Y$.

**Definition A.6.** Given two elements $\omega, \omega' \in \Gamma$, we say that $\omega'$ is the future-follower of $\omega$ (or also that $\omega$ is the past-follower of $\omega'$) if the following happens:

(A) There exists $i_0$ such that $\omega_i = \omega'_i$ for all $i \neq i_0$ and $\omega_{i_0} \neq \omega'_{i_0}$ (this means that $d_H(\omega, \omega') = 1$);

(B) $\Theta_{\omega} \cap \Lambda < \Theta_{\omega'} \cap \Lambda$.

Notice that, in order for $\omega$ and $\omega'$ to satisfy definition above, it must necessarily happen that either $\omega_{i_0} = \alpha$ and $\omega'_{i_0} = \beta +$, or $\omega_{i_0} = \beta -$ and $\omega'_{i_0} = \alpha$. Given $\omega \in \Gamma$, its future-follower (if it exists) will be denoted by $\omega^+$. It is clear that (because of

---

1 $d(\Theta_{\omega} \cap \Lambda, \pi)$ denotes the distance between the two sets $\Theta_{\omega} \cap \Lambda$ and the set $\pi$
property (C2) described above) that we can order elements in \( \Gamma \) as \( \omega^1, \omega^2, \ldots, \omega^M \) in such a way that \( \omega^{r+1} = (\omega^r)^+ \) for every \( r = 1, \ldots, M - 1 \).

Let us consider the example with \( N = 2 \) in Figure A.2: we can order all elements in \( \Gamma \) as follows
\[
(\beta^-, \beta^-)^+ = (\alpha, \beta^-), (\alpha, \beta^-)^+ = (\beta^+, \beta^-), (\beta^+, \beta^-)^+ = (\beta^+, \beta^+).
\]

Let us consider the example with \( N = 2 \) in Figure A.2: we can order all elements in \( \Gamma \) as follows
\[
(\beta^-, \beta^-)^+ = (\alpha, \beta^-), (\alpha, \beta^-)^+ = (\beta^+, \beta^-), (\beta^+, \beta^-)^+ = (\beta^+, \beta^+).
\]

Given \( \omega \in \Gamma \), consider the following subsets of \( \mathbb{R}^N \) (see Fig. A.3):
\[
\mathcal{M}_{\omega}^\epsilon := \{ x \in \Theta_{\epsilon} : \nabla \omega + \Omega z \in \Theta_{\epsilon} \}, \quad \forall \omega \in \Theta \epsilon, \forall \omega \in \Theta \epsilon.
\]
\[
L_{\omega, \omega^+}^\epsilon := \{ x \in \Theta^\epsilon : \mathcal{M}_{\omega}^\epsilon \cap \Lambda < \hat{x} < \mathcal{M}_{\omega^+}^\epsilon \cap \Lambda \}.
\]

(with the implicit assumption that \( L_{\omega, \omega^+}^\epsilon = \emptyset \) if \( \omega^+ \) does not exist.) We clearly have
\( \Theta^\epsilon = \bigcup_{\omega \in \Gamma} \mathcal{M}_{\omega}^\epsilon \cup L_{\omega, \omega^+}^\epsilon \).

\[\text{Figure A.3. Given the couple } (\omega, \omega') \text{ the sets } L_{\omega, \omega'}^\epsilon \text{ and } \mathcal{M}_{\omega}^\epsilon \text{ are visualized.}\]

Notice that, because of property (C1), we can always choose \( \epsilon_0 \in (0, \min_{\omega \in \Gamma} \sigma_{\omega}) \) such that
\[
\hat{\theta}(\omega) = y_i \in \bigcup_{\omega \in \Gamma} \mathcal{M}_{\omega}^\epsilon, \quad \forall \omega \in \Theta^\epsilon, \forall i \in V.
\]

This implies that there exists \( \tilde{\epsilon} > 0 \) such that
\[
d \left( \bigcup_{\omega \in \Gamma} \partial_{\Lambda} (\mathcal{M}_{\omega}^\epsilon \cap \Lambda), \{ \hat{\theta}(\omega), y_i \} \right) \geq \tilde{\epsilon}, \quad \forall \epsilon \leq \epsilon_0 \quad (A.6)
\]

where \( \partial_{\Lambda}(\cdot) \) denotes the boundary of a set in the relative topology of \( \Lambda \).

Fix now \( \epsilon \leq \epsilon_0 \) and choose \( t_\epsilon \) such that \( \hat{\theta}(t) \in \Theta^\epsilon \) for all \( t \geq t_\epsilon \) (it exists by Corollary A.3). From now on we consider times \( t \geq t_\epsilon \). Our aim is to prove through intermediate steps the following facts

F1) if \( \hat{\theta}(\omega) \in \mathcal{M}_{\omega}^\epsilon \) then \( \mathcal{M}_{\omega}^\epsilon \) is an asymptotically invariant set for \( \hat{\theta}(t) \), namely, when \( t \) is sufficiently large, if \( \hat{\theta}(t) \in \mathcal{M}_{\omega}^\epsilon \) then \( \hat{\theta}(t+1) \in \mathcal{M}_{\omega}^\epsilon \);

F2) if \( \hat{\theta}(\omega) \notin \mathcal{M}_{\omega}^\epsilon \) then \( \hat{\theta}(t) \notin \mathcal{M}_{\omega}^\epsilon \) for \( t \) sufficiently large;

F3) \( \hat{\theta}(t) \notin \bigcup_{\omega \in \{ \alpha, \beta^+, \beta^- \}} L_{\omega, \omega^+}^\epsilon \) for \( t \) sufficiently large.
**F1) Asymptotic invariance of** $\mathcal{M}_\omega'$ **when** $\tilde{\theta}(\omega)\| \in \mathcal{M}_\omega'$.

