Large Deviations Rates for Distributed Inference

Submitted in partial fulfillment of the requirements for

the degree of

Doctor of Philosophy

in

Electrical and Computer Engineering

Dragana Bajović

Dipl. Ing., School of Electrical Engineering, University of Belgrade

Carnegie Mellon University

Pittsburgh, PA

Instituto Superior Técnico

Lisbon, Portugal

May 2013

To my parents, Dragoslav and Snežana

Acknowledgement

I would like to thank the members of my thesis committee, Professor Bruno Sinopoli and Professor José M. F. Moura from Carnegie Mellon University, Professor João Xavier and Professor João Pedro Gomes from Instituto Superior Técnico, Lisbon, Portugal, Professor Babak Hassibi from California Institute of Technology, and Professor Antonio Pedro Aguiar from The University of Porto, Porto, Portugal.

I acknowledge the support from the Fundação de Ciência e Tecnologia (FCT), Portugal.

Abstract

This thesis analyzes large deviations performance of linear consensus-based algorithms for distributed inference (detection and estimation). With consensus-based algorithms, agents communicate locally with their neighbors, through intermittently failing links, and assimilate their streaming observations in real time. While existing work usually focuses on asymptotic consistency and asymptotic normality measures, we establish the large deviations rates, thus giving parallels of the classical Chernoff lemma and Cramer's theorem for distributed systems. Our main contributions are two-fold. (1) We find the large deviation rate $\mathcal{J}$ for convergence in probability of products of random stochastic matrices that model the local inter-agent interactions. Our analysis includes a wide range of random stochastic matrix models, including asymmetric matrices and temporal dependencies. Further, for commonly used gossip and link failure models, we show how the rate $\mathcal{J}$ can be computed by solving a min-cut problem. (2) We find tight upper and lower bounds on the large deviations performance of linear consensus-based inference, as well as the full large deviation principle when the underlying network is regular. When translated into distributed detection, our results reveal a phase transition behavior with respect to the network connectivity, measured by $\mathcal{J}$. If $\mathcal{J}$ is above a threshold, each agent is an asymptotically optimal detector with the error exponent equal to the total Chernoff information of the network; if below the threshold, we characterize what fraction of the total Chernoff information can be achieved at each agent. When translated into distributed estimation, our results show that distributed system's performance relative to the performance of an ideal, centralized system, is a highly nonlinear function of the required estimation accuracy. Finally, our methodology develops new tools that are of general interest in the large deviations theory and products of random matrices.

Contents

Chapter 1

Introduction

This thesis is mainly devoted to the large deviations performance analysis of consensus-based distributed algorithms for inference (detection, estimation) in networks. With these algorithms, each agent i in an N-agent network collects (possibly vector) measurements $y_{i,k}$ over time k to perform the inference task. At each time k, agent i updates its state (solution estimate) $x_{i,k} \in \mathbb{R}^d$, by mixing it with the states of immediate neighbors (consensus), and by accounting for the new local measurement $y_{i,k}$ (innovation), [1, 2]. The update of $x_{i,k}$ is generally of the form:

$$\widehat{x}_{i,k} = \sum_{j \in O_{i,k}} W_{ij,k}\, x_{j,k-1} \qquad \text{(consensus)} \tag{1.1}$$

$$x_{i,k} = \widehat{x}_{i,k} - \frac{1}{k}\left(\widehat{x}_{i,k} - Z_{i,k}\right) \quad \text{(innovation)}. \tag{1.2}$$

In (1.1), $\widehat{x}_{i,k}$ is an intermediate state, $O_{i,k}$ is a (possibly random) neighborhood of agent i (including i), and $W_{ij,k} \geq 0$ is the averaging weight, with $\sum_{j \in O_{i,k}} W_{ij,k} = 1$; in (1.2), $Z_{i,k}$ is a deterministic function of $y_{i,k}$. For example, with distributed detection, $Z_{i,k} = \log \frac{f_1(y_{i,k})}{f_0(y_{i,k})}$ is the log-likelihood ratio based on the measurement $y_{i,k}$. We refer to quantity $Z_{i,k}$ as the innovation.

Recently, there has been a strong interest in the literature for distributed algorithms of type (1.1)–(1.2), due to several favorable features of such algorithms: 1) they process data online; 2) utilize only inter-neighbor communications, hence avoiding the fusion-node bottleneck; and 3) exhibit resilience to random communication failures. Algorithms (1.1)–(1.2) have been studied in several application contexts, including detection and estimation in sensor networks [2], modeling swarm behavior of robots/animals [3, 4], detection of a primary user in cognitive radio networks [5], and power grid state estimation [6].

A relevant aspect with (1.1)–(1.2) is the *randomness* in communication links. For example, with wireless

sensor networks, packet dropouts may occur at random times. We analyze (1.1)–(1.2) allowing for the random underlying networks and link failures.

We now explain the main objective of the thesis, while we detail our contributions in the subsequent section. For algorithms (1.1)–(1.2), existing literature provides convergence guarantees to the desired value θ (e.g., true parameter in distributed estimation), under mild conditions on the underlying network and the measurements' distribution. Our objective is to establish the *large deviations rates of convergence*. More specifically, suppose that $x_{i,k}$ converges to θ as $k \to \infty$ in probability, so that, for any (measurable) set $E \subset \mathbb{R}^d$ that does not contain θ:

$$\mathbb{P}\left(x_{i,k} \in E\right) \to 0 \text{ as } k \to \infty. \tag{1.3}$$

We ask how fast $\mathbb{P}\left(x_{i,k} \in E\right)$ converges to zero. In many situations, it turns out that $\mathbb{P}\left(x_{i,k} \in E\right)$ converges to zero exponentially fast in k (per-agent number of measurements), i.e.,

$$\mathbb{P}\left(x_{i,k} \in E\right) \sim e^{-k\, I(E)}, \tag{1.4}$$

with a certain $I(E) \geq 0$. Our goal is to determine $I(E)$ as a function of the underlying system parameters – network topology and distribution of the agents' measurements.

The metric $I(E)$ is highly relevant in distributed inference. For example, in distributed detection – hypothesis testing H_1 versus H_0, $x_{i,k}$ is compared at each time k against the zero threshold to make the decision (H_1 or H_0):

$$x_{i,k} \underset{H_0}{\overset{H_1}{\gtrless}} 0.$$

The relevant detection metrics – probability of false alarm, probability of miss, and average error probability are respectively (here π_0, π_1 are the prior probabilities):

$$\begin{aligned}
\alpha(k) &= \mathbb{P}\left(x_{i,k} \geq 0 \,|\, H_0\right) \\
\beta(k) &= \mathbb{P}\left(x_{i,k} < 0 \,|\, H_1\right) \\
P_{\mathrm{e}}(k) &= \pi_0\, \alpha(k) + \pi_1\, \beta(k),
\end{aligned}$$

and they are of type (1.3).[1] The large deviation rates that correspond to them are known as error exponents and are standard detection metrics, e.g., [7]. For the classical, centralized detection, the rates that corresponds to the three probabilities are the same and are known as the Chernoff information, [7, 8]. In

[1] For example, consider the probability of false alarm where we take $E = [0, +\infty)$. It can be shown that, conditioned on H_0, $x_{i,k}$ converges in probability to $\theta = \mathbb{E}\left[\log \frac{f_1(y_{i,k})}{f_0(y_{i,k})}\right]$ which is negative, so that $\theta \notin E$.

alternative, in distributed estimation of a vector θ, the rate $I(\xi)$ with:

$$\mathbb{P}\left(\|x_{i,k}-\theta\|>\xi\right)\sim e^{-k\,I(\xi)},\ \ \xi\geq 0,$$

is known as the inaccuracy rate [9]. It measures how fast (in k) the probability $\mathbb{P}\left(\|x_{i,k}-\theta\|\leq\xi\right)$ enters a given confidence level, e.g., 98%.