**Lemma A.7.** If $\tilde{\theta}(t) \in \Theta_\omega$ then there exists $c(t) \in [\alpha^2/\beta^2, \beta^2/\alpha^2]$ and $r(t) = o(\gamma(t))$ for $t \to +\infty$ such that

$$
\tilde{\theta}^{(t+1)} = \tilde{\theta}^{(t)} + c(t)\gamma(t) \left( \tilde{\theta}(\omega) - \tilde{\theta}^{(t)} \right) + r(t) \quad (A.7)
$$

**Proof.** If $\tilde{\theta}(t) \in \Theta_\omega$ then

$$
\frac{\mu^{(t+1)}}{\mu^{(t)}} = \frac{\mu^{(t)}}{\mu^{(t)}} \left( 1 - \gamma(t) \right) + \gamma(t) N^{-1} \sum_{i=1}^{N} y_i \omega_i^{-2} - \frac{\mu^{(t)}}{\mu^{(t)}} = \frac{\mu^{(t)} - \gamma(t) N^{-1} \sum_{i=1}^{N} y_i \omega_i^{-2}}{\mu^{(t)}}
$$

Choosing $c(t) = \frac{N^{-1} \sum_{i=1}^{N} y_i \omega_i^{-2}}{\mu^{(t+1)}} \in [\alpha^2/\beta^2, \beta^2/\alpha^2]$ and using Corollary A.3 thesis easily follows. □

**Proposition A.8 (Proof of F1)).** There exists $t' \geq t_\epsilon$ such that, if $\tilde{\theta}(\omega)\| \in \Theta_\omega$, then

$$
\tilde{\theta}(t) \in \mathcal{M}_\omega' \Rightarrow \tilde{\theta}(t+1) \in \mathcal{M}_\omega' \quad \forall t \geq t'.
$$

**Proof.** Consider the relation (A.7). If $\tilde{\theta}(t) \in \mathcal{M}_\omega'$ and if $t$ is large enough so that $c(t)\gamma(t) < 1$, we have, by convexity, that

$$
z := \tilde{\theta}^{(t)} + c(t)\gamma(t) \left( \tilde{\theta}(\omega) - \tilde{\theta}^{(t)} \right) \in \mathcal{M}_\omega'.
$$

Moreover, because of (A.6) and the fact that $c(t)$ is bounded away from 0, there exists $c' > 0$ such that $d(z, \partial(\mathcal{M}_\omega' \cap \Lambda)) \geq c' \gamma(t)$. Proof is then completed by selecting $t' \geq t_\epsilon$ such that $c(t') < 1$ and $|r(t)| < c'\gamma(t)/2$ for all $t \geq t'$. □

**F2) Transitivity of** $\mathcal{M}_\omega'$ **when** $\tilde{\theta}(\omega)\| \notin \mathcal{M}_\omega'$. Our next goal is to prove that if $\tilde{\theta}(\omega)\| \notin \mathcal{M}_\omega'$, then, at a certain time $t$, $\tilde{\theta}(t)$ will definitively be outside $\mathcal{M}_\omega'$. A technical lemma based on convexity arguments is required.

**Lemma A.9.** Let $\omega \in \Gamma$ be such that there exists its future-follower $\omega^+$. Then,

$$
\tilde{\theta}(\omega)\| > \Theta_\omega \cap \Lambda \quad \Rightarrow \quad \tilde{\theta}(\omega^+)\| > \Theta_\omega \cap \Lambda
$$

$$
\tilde{\theta}(\omega^+)\| < \Theta_\omega \cap \Lambda \quad \Rightarrow \quad \tilde{\theta}(\omega)\| < \Theta_\omega \cap \Lambda.
$$

**Proof.** Suppose $\omega_i = \omega_i^+, \forall i \neq i_0$ and $\omega_{i_0} = \beta_-$, $\omega_{i_0}^+ = \alpha$ (the other case can be treated in an analogous way). Pick $x' \in \Theta_\omega \cap \Lambda$ and $x'' \in \Theta_\omega^+ \cap \Lambda$. From $|x'' - y_{i_0}| < \delta$ and $|x' - y_{i_0}| > \delta$ it immediately follows that $x'' > y_{i_0} - \delta$, $x' < y_{i_0} + \delta$ and, in particular, the fact

$$
y_{i_0} > \Theta_\omega \cap \Lambda.
$$

(A.8)
Notice now that

\[
\tilde{\theta}(\omega^+) = \frac{y_{i_0} \left( \frac{1}{\alpha^2} - \frac{1}{\beta^2} \right)}{\sum_{i \in V} \frac{1}{\omega_i^2}} + \frac{\sum_{i \in V \setminus i_0} \frac{1}{\omega_i^2} + \frac{y_{i_0}}{\omega_i^2}}{\sum_{i \in V} \frac{1}{\omega_i^2}}
\]

\[
= y_{i_0} \left( \frac{1}{\alpha^2} - \frac{1}{\beta^2} \right) + \frac{\hat{\theta}(\omega) \sum_{i \in V} \frac{1}{\omega_i^2}}{\sum_{i \in V} \frac{1}{\omega_i^2}}
\]

\[
= y_{i_0} \left( \frac{1}{\alpha^2} - \frac{1}{\beta^2} \right) + \frac{\hat{\theta}(\omega) \left[ \sum_{i \in V} \frac{1}{\omega_i^2} - \left( \frac{1}{\alpha^2} - \frac{1}{\beta^2} \right) \right]}{\sum_{i \in V} \frac{1}{\omega_i^2}}
\]

In Figures A.4 and A.5 a picture of the various points is depicted when \( \hat{\theta}(\omega) > \Theta_\omega \cap \Lambda \).