1.1 Thesis Contributions

We now state the main contributions of the thesis. We explain the contributions chapter by chapter.

Chapter 2: Products of random stochastic matrices: The symmetric i.i.d. case

The network effect on distributed inference (1.1)–(1.2) is captured through the $N\times N$ random, stochastic[2] weight matrices W_k that collect the weights $W_{ij,k}$'s. More precisely, performance of (1.1)–(1.2) is affected by the products $W_k\cdots W_1$. It is known that, under certain conditions on the underlying network, the product $W_k\cdots W_1$ converges in probability to the ideal consensus matrix $J:=\frac{1}{N}11^\top$ (whose all entries are equal to $1/N$). We reveal by our analysis that the key parameter that controls the large deviations performance of (1.1)–(1.2) is:

$$\mathcal{J}:=\lim_{k\to\infty}-\frac{1}{k}\log\mathbb{P}\left(\|W_k\cdots W_1-J\|\geq\epsilon\right),\ \ \epsilon\in(0,1], \tag{1.5}$$

where $\|\cdot\|$ is the spectral norm. We refer to $\mathcal{J}$ as the large deviation rate of consensus. Clearly, the larger $\mathcal{J}$ is, the faster consensus (convergence of product $W_k\cdots W_1$) is. In a sense, $\mathcal{J}$ measures the "connectivity" of the network – the larger $\mathcal{J}$, the better the network connectivity. The quantity $\mathcal{J}$ has not been computed in the literature before. We characterize the rate $\mathcal{J}$ for generic random, symmetric, independent, identically distributed (i.i.d.) matrices W_k, showing that $\mathcal{J}$ is solely a function of the graphs induced by the matrices W_k and the corresponding probabilities of occurrences of these graphs [10, 11]. Further, we show that $\mathcal{J}$ does not depend on $\epsilon\in(0,1]$. As mentioned, calculation of $\mathcal{J}$ is the key to analyzing the large deviations performance of algorithms (1.1)–(1.2), but it is also important in its own right for the theory of products of stochastic matrices [12, 13], non-homogenous Markov chains [14] and the standard averaging (consensus) algorithms, e.g., [15].

In general, computation of the large deviation rate $\mathcal{J}$ is a combinatorial problem. However, we calculate or tightly approximate $\mathcal{J}$ for many important cases. We give a closed form solution for $\mathcal{J}$ for arbitrary

[2]Stochastic means that rows of the matrix sum to one and all its entries are nonnegative.

type of averaging that runs on a tree. For example, for standard gossip (see Subsection 2.5.1) on a tree, $\mathcal{J} = |\log(1 - p_e)|$, where p_e is the probability of the link e that is least likely to occur. Further, we give closed form expressions for $\mathcal{J}$ for standard gossip and link failures over a regular graph. For those, we show that $\mathcal{J}$ equals $|\log p_{\text{isol}}|$, where p_{isol} is the probability that an agent is isolated from the rest of the network.

Further, we give simple formulas for $\mathcal{J}$ for *generic networks* and commonly used gossip and link failure models. For gossip, we show that $\mathcal{J} = |\log(1 - c)|$, where c is the min-cut value (or connectivity [16]) of a graph whose links are weighted by the gossip link probabilities; the higher the connectivity c is (the more costly or, equivalently, less likely it is to disconnect the graph) the larger the rate $\mathcal{J}$ and the faster the convergence are. This is illustrated in Figure 1.1.

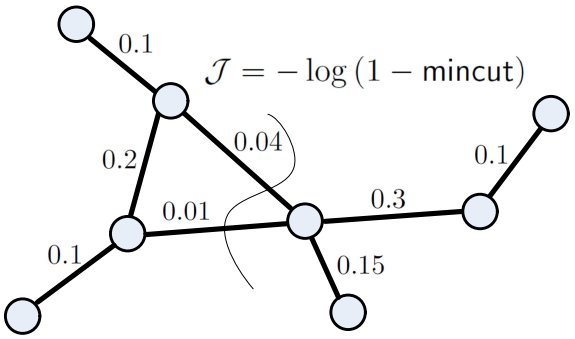

Figure 1.1: The rate $\mathcal{J}$ for random gossip algorithm is given by $\mathcal{J} = -\log(1 - \text{mincut})$ where mincut is the minimum over all cuts of the graph with the edge costs equal to the corresponding link activation probabilities in gossip; it can be seen that for the example in the figure $\text{mincut} = 0.05$.

Similarly, we show that $\mathcal{J}$ is computed efficiently via min-cut for *link failures on general graphs*, with both independent and correlated fading (failing). Finally, we find easily computable tight approximations for $\mathcal{J}$ for a more general class of gossip-like models including symmetrized broadcast gossip.

Chapter 3: Products of random stochastic matrices: Temporal dependencies and directed networks

We extend the results in Chapter 2 in two ways, by considering: 1) temporally dependent sequences of (symmetric) matrices W_k; and 2) temporally i.i.d., asymmetric (not necessarily doubly stochastic) matrices W_k.

1) Temporal dependencies. Our temporally dependent model of the W_k's associates a state of a Markov chain to each of the possible realizations G_t of graphs that supports W_t. The distribution of the graphs G_t, $t \geq 1$, is determined by an $\mathcal{M} \times \mathcal{M}$ transition probability matrix P, where $\mathcal{M}$ is the number of possible realizations of G_t. This model subsumes, e.g., the token-based protocols similar to [17], or temporally dependent link failure models, where the on/off state of each link follows a Markov chain. We characterize the

rate $\mathcal{J}$ as a function of the transition probability matrix P. We briefly convey here the general idea behind the result. The rate $\mathcal{J}$ is determined by the most likely way in which the Markov chain stays "trapped" in a subset of states (graphs) whose union is disconnected. The probability of this event is determined by the spectral radius of the block in the transition matrix P that corresponds to this most likely subset of states, and this spectral radius determines the rate $\mathcal{J}$. We illustrate the results on two examples, namely gossip with Markov dependencies and temporally correlated link failures. The example with temporally correlated link failures shows that "negative temporal correlations" of the links' states (being ON or OFF) increase (improve) the rate $\mathcal{J}$ when compared with the uncorrelated case, while positive correlations decrease (degrade) the rate. This result is in accordance with standard large deviations results on temporally correlated sequences, see, e.g., [[18], exercise V.12, page 59].