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figureA4.png}
\caption{\( \omega_{i_0} = \beta, \hat{\theta}(\omega)1 > \Theta_\omega \cap \Lambda \) (a) \( y_{i_0} < \hat{\theta}(\omega^+) < \hat{\theta}(\omega) \).}
\end{figure}

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figureA5.png}
\caption{\( \omega_{i_0} = \alpha, \hat{\theta}(\omega)1 > \Theta_\omega \cap \Lambda \) (b) \( \hat{\theta}(\omega) < \hat{\theta}(\omega^+) < y_{i_0} \).}
\end{figure}

A convexity argument and the use of (A.8) now allow to conclude. \( \Box \)

**Proposition A.10 (Proof of F2).** If \( \hat{\theta}(\omega)1 \notin \Theta_\omega \), then there exists \( t'' \) such that \( \tilde{\theta}(t') \notin \Theta_\omega \forall t > t'' \).

**Proof.** Suppose \( \hat{\theta}(\omega)1 \notin \Theta_\omega \cap \Lambda \) (the case when \( \leq \) can be treated analogously).

Lemma A.9 implies that \( \hat{\theta}(\omega^+)1 \notin \Theta_\omega \cap \Lambda \). Let \( \tilde{c} \) be the constant given in (A.6) and put

\[ A := \{ x \in \Theta_\omega^+ \cup \Theta_\omega^+ \mid \overline{x} \leq \alpha := \min \{ \hat{\theta}(\omega), \hat{\theta}(\omega^+) \} - \tilde{c}/2 \}. \]

Consider the relation (A.7) and choose \( t_1 \) in such a way that

\[ \frac{z^{(t+1)}}{\tilde{\theta}} - \frac{z^{(t)}}{\tilde{\theta}} \leq c_2(\max \{ y_i \} - \min \{ y_i \})^{(t)} + r(t) < \tilde{c}/2 \quad (A.9) \]

and \( |r(t)| < \alpha^2 \gamma(1)/4 \beta^2 \) for all \( t \geq t_1 \). It also follows from (A.7) that, if for some \( t \geq t_1 \), then,

\[ \frac{z^{(t+1)}}{\tilde{\theta}} \geq \frac{z^{(t)}}{\tilde{\theta}} + \alpha^2 \tilde{c} \gamma(t)/4 \beta^2. \quad (A.10) \]
Owing to the non-summability of $\gamma(t)$ it follows that if $\hat{\theta}(t)$ enters in $\Theta^c_\omega$ for some $t \geq t_1$, then, in finite time it will enter into $A \setminus \Theta^c_\omega$ and then it will finally exit $A$. In particular there must exist $t_2 \geq t_1$ such that $\hat{\theta}(t_2) > \alpha$. We now prove that $\hat{\theta}(t) > \Theta_\omega$ for every $t \geq t_2$. If not there must exist a first time index $t_3 > t_2$ such that $\hat{\theta}(t_3) < \alpha - \tilde{c}$. Because of (A.9), it must be that $\hat{\theta}(t_3-1) < \alpha - \tilde{c}/2$ but this contradicts the fact that on $A$, $\hat{\theta}(t)$ is increasing (A.10).

**F3) Transitivity of $\bigcup_{\omega, \omega+}^{\alpha, \beta+} \in \{\alpha, \beta+\}_{N \in \mathbb{N}} L_{\omega, \omega+}$.** We start with the following technical result concerning the general system (A.1).

**Lemma A.11.** Let $x^{(t)}$ be the sequence defined in (A.1) and suppose that there exists a strictly increasing sequence of switching times $\{\tau_k\}_{k=0}^{+\infty}$ such that

$$u(s+1) = u(s) \quad \forall i \neq i_0 \quad \text{and} \quad \forall s \in [\tau_0, +\infty[,$$

$$u_{\tau_0}^{(s)} = \begin{cases} v' & \forall s \in I' := \bigcup_{k=0}^{+\infty} [\tau_{2k}, \tau_{2k+1}) \\ v'' & \forall s \in I'' := \bigcup_{k=0}^{+\infty} [\tau_{2k+1}, \tau_{2k+2}) \end{cases}.$$

Then, for every $\delta > 0$, there exists $\tilde{t}_\delta$ and two sequences $a^{(t)}_\delta \geq 0$ and $b^{(t)}_\delta \leq \delta \gamma(t)$, such that

$$\left( \Omega \left( x^{(t+1)} - x^{(t)} \right) \right)_{\tau_0} = a^{(t)}_\delta \gamma(t) (v' - v'') + b^{(t)}_\delta$$

for $t \in I'$ with $t \geq \tilde{t}_\delta$.

**Proof.** Let $\phi_i \in \mathbb{R}^N$ be an orthonormal basis of eigenvectors for $P$ relative to the eigenvalues $1 = \lambda_1 > \lambda_2 \geq \cdots \geq \lambda_N \geq 0$. Also assume we have chosen $\phi_1 = N^{-1/2}$. We put

$$F^{(t)} := \prod_{k=0}^{t} \left( 1 - \gamma^{(k)} \right) / \gamma^{(t)}$$

and we notice that

$$\frac{F^{(s+1)}}{F^{(s)}} = \left( 1 - \gamma^{(s+1)} / \gamma^{(s)} \right) \rightarrow 1, \text{ for } s \rightarrow +\infty.$$  