2) Directed networks – Asymmetric matrices. We study temporally i.i.d. asymmetric matrices W_k, hence relaxing the symmetricity assumption from Chapter 2. It is well known that, under certain conditions, rather than converging to $J = \frac{1}{N}11^\top$, the product $W_k \cdots W_1$ here converges almost surely to a random rank-one matrix $1v^\top$. (Here, the vector $v \in \mathbb{R}^N$ is random.) A natural measure of the "distance" of the product $W_k...W_1$ to its limiting space is $|\lambda_2(W_k \cdots W_1)|$ – the modulus of the second largest (in modulus) eigenvalue of the product $W_k \cdots W_1$. Henceforth, a natural counterpart to (1.5) is the following rate:

$$\mathcal{J}_{\text{dir}} := \lim_{k \to +\infty} -\frac{1}{k} \log \mathbb{P}\left(|\lambda_2(W_k \cdots W_1)| \geq \epsilon\right), \ \epsilon \in (0, 1]. \tag{1.6}$$

We fully characterize the limit $\mathcal{J}_{\text{dir}}$ and show that, similarly as in the case of symmetric W_k, $\mathcal{J}_{\text{dir}}$ depends on the distribution of matrices W_k only through their support graphs. More precisely, $\mathcal{J}_{\text{dir}}$ is determined by the probability of the most likely set of support graphs whose union fails to form a directed spanning tree. Thus, the characterization of $\mathcal{J}_{\text{dir}}$ exhibits full consistency with the result for symmetric matrices in Chapter 2: for undirected networks a collection of topologies is jointly tree-free[3] if and only if it is disconnected. Thus, when the matrices are symmetric, the two rates $\mathcal{J}$ in (1.5) and $\mathcal{J}_{\text{dir}}$ in (1.6) coincide. Finally, we illustrate our results on a commonly used broadcast gossip protocol [19] in sensor networks, where (only one) node u activates at a time with probability p_u, and broadcasts its state to all single-hop neighbors. For this model, the rate $\mathcal{J} = |\log 1 - p_{\min}|$, where $p_{\min}$ is the probability of the most rarely active node.

[3]We say that a collection of topologies (graphs) is tree-free if the graph that contains the edges of all the graphs in the collection does not contain a directed spanning tree.

Chapter 4: Large deviations for distributed inference

We consider algorithms (1.1)–(1.2) for vector innovations $Z_{i,k}$ with generic distributions, hence encompassing distributed vector-parameter estimation in sensor networks. The matrices W_k are assumed i.i.d., but they may be asymmetric (directed networks). We study the large deviation rates $I(E)$:

$$\mathbb{P}\left(x_{i,k} \in E\right) \sim e^{-k\, I(E)}$$

for generic sets $E \subset \mathbb{R}^d$. We obtain several results, as we outline below.

1) Spatio-temporally i.i.d. innovations and asymmetric W_k's. For spatio-temporally i.i.d. observations and asymmetric matrices W_k, we show that performance $I(E)$ of distributed inference (1.1)–(1.2) is always better (or at worse the same) as the performance of the agents working in isolation. Further,distributed inference (1.1)–(1.2) is always worse (or at best equal) to the performance of a centralized, fusion-based inference that has access to all agents' innovations at all times. The result is intuitive, as it says that cooperation cannot "hurt." Likewise, a distributed system that does not have access to full information cannot outperform the centralized system with full information. Although very intuitive, the result was surprisingly difficult to prove.

2) Spatio-temporally i.i.d. innovations and symmetric W_k's. When the W_k's are symmetric (and still i.i.d. in time), we establish tight upper and lower bounds on the large deviations performance $I(E)$. The results reveal a very interesting interplay between the underlying network and the distribution of the agents' innovations, which we explain here at a qualitative level. To make our point clear, consider the sets E of type

$$E_\xi = \{x \in \mathbb{R}^d : \|x - \theta\| > \xi\}, \quad \xi > 0,$$

where θ is the mean of the innovations $Z_{i,k}$'s. Hence, requiring that the estimate $x_{i,k} \notin E_\xi$ for a very small ξ means requiring a very high estimation precision (high confidence); conversely, a large ξ corresponds to a coarse estimation. Our results show the following nonlinear behavior. Distributed inference (1.1)–(1.2) is close to the centralized performance for very high precisions (very small ξ's) and becomes much worse from the centralized performance for very coarse precisions. Intuitively, reaching high precisions is intrinsically difficult even for the centralized system, and hence the network-wide averaging process in (1.1)–(1.2) has sufficient time to "catch up" with the centralized system. On the other hand, the centralized system reaches a coarse accuracy very quickly, so that the distributed system cannot "catch up." The point $\xi^\star$ where the behavior significantly changes depends on the quantity $\mathcal{J}$ in (1.5), number of agents N, and distribution of

the $Z_{i,k}$'s.

3) Spatio-temporally i.i.d. innovations and regular random networks. When we additionally assume regular networks, we establish the full large deviations principle (see Chapter 4 for details) for (1.1)–(1.2) and we characterize explicitly the corresponding rate function. More precisely, for generic sets E and certain additional conditions, we show that for any node i:

$$\lim_{k\to\infty} \frac{1}{k} \log \mathbb{P}\left(x_{i,k} \in E\right) = -I_{\mathcal{J},N}(E),$$

where $I_{\mathcal{J},N}(\cdot)$ is the same at all nodes. Our result reveal that $I_{\mathcal{J},N}$ has a very neat form: $I_{\mathcal{J},N}$ is the convex hull of two functions, $I(\cdot) + \mathcal{J}$ and $NI(\cdot)$, where $I(\cdot)$ is the rate function of a node working in isolation, $\mathcal{J}$ is the large deviation rate of consensus in (1.5), and $NI(\cdot)$ is the optimal rate function, i.e., the rate function of a (fictional) fusion node. Figure 1.2 illustrates $I(\cdot) + \mathcal{J}$, $NI(\cdot)$ and $I_{\mathcal{J},N}(\cdot)$ for the case when the nodes' observations are Gaussian.

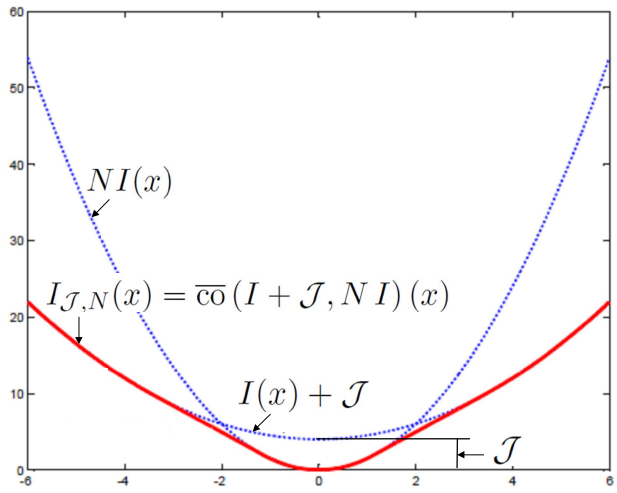

Figure 1.2: Illustration of $I_{\mathcal{J},N}$ for a network of size $N = 3$, with $\mathcal{J} = 5$, and $Z_{i,t} \sim \mathcal{N}(0,1)$. The more curved blue dotted line plots $NI(x) = \frac{1}{2}Nx^2$, the less curved blue dotted line plots $I(x) + \mathcal{J} = \frac{1}{2}x^2 + \mathcal{J}$. The solid red line plots $I_{\mathcal{J},N} = \overline{\text{co}}\left(NI(\cdot), I(\cdot) + \mathcal{J}\right)$.

4) Spatially different innovations and symmetric W_k's. We extend the above conclusions to spatially non-identically distributed observations, hence encompassing distributed Gaussian estimation in sensor networks. We show that distributed estimation exhibits qualitatively similar behavior as described above. For sufficiently high accuracy (sufficiently small ξ), it achieves the performance of the centralized estimator; for a very coarse accuracy, distributed estimator can be significantly poorer than the centralized estimator.

Chapter 5: Large deviations for Distributed detection

We establish the large deviations performance of distributed detection of type (1.1)–(1.2) for random networks. Specifically, we consider distributed detector proposed in [1]. This result for distributed detection may be seen as a counterpart result to the (centralized detection's) Chernoff lemma [8]. We allow for random, symmetric, i.i.d. matrices W_k and generic (non-Gaussian) distributions of the agents' measurements. We show that distributed detection exhibits a phase transition behavior with respect to the large deviations rate of consensus $\mathcal{J}$ in (1.5) (network connectivity). If $\mathcal{J}$ is above a threshold, then the large deviations rate of detection error probability with distributed detector equals the Chernoff information–the best possible rate of the optimal centralized detector. Thus, when the network connectivity is above the threshold, distributed detection is as good as the optimal centralized detector. When $\mathcal{J}$ is below the threshold, we find what fraction of the centralized detector's performance can distributed detector achieve.