Fix $\epsilon$ in such a way that $\lambda_2(1 + \epsilon) < 1$ and choose $s_0$ such that

$$\frac{F^{(s+1)}}{F^{(s)}} \leq 1 + \epsilon, \quad \forall s \geq s_0.$$  

Let $t_0 \geq s_0$ to be fixed later. From (A.3) we can write
\( \Omega(x^{(t+1)} - x^{(t)}) = \)
\[= \prod_{s=t_0}^{t-1} (1 - \gamma^{(s)}) \left[ (1 - \gamma^{(t)}) P - I \right] p^{t-t_0} \Omega x^{(t_0)} \]
\[+ \sum_{s=t_0}^{t} \prod_{k=s+1}^{t} (1 - \gamma^{(k)}) \gamma^{(s)} p^{t-s} \Omega u^{(s)} - \sum_{s=t_0}^{t-1} \prod_{k=s+1}^{t-1} (1 - \gamma^{(k)}) \gamma^{(s)} p^{t-s-1} \Omega u^{(s)} v \]
\[= \prod_{s=t_0}^{t-1} (1 - \gamma^{(s)}) \left[ (1 - \gamma^{(t)}) P - I \right] p^{t-t_0} \Omega x^{(t_0)} \]
\[+ \gamma^{(t)} \sum_{s=t_0}^{t-1} p^{t-s-1} \frac{F^{(t)}}{F(s+1)} \Omega u^{(s+1)} - \gamma^{(t-1)} \sum_{s=t_0}^{t-1} p^{t-s-1} \frac{F^{(t-1)}}{F(s)} \Omega u^{(s)} \]
\[= \prod_{s=t_0}^{t-1} (1 - \gamma^{(s)}) \left[ (1 - \gamma^{(t)}) P - I \right] p^{t-t_0} \Omega x^{(t_0)} \]  \quad \text{(A.11)}
\[+ (\gamma^{(t)} - \gamma^{(t-1)}) \sum_{s=t_0}^{t-1} p^{t-s-1} \frac{F^{(t-1)}}{F(s)} \Omega u^{(s)} + \gamma^{(t)} p^{t-t_0} \frac{F^{(t-1)}}{F(t_0)} \Omega u^{(t_0)} \]  \quad \text{(A.12)}
\[+ \gamma^{(t)} \sum_{s=t_0}^{t-1} p^{t-s-1} \left( \frac{F^{(t)}}{F(s+1)} - \frac{F^{(t-1)}}{F(s)} \right) \Omega u^{(s+1)} \]  \quad \text{(A.13)}
\[+ \gamma^{(t)} \sum_{s=t_0}^{t-1} p^{t-s-1} \frac{F^{(t-1)}}{F(s)} \Omega \left( u^{(s+1)} - u^{(s)} \right) . \]  \quad \text{(A.14)}

It follows from the assumptions on \( P \), the assumptions on \( \gamma^{(t)} \) and relation (A.2) that the terms (A.11) and (A.12) are both \( o(\gamma^{(t)}) \) for \( t \to +\infty \). We now estimate (A.13):

\[ \left\| \sum_{s=t_0}^{t-1} p^{t-s-1} \left( \frac{F^{(t)}}{F(s+1)} - \frac{F^{(t-1)}}{F(s)} \right) \Omega u^{(s+1)} \right\| _2 = \]
\[ \leq \left\| \sum_{s=t_0}^{t-1} \frac{F^{(t-1)}}{F(s)} \left( \frac{F^{(t)}}{F(t+1)} - \frac{F^{(s)}}{F(s+1)} \right) - 1 \right\| _2 p^{t-s-1} \Omega u^{(s+1)} \]
\[ \leq \sum_{s=t_0}^{t-1} \left| \lambda_2 (1 + \epsilon) \right|^{t-s-1} \left| \left( \frac{F^{(s)}}{F(s+1)} - 1 \right) \right| K \leq \]
\[ \frac{K}{1 - \lambda_2 (1 + \epsilon) \beta_{t_0}} \]  \quad \text{(A.15)}

where

\[ K = \max \| u^{(s)} \| _2 , \quad \beta_{t_0} := \sup_{t \geq s \geq t_0} \left| \frac{F^{(t)}}{F(t+1)} - \frac{F^{(s)}}{F(s+1)} \right| . \]

We now concentrate on the component \( t_0 \) of the term (A.14). Using the spectral
decomposition of $P$ and the assumptions on $u(t)$, we can write,

$$
\left[ \sum_{s=t_0}^{t-1} P^{t-s-1} \frac{F(t-1)}{F(s)} \Omega \left( u(s+1) - \nu(s) \right) \right]_{t_0} = \tag{A.16}
$$

$$
\sum_{j \geq 2} (\phi_j)_{t_0}^2 \sum_{h:t_0 \leq \tau_h \leq t-1} \lambda_j^{t-\tau_h} \frac{F(t-1)_{\tau_h - 1}}{F(\tau_h - 1)} (t') - v''). \tag{A.17}
$$

If $t \in T'$, the above expression can be rewritten as

$$
\sum_{j \geq 2} (\phi_j)_{t_0}^2 \sum_{k:t_0 \leq \tau_k \leq t-1} \left[ \lambda_j^{t-\tau_k} \frac{F(t-1)}{F(\tau_k - 1)} - \lambda_j^{t-\tau_k - 1} \frac{F(t-1)}{F(\tau_k - 1 - 1)} \right] (v' - v'').
$$

Notice that

$$
\lambda_j^{t-\tau_k} \frac{F(t-1)}{F(\tau_k - 1)} - \lambda_j^{t-\tau_k - 1} \frac{F(t-1)}{F(\tau_k - 1 - 1)} = \lambda_j^{t-\tau_k} \frac{F(t-1)}{F(\tau_k - 1)} \left( 1 - \lambda_j^{\tau_k - \tau_k - 1} \frac{F(\tau_k - 1)}{F(\tau_k - 1 - 1)} \right) > 0
$$

(we have used the fact that $0 \leq \lambda_j(1 + \epsilon) < 1$ for all $j \geq 2$). To complete the proof now proceed as follows. For a fixed $\delta > 0$, choose $t_0 \geq s_0$ in such a way that (A.15) is below $\delta/2$. Then, fix $t_0 \geq t_0$ in such a way that the summation of (A.11) and (A.12) is below $\delta \gamma(t)/2$ for $t \geq t_0$. It is now sufficient to define

$$a_{t_0}^{(t)} := \sum_{j \geq 2} (\phi_j)_{t_0}^2 \sum_{k:t_0 \leq \tau_k \leq t-1} \left[ \lambda_j^{t-\tau_k - 1} \frac{F(t-1)}{F(\tau_k - 1)} - \lambda_j^{t-\tau_k - 1} \frac{F(t-1)}{F(\tau_k - 1 - 1)} \right]
$$

and $b_{t_0}^{(t)}$ equal to the sum of the terms (A.11), (A.12), and (A.13). $\square$

**Proposition A.12 (Proof of F3).** There exists $t'' \in \mathbb{N}$ such that

$$\tilde{\theta}^{(t)} \notin \bigcup_{(\omega, \omega') \notin \{a, b\}} \mathcal{L}_{\omega, \omega'} \tag{A.18}
$$

for all $t > t''$.

**Proof.** In view of the results in Propositions A.8 and A.10, and the fact that $\tilde{\theta}^{(t+1)} - \tilde{\theta}^{(t)}$ goes to 0 for $t \to +\infty$, negation of (A.18) yields that there exists $\omega \in \Gamma$ such that $\tilde{\theta}^{(t)} \in \mathcal{L}_{\omega, \omega^*}$ for $t$ large enough. Now, if $\tilde{\theta}^{(t)} \in \mathcal{L}_{\omega, \omega^*} \cap \Theta_\omega$ (or if $\tilde{\theta}^{(t)} \in \mathcal{L}_{\omega, \omega^*} \cap \Theta_{\omega^+}$) for $t$ sufficiently large, a straightforward application of (A.7) would imply that $\tilde{\theta}^{(t)}$ would necessarily exit $\mathcal{L}_{\omega, \omega^*}$ in finite time. Therefore, it must hold that $\tilde{\theta}^{(t)}$ keeps switching, for large $t$, between $\mathcal{L}_{\omega, \omega^*} \cap \Theta_\omega$ and $\mathcal{L}_{\omega, \omega^*} \cap \Theta_{\omega^*}$.

From Lemma A.2 and Corollary A.3 we can write

$$\tilde{\theta}^{(t+1)} - \tilde{\theta}^{(t)} = \left( \frac{\nu^{(t+1)}}{\nu^{(t)} - \nu^{(t)}} \right) I + \frac{1}{\nu^{(t)}} \left[ \Omega \left( \mu^{(t+1)} - \mu^{(t)} \right) - \frac{\bar{\mu}^{(t)}}{\nu^{(t)}} \Omega \left( \nu^{(t+1)} - \nu^{(t)} \right) \right] + o(\gamma(t)).
$$

Define now

$$t' := \{ t \mid \tilde{\theta}^{(t)} \in \Theta_\omega \}, \quad t'' := \{ t \mid \tilde{\theta}^{(t)} \in \Theta_{\omega^*} \}$$
and put $v' = 1/\omega_{i_0}^2$ and $v'' = 1/\omega_{i_0}^+$. From Lemma A.7, and applying Lemma A.11 to $\mu^{(t)}$ and $\nu^{(t)}$, we get that for $t \in I'$ sufficiently large, it holds

$$\hat{\theta}^{(t+1)}_{i_0} - \hat{\theta}^{(t)}_{i_0} = c^{(t)}\gamma^{(t)}(\hat{y}_{\omega} - \hat{\theta}^{(t)}) + \frac{1}{\rho^{(t)}}\gamma^{(t)}(v' - v'') (y_{i_0} - \hat{\theta}^{(t)}) + a^{(t)}_{k} + r^{(t)}. \quad (A.19)$$

If $\hat{\theta}(\omega) > \Theta_v \cap A$, then also, by Lemma A.9, $\tilde{y}_{\omega} > \Theta_v \cap A$. This, using (A.7), would imply that $\hat{\theta}(t)$ would necessarily exit $L_{\omega,\omega^+}$ in finite time. Therefore, we must have $\hat{\theta}(\omega) < \Theta_v \cap A$. Hence, $\tilde{y}_{\omega} - \hat{\theta}(t) < 0$. Moreover, it is easy to check that in any case $(v' - v'') (y_{i_0} - \hat{\theta}^{(t)}) < 0$. Recall now the definition of the constant $\tilde{c}$ in (A.6) and notice that, since $\hat{\theta}(t) \in L_{\omega,\omega^+}$,

$$c^{(t)}\gamma^{(t)}(\hat{y}_{\omega} - \hat{\theta}^{(t)}) \leq -\alpha^2 \tilde{c}/4 \beta^2 \gamma^{(t)}.$$

Choose now $\delta$ such that $\delta < \alpha^2 \tilde{c}/16 \beta^2$ and $t \geq t_\delta$ such that $r(t) < \delta \gamma^{(t)}$. It then follows from (A.19) that for $t \in I'$ and $t \geq t_\delta$, it holds

$$\hat{\theta}^{(t+1)}_{i_0} - \hat{\theta}^{(t)}_{i_0} \leq -\alpha^2 \tilde{c}/8 \beta^2 \gamma^{(t)} < 0. \quad (A.20)$$

This says that as long as $\hat{\theta}(t) \in \Theta_v$, its $i_0$-th component decreases. But this entails that $\hat{\theta}(t)$ can never leave $\Theta_v$, which contradicts the infinite switching assumption and thus implies the thesis. \[\Box\]

**A.3. Proof of Theorem 4.1.** Propositions A.8, A.10, and A.12 imply that there exists $\hat{\Theta}_{\omega} \in \{\alpha, \beta\}^\mathcal{V}$ such that $\hat{\theta}(t) \in \hat{\Theta}_{\omega} \cap A$ for $t$ sufficiently large. This immediately implies that $\hat{\omega}^{(t)} = \hat{\Theta}_{\omega} \cap A$ for $t$ sufficiently large. Corollary A.5 implies that $\hat{\omega}^{(t)} = \lim_{t \to +\infty} \hat{\theta}(t) = \hat{\Theta}_{\omega} \cap A$. Finally, since $\hat{\theta}(\hat{\omega}) \in \hat{\Theta}_{\omega} \cap A$, we also have that $\hat{\omega}^{(t)} = \hat{\omega}(\hat{\omega})$.