We demonstrate how the optimality threshold is a function of the logarithmic moment generating function of the measurements' log-likelihood ratios and of the number of agents N. This reveals for the performance of distributed detection a very interesting interplay between the distribution of the agents' measurements (e.g., Gaussian or Laplace) and the network connectivity (the value of $\mathcal{J}$). We show that, for the same network connectivity (same $\mathcal{J}$), a distributed detector with given observations distributions, say, Laplace, may match the optimal asymptotic performance of the centralized detector, while the distributed detector for Gaussian observations may be suboptimal, even though the centralized detectors for the two distributions, Laplace and Gaussian, have the same optimal asymptotic performance. (See Figure 1.3 for an illustration.) This is a very interesting effect with distributed detection that does not have a parallel in the classical, centralized detection. Figure 1.3 illustrates the dependence of distributed detection's large deviation rate on $\mathcal{J}$ (network connectivity) and the observations' distribution. We can see that, at the value $\mathcal{J} = \mathcal{J}^\star$, further increase of $\mathcal{J}$ does not pay off in terms of detection performance, as we have already reached the best, centralized detector's level. Hence, in a sense, the threshold value $\mathcal{J}^\star$ represents the optimal operating point of the network. Finally, we address the problem of "targeting" the value $\mathcal{J}^\star$ when the inter-agent links are fading, and the fading probabilities depend on the allocated transmission powers. We optimally allocate the agents' transmission powers such that the value $\mathcal{J}^\star$ is achieved with the minimal overall (across agents) invested power.

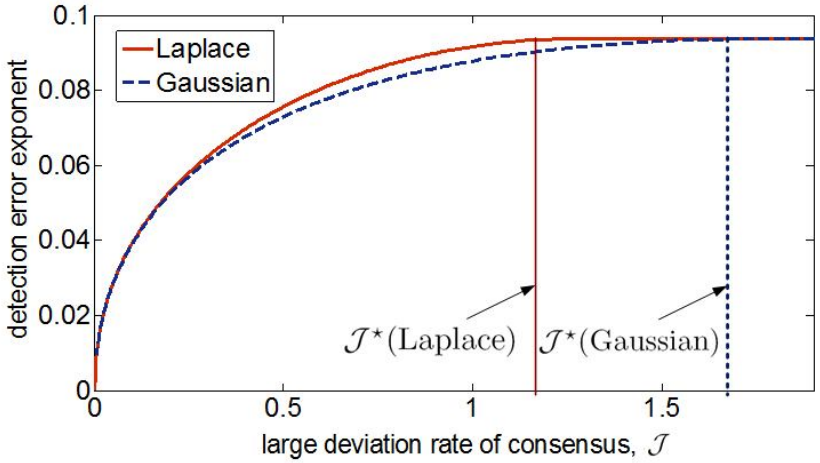

Figure 1.3: Error exponent versus the large deviations rate of consensus $\mathcal{J}$ for the Gaussian and Laplace sensor observations. The saturation level of the error exponent in the figure is the optimal centralized detector's error exponent. The centralized Gaussian and Laplace detectors are equivalent, while distributed detectors are not equivalent: Laplace distributed detector has a lower value of the threshold $\mathcal{J}^\star$. Simulated data are: $N = 20$, $C_{\rm ind} = C_{\rm ind,L} = C_{\rm ind,G} = 0.005$, $b_{\rm L} = 1$, $m_{\rm L} = 0.2$, and $m_{\rm G}^2/\sigma_{\rm G}^2 = 0.04 = 8C_{\rm ind}$. (See Section 5.3 for details.)

1.2 Literature review

We now provide a literature review that help us contrast our contributions with existing work. We consider separately the literature on standard consensus and products of stochastic matrices and the literature on distributed inference.

Literature on consensus and products of stochastic matrices

There has been a large amount of work on distributed averaging (consensus) algorithms and products of stochastic matrices. In distributed averaging, each agent has a scalar measurement y_i, and the goal is for all agents to find the global average $\frac{1}{N}\sum_{i=1}^N y_i$. This task can be done via the consensus algorithm, where the network-wide state $x_k = (x_{1,k}, ..., x_{N,k})^\top$ updates as $x_{k+1} = W_k\, x_k$, and W_k is the weight matrix that respects the sparsity pattern of the network (as in (1.1)–(1.2)). Early work on consensus includes [20, 21], and the topic received renewed interest in the past decade [22, 15]. Reference [15] analyzes convergence of the consensus algorithm under deterministic time-varying matrices W_k. Reference [23] provides a detailed study of the standard gossip model, that has been further modified, e.g., in [24, 25]; for a recent survey, see [26]. Reference [27] analyzes convergence under random matrices W_k, not necessarily symmetric, and ergodic – hence not necessarily independent in time. Reference [28] studies effects of delays,

while reference [29] studies the impact of quantization. Reference [30] considers random matrices W_k and addresses the issue of the communication complexity of consensus algorithms. The recent reference [31] surveys consensus and averaging algorithms and provides tight bounds on the worst case averaging times for deterministic time varying networks.

Existing works mostly study the products $W_k \cdots W_1$ in the context of (standard) consensus or gossip algorithms, and not with the *consensus* and *innovations* algorithms of type (1.1)–(1.2). Hence, these works consider certain convergence metrics different than $\mathcal{J}$. In contrast, our main concern are the algorithms of type (1.1)–(1.2), for which $\mathcal{J}$ appears as a natural convergence metric. For example, references [23, 32] consider the ϵ-averaging time, and $\lambda_2(\mathbb{E}[W_k^2])$. Further, [33] considers $\lim_{k\to\infty} \frac{1}{k} \log \mathbb{E}[\|W_k ... W_1 - J\|^2]$. To our best knowledge, the *exact* large deviations rate $\mathcal{J}$ in (1.5) has not been computed for i.i.d. averaging matrices W_k, nor for the commonly used sub-classes of gossip and link failure models. From existing results, one can deduce upper bounds on $\mathcal{J}$, but not the exact rate $\mathcal{J}$. (See [10] for an explanation how this can be done.)

Products of random matrices appear also in many other fields that use techniques drawn from Markov process theory. Examples include repeated interaction dynamics in quantum systems [13], inhomogeneous Markov chains with random transition matrices [34, 13], infinite horizon control strategies for Markov chains and non-autonomous linear differential equations [12], or discrete linear inclusions [35]. These papers are usually concerned with proving convergence of the products and determining the limiting matrix. Reference [13] studies the product of matrices belonging to a class of complex contraction matrices and characterizes the limiting matrix by expressing the product as a sum of a decaying process, which exponentially converges to zero, and a fluctuating process. Reference [12] establishes conditions for strong and weak ergodicity for both forward and backward products of stochastic matrices, in terms of the limiting points of the matrix sequence. Using the concept of infinite flow graph, which the authors introduced in previous work, reference [34] characterizes the limiting matrix for the product of stochastic matrices in terms of the topology of the infinite flow graph. For more structured matrices, [36] studies products of nonnegative matrices. For nonnegative matrices, a comprehensive study of the asymptotic behavior of the products can be found in [14]. A different line of research, closer to our work, is concerned with the limiting distributions of the products (in the sense of the central limit theorem and large deviations). The classes of matrices studied are: invertible matrices [37, 38] and its subclass of matrices of determinant equal to 1 [39] and, also, positive matrices [40]. None of these apply to our case, as the matrices that we consider are not invertible ($W_k - J$ has a zero eigenvalue, for every realization of W_k) and, also, we allow the entries of W_k to be zero, and therefore the entries of $W_k - J$ might be negative with positive probability. Furthermore, as pointed out

in [41], the results obtained in [37, 38, 39] do not provide ways to effectively compute the rates of convergence. Reference [41] improves on the existing literature in that sense by deriving more explicit *bounds* on the convergence rates, while showing that, under certain assumptions on the matrices, the convergence rates do not depend on the size of the matrices; the result is relevant from the perspective of large scale dynamical systems, as it shows that, in some sense, more complex systems are not slower than systems of smaller scale, but again it does not apply to our study.