**Appendix B. Proof of concentration results.**

**B.1. Preliminaries.** For a more efficient parametrization of the stationary points, we introduce the notation:

$$\omega \in \{\alpha, \beta\}^\mathcal{V} \quad \Theta_v := \{x \in \mathbb{R} | |x - y_i| < \delta \iff \omega_i = \alpha \text{ for any } i \in \mathcal{V}\} \quad (B.1)$$

It is then straightforward to check from (3.11) that the set of local maxima $S_N$ can be represented as

$$S_N := \{\theta = \hat{\theta}(\omega) | \omega \in \{\alpha, \beta\}^\mathcal{V}, \hat{\theta}(\omega) \in \Theta_v\}. \quad (B.2)$$

Since,

$$\Theta_v \neq \emptyset \iff \omega = \hat{\omega}(x) \text{ for some } x \in \mathbb{R} \quad (B.3)$$

for analyzing the set $S_N$ we can restrict to consider $\omega$ of type $\omega = \hat{\omega}(x)$. Consider the sequence of random functions $\gamma_N(x) := \hat{\theta}(\hat{\omega}(x))$.

From (3.8), applying the strong law of large numbers, we immediately get that

$$\lim_{N \to +\infty} \gamma_N(x) \overset{a.s.}{=} \gamma_{\infty}(x) := \frac{E(y_i \hat{\omega}(x))^{-2}}{E(\hat{\omega}(x))^{-2}}. \quad (B.4)$$
Something stronger can indeed be said by a standard use of Chernoff bound [30]:

**Lemma B.1.** For every \( \epsilon > 0 \), there exists \( q < 1 \) such that, for any \( x \in \mathbb{R} \),

\[
\mathbb{P}(|\gamma_N(x) - \gamma_\infty(x)| > \epsilon) \leq 2q^N.
\]

**Proof.** Let \( a_i = y_i\omega_N(x)_{i}^{-2} \) and \( b_i = \omega_N(x)_{i}^{-2} \) with \( i \in \{1, \ldots, N\} \) and let \( a \) and \( b \) denote the corresponding expected values.

By Chernoff’s bound and by Hoeffding’s inequality we have, respectively, that

\[
\mathbb{P} \left( \left| \frac{1}{N} \sum_{i=1}^{N} a_i - a \right| \geq \epsilon_1 \right) \leq q_1^N \quad \mathbb{P} \left( \left| \frac{1}{N} \sum_{i=1}^{N} b_i - b \right| \geq \epsilon_2 \right) \leq 2q_2^N
\]

with

\[
q_1 = e^{-\alpha^2 \epsilon_1^2} \quad q_2 = e^{-2\epsilon_2^2(\alpha^2 - \beta^2)^{-2}}.
\]  (B.5)

Fix \( \epsilon_1 < \frac{\alpha^2}{2\beta^2} \) and \( \epsilon_2 < \frac{1}{2|\alpha|} \), then

\[
\mathbb{P} \left( |\tilde{y}_N(x) - y_\infty(x)| > \epsilon \right) \leq \mathbb{P} \left( \left| \frac{1}{N} \sum_{i=1}^{N} a_i - a \right| b + |a| \left| b - \frac{1}{N} \sum_{i=1}^{N} b_i \right| > \epsilon \right)
\]

\[
\leq q_1^N + q_2^N + 1_{\{|\beta a + |a|\epsilon_2| > \epsilon\}}
\]

\[
= q_1^N + q_2^N
\]

where the last step follows by the way \( \epsilon_1 \) and \( \epsilon_2 \) have been chosen.

There is still a point to be understood: in our derivation \( q_1 \) and \( q_2 \) depend on the choice of \( x \) through \( a \) and \( b \). However, it is immediate to check that \( a \) and \( b \) are both bounded in \( x \). This allows to conclude. \( \square \)

From (B.4) is immediate to see that \( \gamma_\infty \) is a bounded function of class \( C^1 \) and it has an important property which will be useful later on.

**Lemma B.2.** There exists a constant \( C > 0 \) such that

\[
x - \gamma_\infty(x) \geq C(x - \theta^*) \quad \text{if } x \in (\theta^*, +\infty)
\]

\[
\gamma_\infty(x) - x \geq C(\theta^* - x) \quad \text{if } x \in (-\infty, \theta^*)
\]

\[
\gamma_\infty(\theta^*) = \theta^*
\]

**Proof.**

If \( x \in (\theta^*, +\infty) \) and \( f \) is the density of each \( y_i \) (a mixture of two Gaussians) then

\[
x - \gamma_\infty(x) = \frac{1}{2\pi^3} \int_{x-\delta}^{x+\delta} (x-t) f(t) dt + \frac{1}{2\pi^3} \int_{R \setminus (x-\delta, x+\delta)} (x-t) f(t) dt
\]

\[
\geq \frac{1}{2\pi^3} \int_{x-\delta}^{x+\delta} f(t) dt + \frac{1}{2\pi^3} \int_{R \setminus (x-\delta, x+\delta)} f(t) dt
\]

where the last inequality follows from the fact that \( \int_{x-\delta}^{x+\delta} (x-t) f(t) dt \geq 0 \). We conclude that

\[
x - \gamma_\infty(x) \geq \frac{1}{2\pi^3} (x - \theta^*)
\]

\[
\geq \frac{1}{2\pi^3} \int_{x-\delta}^{x+\delta} f(t) dt + \frac{1}{2\pi^3} \int_{R \setminus (x-\delta, x+\delta)} f(t) dt > 0.
\]
Second statement if \( x \in (\infty, \theta^*) \) can be verified in a completely analogous way. The third statement then simply follows by continuity.  