Literature on distributed inference

Distributed inference has been extensively studied, in the context of parallel fusion architectures, e.g., [42, 43, 44, 45, 46, 47, 48], consensus-based inference [49, 50, 51, 52], and, more recently, consensus+innovations distributed inference, see, e.g., [53, 2, 54, 55, 56] for distributed estimation, and [57, 58, 1, 5, 59, 60, 61] for distributed detection. Different variants of consensus+innovations distributed detection algorithms have been proposed; we analyze here the algorithm in [1]. In [62], we considered deterministically time varying networks, where the union networks over finite time windows are connected. In [63, 64], we study random networks, where [63] considers Gaussian agents' measurements, while in [64] we consider generic agents' measurements. Reference [65] considers the large deviations performance of a different consensus+innovations detection algorithm when the noise is Gaussian and the communications among sensors are noisy (additive noise).

Reference [1] considers distributed detection's asymptotic optimality, but in a framework that is very different from ours. Reference [1] studies the asymptotic performance of the distributed detector where the means of the sensor observations under the two hypothesis become closer and closer (vanishing signal to noise ratio (SNR)), at the rate of $1/\sqrt{k}$, where k is the number of observations. For this problem, there is an asymptotic, non-zero, probability of miss and an asymptotic, non-zero, probability of false alarm. Under these conditions, distributed detector is as efficient as the optimal centralized detector, [66], as long as the network is connected on average. Here, we assume that the means of the distributions stay fixed as k grows. We establish the large deviations rate of detection error probability, showing that detection error decays to zero exponentially fast as k goes to infinity. We show that connectedness on average is not sufficient for distributed detector to achieve the optimality of centralized detection; rather, phase transition occurs, with distributed becoming as good as centralized, when the network connectivity (the value of $\mathcal{J}$), exceeds a certain threshold.

We now contrast our work with reference [2]. The latter reference considers distributed estimation algorithms of type (1.1)–(1.2) under very general conditions on the underlying network and the agents'

measurements; it also allows for the inter-agent additive communication noise and for the nonlinear state updates. The reference proves convergence of the state $x_{i,k}$ to the true parameter θ in the sense of: 1) consistency: $x_{i,k} \to \theta$, almost surely; 2) asymptotic unbiasedness: $\mathbb{E}[x_{i,k}] \to \theta$; and 3) *asymptotic normality*: $\frac{1}{\sqrt{k}}(x_{i,k} - \theta) \to \mathcal{N}(0, S)$ in distribution, where $\mathcal{N}(0, S)$ is a Gaussian random variable with zero mean and covariance matrix S. In contrast, we study the *large deviations rates of convergence*, in the sense of (1.4). Among the noted three aspects of convergence studied in [2], the closest study to ours is that of asymptotic normality; but, the large deviations rate and asymptotic normality are different; see, e.g., [8]. While asymptotic normality captures only information about the first and second moments of $x_{i,k}$, the large deviations rates capture the information about all moments (full distribution) of $x_{i,k}$. The two metrics are equivalent only when $x_{i,k}$ is Gaussian, but, due to randomness of the underlying network (randomness of W_k) assumed here, $x_{i,k}$ is not Gaussian even when the agents' measurements are Gaussian.

Chapter 2

Products of Random Stochastic Matrices: The Symmetric I.I.D. Case

2.1 Introduction

We study the convergence in probability of products $W_k \cdots W_1$ of (doubly) stochastic symmetric $N \times N$ matrices W_t, $t \geq 1$. These products arise in many contexts; we consider a power allocation application in distributed detection in Chapter 5. When 1) the matrices W_t are independent and identically distributed (i.i.d.), 2) the support graph of the expected matrix $\mathbb{E}[W_t]$ is connected, and 3) each W_t has positive diagonals almost surely, it is well known that these products converge to the average consensus matrix $J = \frac{1}{N} 1 1^\top$ almost surely [27], hence in probability. The goal of the current chapter is to study the rate of this convergence in probability – namely, we establish that this convergence in probability is *exponentially fast*, and we determine the *exact exponential rate* of this convergence.

We explain our problem in intuitive terms. Consider the static (deterministic) case $W_t = A$, for all $t \geq 1$, where A is a (doubly) stochastic symmetric matrix with $|\lambda_2(A)| < 1$; let $\|\cdot\|$ denote the spectral norm. Then $\|W_k \cdots W_1 - J\| = |\lambda_2(A)|^k$, or, in words, the spectral norm of the error matrix $W_k \cdots W_1 - J = A^k - J$ decays exponentially fast with exponent $|\lambda_2(A)|$. When the W_k's are random i.i.d. matrices, a similar behavior occurs, but now the role of $|\lambda_2(A)|$ is taken by the Lyapunov exponent $\gamma < 1$, i.e., the path of the norm $\|W_k \cdots W_1 - J\|$, $k \geq 1$, behaves as γ^k [67, 33, 68][1]. But, contrary to the deterministic case, because the W_k's are random, there are paths of the norm $\|W_k \cdots W_1 - J\|$, $k \geq 1$, that decay slower than γ^k, although with vanishing probability as the size k of the product increases. To be specific, consider an

[1] More precisely, $\lim_{k \to +\infty} (\|W_k \cdots W_1 - J\|)^{\frac{1}{k}} = \gamma$, almost surely. We also remark that γ is a constant that depends only on the statistics of the matrices W_k (and not on the particular choice of the sequence realization $W_k, ..., W_1$), see also [67].

arbitrary $\epsilon \in (0,1]$ and, for large k, the *rare* event $\{\|W_k \cdots W_1 - J\| \geq \epsilon\}$. In this chapter, we consider the probability of such rare events and the rate at which the sequence of these probabilities vanishes with k; in particular, we show that the following *large deviation rate* $\mathcal{J}$ exists

$$\mathcal{J} = \lim_{k\to\infty} -\frac{1}{k} \log \mathbb{P}\left(\|W_k \cdots W_1 - J\| \geq \epsilon\right) \tag{2.1}$$

and we show how it can be computed in terms of network parameters and the statistics of the W_k's. In fact, we provide a stronger result on the rate $\mathcal{J}$. We show that the same large deviation rate $\mathcal{J}$ holds for the following events. Let d_k, $k \geq 1$, be a sequence with a decay rate slower than exponential; e.g., $d_k = \frac{1}{k}$, for $k \geq 1$. Similarly to the case when $d_k \equiv \epsilon$, consider the rare event $\{\|W_k \cdots W_1 - J\| \geq d_k\}$. This is a rare event because $\|W_k \cdots W_1 - J\| \sim \gamma^k << d_k$. We show that the large deviation rate at which the probabilities of these rare events vanish with k is the same as the rate $\mathcal{J}$ in (2.1). More precisely, for any sequence d_k, $k \geq 1$, $d_k \in (0,1]$, $\log d_k = o(k)$,

$$\mathcal{J} = \lim_{k\to\infty} -\frac{1}{k} \log \mathbb{P}\left(\|W_k \cdots W_1 - J\| \geq d_k\right), \tag{2.2}$$

and the rate $\mathcal{J}$ is the same for any such sequence d_k.

Our results reveal that the large deviation rate $\mathcal{J}$ is solely a function of the graphs induced by the matrices W_k and the corresponding probabilities of occurrences of these graphs. In general, the computation of the rate $\mathcal{J}$ is a combinatorial problem. However, for special important cases, we can get particularly simple expressions. For example, when the matrices W_k are the weight matrices for gossip consensus on a tree, the rate $\mathcal{J}$ is equal to $|\log(1-p_{ij})|$, where p_{ij} is the probability of the link $\{i,j\}$ that is least likely to occur in the gossip protocol. Another example is with gossip consensus over a regular graph when $p_{ij} \equiv p$ in which case we show that the rate $\mathcal{J}$ equals $|\log p_{\text{isol}}|$, where $p_{\text{isol}} = 1 - dp$ is the probability that a node is isolated from the rest of the network, and d is the degree of a node. For gossip over more general graph structures, we show that $\mathcal{J} = |\log(1-c)|$ where c is the min-cut value (or connectivity [16]) of a graph whose links are weighted by the gossip link probabilities; the higher the connectivity c is (the more costly or, equivalently, less likely it is to disconnect the graph) the larger the rate $\mathcal{J}$ and the faster the convergence are. Similarly, for consensus algorithms running on networks with link failures on general graphs, the rate is computed by solving a min-cut problem and is hence computable in polynomial time.

We further establish that for a generic model of W_k's, calculation of the rate $\mathcal{J}$ is equivalent to solving a generalized min-cut problem. Albeit solving the latter is computationally hard in general, we approximate the rate $\mathcal{J}$ efficiently for a class of gossip-like models that subsumes, e.g., standard pairwise gossip and

symmetrized broadcast gossip. For this class, we provide easily computable tight approximations of $\mathcal{J}$. We also explicitly calculate the rate $\mathcal{J}$ for the correlated fading. Namely, we show that, with this model, there is a single critical link that determines the rate; this link marks the transition between the connected and disconnected regime of network operation. Finally, we give a closed form solution for $\mathcal{J}$ for arbitrary type of averaging that runs on a tree.

The results from this chapter on the rate $\mathcal{J}$ are based on our work in [10] and [11].

Chapter organization. Section 2.2 introduces the model for random matrices W_k and defines relevant quantities needed in the sequel. Section 2.3 states and proves the result on the exact large deviation rate $\mathcal{J}$ of consensus. Section 2.4 formulates a generalized min-cut problem and shows that its solution gives the rate $\mathcal{J}$. In Section 2.5, we detail gossip and link failure averaging models, and we show how to compute the rate $\mathcal{J}$ for each of the studied models.

Notation. We denote by A_{ij} or $[A]_{ij}$ the entry ij of a matrix A. For $N \in \mathbb{N}$, we denote by $\mathbb{S}^N$ the set of stochastic symmetric N by N matrices; by $\mathbb{G}^N$ the set of all undirected graphs on the set of vertices $V = \{1, ..., N\}$; by I_N the identity matrix of size $N \times N$. For a graph $G \in \mathbb{G}^N$ we denote with $\lambda_F(G)$ the Fiedler value of G, i.e., the second smallest eigenvalue of the Laplacian matrix of G; by $A(G)$ the adjacency matrix of G, defined by $[A(G)]_{ij} = 1$ if $\{i, j\}$ belongs to G, and $[A(G)]_{ij} = 0$ otherwise. $U(0, 1)$ denotes the uniform distribution on the interval $[0, 1]$; $\lceil x \rceil$ denotes the smallest integer not less than x. For a finite set S we denote by $\binom{V}{2}$ the set of all two-element subsets of V; by $|S|$ the cardinality of S.

2.2 Problem setup

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, where Ω is the set of outcomes, $\mathcal{F}$ is a sigma algebra on Ω, and $\mathbb{P}$ is a probability measure on $(\Omega, \mathcal{F})$. Let $W_t : \Omega \mapsto \mathbb{S}^N$, $t \geq 1$, be a sequence of maps that are $(\mathcal{F}, \mathcal{B}(\mathbb{R}^{N \times N}) \cap \mathbb{S}^N)$-measurable, that is, for any $B \in \mathcal{B}(\mathbb{R}^{N \times N}) \cap \mathbb{S}^N$, $\{W_t \in B\}$ belongs to $\mathcal{F}$, for all $t \geq 1$. In other words, $\{W_t\}_{t \geq 1}$ is a sequence of random matrices on $(\Omega, \mathcal{F}, \mathbb{P})$.

Assumption 2.1

1. Random matrices W_t, $t \geq 1$, are independent and identically distributed (i.i.d.);
2. Diagonal entries of W_t are almost surely positive, i.e., for each t, almost surely $[W_t]_{ii} > 0$ for all $i = 1, ..., N$.

Let $\Phi(s, t)$ denote the product of the matrices that occur from time $t + 1$ until time s, $1 \leq t < s$, $\Phi(s, t) = W_s \cdots W_{t+1}$. Also, let $\widetilde{\Phi}(s, t) = \Phi(s, t) - J$; we call $\widetilde{\Phi}(s, t)$ the error matrix from time t until time s.

To analyze the products $\Phi(s,t)$, we introduce the induced graph operator $G : \mathbb{S}^N \mapsto \mathbb{G}^N$. For $W \in \mathbb{S}^N$, we define $G(W)$ by

$$G(W) = \left(V, \left\{\{i,j\} \in \binom{V}{2} : W_{ij} > 0\right\}\right). \tag{2.3}$$

Thus, for a given matrix W, $G(W)$ is a graph on N nodes, without self-loops, with edges between those nodes i and j for which the entries W_{ij}'s are positive. As W is symmetric, $G(W)$ is undirected.

Sequence of induced graphs. Using the induced graph operator, from the sequence $\{W_t\}_{t\geq 1}$, we derive the sequence of random graphs $\{G_t\}_{t\geq 1}$ by assigning $G_t = G(W_t)$, for $t = 1, 2, \ldots$. More precisely, $G_t : \Omega \mapsto \mathbb{G}^N$, for $t \geq 1$, is defined by $G_t(\omega) = G(W_t(\omega))$, for any $\omega \in \Omega$. Intuitively, the graph G_t underlying W_t at some time $t \geq 0$ is the collection of all communication links that are active at time t. Note that, for any $t \geq 1$, G_t is $(\mathcal{F}, 2^{\mathbb{G}^N})$-measurable, that is, for any $\mathcal{H} \subseteq \mathbb{G}^N$, the event $\{G_t \in \mathcal{H}\}$ belongs to $\mathcal{F}$.