We now come to a key result.

**Lemma B.3.** For any fixed \( \epsilon > 0 \), there exist \( \bar{q} \in (0, 1) \) and \( \chi > 0 \) such that

\[
\mathbb{P} \left( \gamma_N(x) \in \Theta_{\mathbb{Z}}(x) \right) \leq \chi \bar{q}^N \tag{B.6}
\]

for all \( x \) such that \( |x - \theta^*| > \epsilon \).

**Proof.** We assume \( x > \theta^* + \epsilon \) (the other case \( x < \theta^* - \epsilon \) being completely equivalent). Fix \( \epsilon' \in (0, C\epsilon) \) where \( C \) was defined in Lemma B.2 and estimate as follows

\[
\mathbb{P} \left( \gamma_N(x) \in \Theta_{\mathbb{Z}}(x) \right) \leq \mathbb{P} \left( \gamma_N(x) \in \Theta_{\mathbb{Z}}(x), \ |\gamma_N(x) - \gamma_{\infty}(x)| \leq \epsilon' \right) \\
+ \mathbb{P} \left( \gamma_N(x) - \gamma_{\infty}(x) > \epsilon' \right). \tag{B.7}
\]

Using Lemma B.2 we get

\[
\{ |\gamma_N(x) - \gamma_{\infty}(x)| \leq \epsilon' \} \subset \{ \gamma_N(x) \leq x - (C\epsilon - \epsilon') \}.
\]

Thus

\[
\{ \gamma_N(x) \in \Theta_{\mathbb{Z}}(x), |\gamma_N(x) - \gamma_{\infty}(x)| \leq \epsilon' \} \\
\subset \{ \exists i : y_i \in (\gamma_N(x) - \delta, \gamma_N(x) - \delta + \min(C\epsilon - \epsilon', \delta)) \}
\]

and, consequently, the first term in (B.7) can be estimated as

\[
\mathbb{P} \left( \gamma_N(x) \in \Theta_{\mathbb{Z}}(x), \ |\gamma_N(x) - \gamma_{\infty}(x)| \leq \epsilon' \right) \leq \left( 1 - \int_{\gamma_N(x) - \delta}^{\gamma_N(x) - \delta + \min(C\epsilon - \epsilon', \delta)} f(y)dy \right)^N \tag{B.8}
\]

where \( f(y) \) is the density of each \( y_i \). Considering now that \( f(y) \) is bounded away from 0 on any bounded interval, that \( |\gamma_N(x) - \gamma_{\infty}(x)| \leq \epsilon' \) and that \( \gamma_{\infty}(x) \) is a bounded function, we deduce that the right hand side of (B.8) can be uniformly bounded as \( \bar{q}^N \) for some \( \bar{q} \in (0, 1) \). Substituting in (B.7), and using Lemma B.1 we finally obtain the thesis. 

**B.2. Proof of Theorem 4.2.** Define

\[
\mathcal{A}_N(\epsilon) := \{ \exists \omega \in (\alpha, \beta)^V : \tilde{\theta}(\omega) \in \Theta_{\omega}, |\tilde{\theta}(\omega) - \theta^*| > \epsilon \}
\]

for any \( \epsilon > 0 \) and

\[
\begin{align*}
B_1 &:= \{ \exists i \in V : |y_i - \theta^*| > N \} \\
B_2 &:= \{ \exists (i, j) \in V \times V : |y_i - y_j| < N^{-4} \} \\
B_3 &:= \{ \exists (i, j) \in V \times V : |y_i - y_j| \in (2\delta, 2\delta + N^{-4}) \}
\end{align*}
\]

and estimate \( \mathbb{P} (\mathcal{A}_N(\epsilon)) \leq \mathbb{P} (\mathcal{A}_N(\epsilon) \cap B_1 \cap B_2 \cap B_3) + \mathbb{P}(B_1) + \mathbb{P}(B_2) + \mathbb{P}(B_3) \). Standard considerations allow to upper bound the probability of each event \( B_i \) by a common term \( K/N^2 \). We now focus on the estimation of the first term. The crucial point is that, the condition \( B_1 \cap B_2 \cap B_3 \) allow us to reinforce condition (B.3) in the sense
that all $\omega$ for which $\Theta_\omega \neq \emptyset$ can be obtained as $\omega = \tilde{\omega}(x)$ as $x$ varies in a set whose cardinality is polynomial in $N$. Specifically, define

$$Z = \{\zeta_j = \theta^* - N - \delta + jN^{-4} : j \in \mathbb{N}, j \leq j_{\text{max}}\}$$

where $j_{\text{max}} := [N^4(2N + 2\delta)]$ and notice that, assuming that the $y_i$’s satisfy $\mathcal{B}_2 \cap \mathcal{B}_3$, we have that $\tilde{\omega}(\zeta_j)$ and $\tilde{\omega}(\zeta_{j+1})$ differ in at most one component and that $\tilde{\omega}(x) \in \{\tilde{\omega}(\zeta_j), \tilde{\omega}(\zeta_{j+1})\}$ for every $x \in [\zeta_j, \zeta_{j+1}]$. Moreover, because of $\mathcal{B}_1$, we have that $\tilde{\omega}(x)_i = \tilde{\omega}(\zeta_0)_i = \beta$ for all $x \leq \theta^* - \delta$ and for all $i$. Similarly, $\tilde{\omega}(x)_i = \tilde{\omega}(\zeta_{\text{max}})_i = \beta$ for all $x \geq \theta^* + N + \delta$ and for all $i$. In other terms, under the assumption that the $y_i$’s satisfy $\mathcal{B}_1 \cap \mathcal{B}_2 \cap \mathcal{B}_3$, it holds $\{\omega \in \{\alpha, \beta\}^\gamma \mid \Theta_\omega \neq \emptyset\} = \{\tilde{\omega}(x) \mid x \in Z\}$. Hence,