As the random matrices W_t, $t \geq 1$ are independent, it follows by the disjoint block theorem [69] that the random graphs G_t, $t \geq 1$ are also independent. Furthermore, as W_t are identically distributed, it follows that, for any $H \in \mathbb{G}^N$, the probability $\mathbb{P}(G_t = H) = \mathbb{P}(G(W_t) = H)$ is the same at all times. Thus, the sequence $\{G_t\}_{t\geq 1}$ is i.i.d., and each G_t is distributed according to the same probability mass function p_H, $H \in \mathbb{G}^N$, where

$$p_H = \mathbb{P}\left(G(W_t) = H\right).$$

Further, for a collection $\mathcal{H} \subseteq \mathbb{G}^N$, let $p_{\mathcal{H}}$ denote the probability that the induced graph of W_t belongs to $\mathcal{H}$, that is, $p_{\mathcal{H}} = \mathbb{P}(G_t \in \mathcal{H})$. Then, $p_{\mathcal{H}} = \sum_{H\in\mathcal{H}} p_H$. Finally, we collect in the set $\mathcal{G}$ all the graphs that occur with positive probability:

$$\mathcal{G} := \left\{H \in \mathbb{G}^N : p_H > 0\right\}, \tag{2.4}$$

and we call $\mathcal{G}$ the set of realizable graphs. For example, if H contains a link $\{i,j\}$ such that $\mathbb{P}([W_t]_{ij} > 0) = 0$, then $H \notin \mathcal{G}$; similarly, if for some $\{i,j\}$ $\mathbb{P}([W_t]_{ij} > 0) = 1$, then all realizable graphs must contain this link. The complete graph $\mathcal{G} = \left(V, \binom{V}{2}\right)$ is obtained whenever W_t has a joint probability density function that is strictly positive on $\mathbb{S}^N$. We next give examples of sequences of random matrices that satisfy Assumption 2.1 and, for each of the examples, we derive the set of realizable graphs and compute the distribution of the corresponding induced graphs.

Example 2.2 (Gossip with uniformly distributed weights) Let $\widehat{G} = (V, \widehat{E})$ be an arbitrary connected graph on N vertices. At each time $t \geq 1$ a node in V is chosen independently from the previous choices and according to the probability mass function $r_u > 0$, $u \in V$, $\sum_{u\in V} r_u = 1$. The chosen node then randomly chooses a neighbor in $\widehat{G}$ according to the probability mass function $q_{uv} > 0$, $\{u,v\} \in \widehat{E}$, $\sum_{\{u,v\}\in\widehat{E}} q_{uv} = 1$,

$u \in V$. Denote the node chosen at time t and its chosen neighbor by u_t and v_t, respectively. With gossip with uniformly distributed weights, averaging occurs only at the edge that is active at time t, $\{u_t, v_t\}$, and with weight equal to the realization of a uniformly distributed parameter $\alpha_t \sim U[0,1]$. Correspondingly, the weight matrix at time t is $W_t = I_N - \alpha_t(e_{u_t} - e_{v_t})(e_{u_t} - e_{v_t})^\top$. We assume that α_t, $t \geq 1$, are independent random variables, and, also, that α_t is independent of u_s, v_s, for all s, t, implying that the sequence W_t, $t \geq 1$ is i.i.d. Also, since $\alpha_t = 1$ with probability zero, diagonal entries of W_t are almost surely positive, and we conclude that the sequence of random matrices $\{W_t\}_{t\geq 1}$ constructed in this way satisfies Assumption 2.1.

By construction, every realization of W_t is of the form $I_N - \alpha(e_u - e_v)(e_u - e_v)^\top$, for some $\alpha \in [0,1]$ and $u, v \in V$ such that $\{u, v\} \in \widehat{E}$. Thus, every realization of G_t is of the form: 1) $(V, \emptyset)$, when $\alpha_t = 0$; or 2) $(V, \{u, v\})$, for $\{u, v\} \in \widehat{E}$. Since $\alpha_t = 0$ with probability 0, we have that $p_{(V,\emptyset)} = 0$, and, so, the potential candidates for realizable graphs are only graphs from the second category. Now, for $\{u, v\} \in \widehat{E}$, $p_{(V,\{u,v\})} = \mathbb{P}(\alpha_t > 0, u_t = u \text{ and } v_t = v \text{ or } u_t = v \text{ and } v_t = u)$. Since α_t is independent of u_t and v_t, it follows that $p_{(V,\{u,v\})} = r_u q_{uv} + r_v q_{vu} > 0$, showing that $(V, \{u, v\})$ is a realizable graph. Summing up, the set of realizable graphs for gossip with uniformly distributed weights running on $\widehat{G}$ is the set of all one-link subgraphs of $\widehat{G}$

$$\mathcal{G}^{\text{Gossip}}(\widehat{G}) = \left\{(V, \{u, v\}) : \{u, v\} \in \widehat{E}\right\}. \tag{2.5}$$

We remark that the same conclusions would be obtained if the uniform distribution, which generates α_t, was replaced by an arbitrary distribution $\mu : \mathcal{B}([0,1]) \mapsto [0,1]$ satisfying $\mu((0,1)) = 1$.

Example 2.3 (Link failure model with Metropolis weights) Consider a connected network defined by $\widehat{G} = (V, \widehat{E})$. We assume that, at any time $t \geq 1$, only edges in $\widehat{E}$ can occur, and, also, that occurrence of edge $e \in \widehat{E}$ at time t is modeled as a Bernoulli random variable $Z_{e,t} \sim \text{Ber}(p_e)$, for $e \in \widehat{E}$, where $p_e \in (0,1)$. We assume that occurrences of edges are independent across space and in time. For $t \geq 1$ and $i = 1, .., N$, let $d_{i,t} = \sum_{j:\,\{i,j\}\in\widehat{E}} Z_{\{i,j\},t}$, that is, $d_{i,t}$ is the degree of node i at time t. The weight matrix at time t is chosen by $[W_t]_{ij} = \frac{1}{1+\max\{d_{i,t}, d_{j,t}\}}$, for all $\{i, j\} \in E_t$, $[W_t]_{ii} = 1 - \sum_{j=1}^N [W_t]_{ij}$, $i = 1, .., N$ and $[W_t]_{ij} = 0$, otherwise. It can be easily shown that, for every realization of $Z_{e,t}$, $e \in \widehat{E}$, diagonal entries of W_t are positive. Further, since $\{Z_{e,t}\}_{e\in E}$ are independent (in time), and for any $e \in E$, $Z_{e,t}$ for $t \geq 1$ are identically distributed, it follows that random matrices W_t are i.i.d. Thus, the sequence $\{W_t\}_{t\geq 1}$ satisfies Assumption 2.1.

For each time t, let E_t collect all the edges that are online at time t, $E_t = \{e : Z_{e,t} = 1\}$. Then, by construction of W_t, $G_t = (V, E_t)$, for all t. Using this fact, for any $H = (V, E) \in \mathbb{G}^N$ such that $E \subseteq \widehat{E}$, we get $p_H = \mathbb{P}(Z_{e,t} = 1,\ e \in E \text{ and } Z_{e,t} = 0,\ e \notin E)$, which by the independence assumption yields

$p_H = \prod_{e\in E} p_e \prod_{f\notin E}(1-p_f) > 0$. We conclude that the set of realizable graphs for the link failure model on $\widehat{G}$ is the set of all subgraphs of $\widehat{G}$:

$$\mathcal{G}^{\text{Link fail.}}(\widehat{G}) = \left\{(V,E) : E \subseteq \widehat{E}\right\}. \tag{2.6}$$

Accumulation graph and disconnected collections. For a collection of graphs $\mathcal{H} \subseteq \mathbb{G}^N$, we denote by $\Gamma(\mathcal{H})$ the union graph which contains all the edges of all the graphs in $\mathcal{H}$:

$$\Gamma(\mathcal{H}) := (V, \bigcup_{G\in\mathcal{H}} E(G)), \tag{2.7}$$

where $E(G)$ denotes the set of edges of a graph G.

Specifically, for any $1 \le t < s$, we denote by $\Gamma(s,t)$[2] the random graph that collects the edges from all the graphs G_r that appeared from time $r = t+1$ to $r = s$, $s > t$, i.e.,

$$\Gamma(s,t) := \Gamma(\{G_s, G_{s-1}, \ldots, G_{t+1}\}),$$

and we call $\Gamma(s,t)$ the accumulation graph from time t until time s.