$$\Pr(A_N(\epsilon), \mathcal{B}_1 \cap \mathcal{B}_2 \cap \mathcal{B}_3^c) \leq \Pr\left( \bigcup_{\zeta \in Z} \left\{ \gamma_N(\zeta) \in \Theta_{\tilde{\omega}(\zeta)}, |\gamma_N(\zeta) - \theta^*| > \epsilon \right\} \right)$$

$$\leq \Pr\left( \bigcup_{\zeta \in Z} \left\{ \gamma_N(\zeta) \in \Theta_{\tilde{\omega}(\zeta)}, |\gamma_N(\zeta) - \theta^*| > \epsilon/2 \right\} \right) + \Pr(|\gamma_N(\zeta) - \gamma_\infty(\zeta)| \leq \epsilon/2).$$

Notice that, because of the continuity of $\gamma_\infty$, there exists $\bar{\epsilon} > 0$ such that $|\gamma_\infty(\zeta) - \theta^*| > \epsilon/2 \Rightarrow |\zeta - \theta| > \bar{\epsilon}$. We can then use Lemma B.3,

$$\Pr\left( \bigcup_{\zeta \in Z} \left\{ \gamma_N(\zeta) \in \Theta_{\tilde{\omega}(\zeta)}, |\gamma_N(\zeta) - \theta^*| > \epsilon/2 \right\} \right)$$

$$\leq |Z| \Pr(\gamma_N(\zeta) \in \Theta_{\tilde{\omega}(\zeta)}, |\gamma_N(\zeta) - \theta^*| > \epsilon) \leq cN^5\tilde{q}^N$$

where $c$ and $\tilde{q}$ are those coming from Lemma B.3 relatively to $\bar{\epsilon}$. Putting together all the estimations we have obtained and using Lemma B.1, we finally obtain that there exists $\chi > 0$ such that $\Pr(A_N(\epsilon)) \leq \chi/N^2$. Using Borel-Cantelli Lemma and standard arguments, it follows now that the relation (4.5) hold in an almost surely sense.

It remains to be shown convergence in mean square sense. For this we need to go back to the form (3.10) of the derivative of $L(\theta, \tilde{\omega}(\theta))$. The key observation is that the second additive term in the right hand side of (3.10) can be bounded uniformly in modulus by some constant $C$. If we denote $\tilde{\gamma}_N = N^{-1} \sum y_i$, this implies that the function is increasing for $\theta > \tilde{\gamma}_N + \beta^2C$ and decreasing for $\theta < \tilde{\gamma}_N - \beta^2C$. Hence, necessarily,

$$|\xi - \tilde{\gamma}_N| \leq \beta^2C \ \forall \xi \in \mathcal{S}_N. \quad (B.9)$$

On the other hand, by the law of large numbers, $\tilde{\gamma}_N$ almost surely converges to $\theta^*$ and this implies, by the previous part of the theorem that $\max_{\xi \in \mathcal{S}_N} |\xi - \tilde{\gamma}_N|$ converges to 0. This, together with (B.9), yields $\mathbb{E} \max_{\xi \in \mathcal{S}_N} |\xi - \tilde{\gamma}_N|^2 \to 0$ for $N \to +\infty$. Since by the ergodic theorem also $\mathbb{E}|\tilde{\gamma}_N - \theta^*|^2 \to 0$ for $N \to +\infty$, the proof is complete.
B.3. Proof of Proposition 4.4. We prove it for $\hat{\omega}^{IA}$, the other verification being completely equivalent). If $\sigma \in \{\alpha, \beta\}$, we define

$$f(\theta, \sigma) = \mathbb{P}(\hat{\omega}(\theta)_i \neq \sigma | \omega^*_i = \sigma)$$

$$= \begin{cases} 
\frac{1}{\sqrt{2\pi\sigma^2}} \int_{\theta - \delta}^{\theta + \delta} e^{-\frac{(s - \theta^*)^2}{2\sigma^2}} ds & \text{if } \sigma = \beta \\
1 - \frac{1}{\sqrt{2\pi\sigma^2}} \int_{\theta - \delta}^{\theta + \delta} e^{-\frac{(s - \theta^*)^2}{2\sigma^2}} ds & \text{if } \sigma = \alpha
\end{cases}$$

(notice that $f$ does not depend on $i$). We can compute

$$\frac{1}{N} \mathbb{E}d_H(\hat{\omega}^{IA}, \omega^*) = \frac{1}{N} \sum_i \mathbb{P}(\hat{\omega}^{IA}_i \neq \omega^*_i)$$

$$= p \mathbb{E}f(\hat{\theta}^{IA}, \beta) + (1 - p) \mathbb{E}f(\hat{\theta}^{IA}, \alpha).$$

Since $f(\theta, \sigma)$ is a $C^1$ function of $\theta$, we immediately obtain that

$$|\mathbb{E}f(\hat{\theta}^{IA}, \sigma) - f(\theta^*, \sigma)| \leq C\mathbb{E}|\hat{\theta}^{IA} - \theta^*|$$

and, by Corollary 4.3, this last expression converges to 0, for $N \to +\infty$. Hence,

$$\frac{1}{N} \mathbb{E}d_H(\hat{\omega}^{IA}, \omega^*) = pf(\theta^*, \beta) + (1 - p) f(\theta^*, \alpha).$$

Straightforward computation now proves the thesis.