We next define collections of realizable graphs of certain types that will be important in computing the rate in (2.2) and (2.1).

Definition 2.4 The collection $\mathcal{H} \subseteq \mathcal{G}$ is a disconnected collection on $\mathcal{G}$ if its accumulation graph $\Gamma(\mathcal{H})$ is disconnected.

Thus, a disconnected collection is any collection of realizable graphs such that the union of all of its graphs yields a disconnected graph. We also define the set of all possible disconnected collections on $\mathcal{G}$:

$$\Pi(\mathcal{G}) = \{\mathcal{H} \subseteq \mathcal{G} : \mathcal{H} \text{ is a disconnected collection on } \mathcal{G}\}. \tag{2.8}$$

Example 2.5 (Gossip model) Consider the gossip algorithm from Example 2.2 when $\widehat{G}$ is the complete graph on N vertices. In this case $\mathcal{G} = \left\{(V,\{i,j\}) : \{i,j\} \in \binom{V}{2}\right\}$, that is, $\mathcal{G}$ is the set of all possible one-link graphs on N vertices. An example of a disconnected collection of $\mathcal{G}$ is $\mathcal{G} \setminus \{(V,\{i,j\}) : j = 1,\ldots N\}$, where i is a fixed vertex, or, in words, the collection of all one-link graphs except of those whose link is adja-

[2]Graph $\Gamma(s,t)$ is associated with the matrix product $W_s \cdots W_{t+1}$ going from time $t+1$ until time $s > t$. The notation $\Gamma(s,t)$ indicates that the product is backwards; see also the definition of the product matrix $\Phi(s,t)$ after Assumption 2.1 at the beginning of this section.

cent to i. Another example is $\mathcal{G} \setminus (\{(V, \{i,k\}) : k = 1, \dots N, k \neq j\} \cup \{(V, \{j,l\}) : l = 1, \dots N, l \neq i\})$, where $\{i, j\}$ is a fixed link.

Example 2.6 (Toy example) Suppose that, for some sequence of random matrices taking values in $\mathbb{S}^5$, the set of realizable graphs is $\mathcal{G} = \{G_1, G_2, G_3\}$, where graphs G_i, $i = 1, 2, 3$ are given in Figure 2.1. In this model each realizable graph is a two-link graph and the supergraph of all the realizable graphs $\Gamma(\{G_1, G_2, G_3\})$ is connected. If we scan over the supergraphs $\Gamma(\mathcal{H})$ of all subsets $\mathcal{H}$ of $\mathcal{G}$, we see that $\Gamma(\{G_1, G_2\})$,

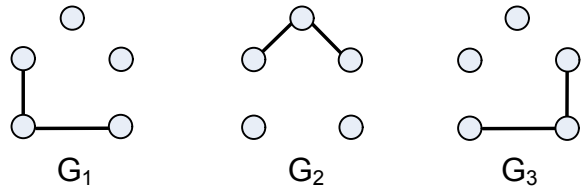

Figure 2.1: Example of a five node network with three possible graph realizations, each realization being a two-link graph

$\Gamma(\{G_2, G_3\})$ and $\Gamma(\{G_1, G_2, G_3\})$ are connected, whereas $\Gamma(\{G_1, G_3\})$ and $\Gamma(G_i) = G_i$, $i = 1, 2, 3$, are disconnected. It follows that $\Pi(\mathcal{G}) = \{\{G_1\}, \{G_2\}, \{G_3\}, \{G_1, G_3\}\}$.

2.3 Convergence in probability - exponential rate

This Section states and proves the main result of this chapter, Theorem 2.7. We prove Theorem 2.7 by proving the corresponding large deviation upper and lower bound; the proof of the lower bound is given in Subsection 2.3.1, whereas the proof of the upper bound is given in Subsection 2.3.2.

Theorem 2.7 Let d_k be a sequence of real numbers such that $d_k \in (0, 1]$ and $\log d_k = o(k)$. Then:

$$\lim_{k\to\infty} \frac{1}{k} \log \mathbb{P}\left(\left\| \widetilde{\Phi}(k, 0) \right\| \geq d_k \right) = -\mathcal{J},$$

where

$$\mathcal{J} = \begin{cases} +\infty & \text{if } \Pi(\mathcal{G}) = \emptyset \\ |\log p_{\max}| & \text{otherwise} \end{cases}, \tag{2.9}$$

and $p_{\max} = \max_{\mathcal{H} \in \Pi(\mathcal{G})} p_{\mathcal{H}}$ is the probability of the most likely disconnected collection.

We prove Theorem 2.7, by proving separately the lower bound (2.10) and the upper bound (2.11)[3]:

$$\liminf_{k\to\infty} \frac{1}{k} \log \mathbb{P}\left(\left\| \widetilde{\Phi}(k,0) \right\| \geq d_k \right) \quad \geq \quad -\mathcal{J} \tag{2.10}$$

$$\limsup_{k\to\infty} \frac{1}{k} \log \mathbb{P}\left(\left\| \widetilde{\Phi}(k,0) \right\| \geq d_k \right) \quad \leq \quad -\mathcal{J}. \tag{2.11}$$

Subsection 2.3.1 proves the lower bound (2.10), and Subsection 2.3.2 proves the upper bound (2.11).

2.3.1 Proof of the lower bound (2.10)

We first show that, for any $k \geq 1$, a sufficient condition for the norm $\left\| \widetilde{\Phi}(k,0) \right\|$ being above d_k is that the supergraph $\Gamma(k,0)$ is disconnected. In fact, we prove the following stronger claim.

Lemma 2.8 For any fixed $\omega \in \Omega$ and any $k \geq 1$

$$\Gamma(k,0) \text{ is disconnected } \Rightarrow \left\| \widetilde{\Phi}(k,0) \right\| = 1.$$

Proof Fix $\omega \in \Omega$ and $k \geq 1$ and suppose that $\Gamma(k,0)$ is not connected. Suppose further (without loss of generality) that $\Gamma(k,0)$ has exactly two components and denote them by C_1 and C_2. Then, for all i, j such that $i \in C_1$ and $j \in C_2$, we have $\{i,j\} \notin \Gamma(k,0)$, and, consequently, $\{i,j\} \notin G_t$, for all $1 \leq t \leq k$. By definition of G_t, this implies that the corresponding entries in the matrices W_t, $1 \leq t \leq k$, are equal to zero, i.e.,

$$\forall t,\, 1 \leq t \leq k : [W_t]_{ij} = 0, \forall \{i,j\} \text{ s.t. } i \in C_1,\, j \in C_2.$$

Thus, every matrix realization W_t from time 1 to time k has a block diagonal form (up to a symmetric permutation of rows and columns, the same for all W_t)

$$W_t = \begin{bmatrix} [W_t]_{C_1} & 0 \\ 0 & [W_t]_{C_2} \end{bmatrix},$$

where $[W_t]_{C_1}$ is the block of W_t corresponding to the nodes in C_1, and similarly for $[W_t]_{C_2}$. This implies that $\Phi(k,0)$ has the same block diagonal form, which, in turn, proves that $\left\| \widetilde{\Phi}(k,0) \right\| = 1$. □

[3]Note that we need to prove the lower bound (2.10) only for the case when $\Pi(\mathcal{G}) \neq \emptyset$, as the bound trivially holds when $\mathcal{J} = +\infty$.

Using the result of Lemma 2.8, we get:

$$\mathbb{P}\left(\left\|\widetilde{\Phi}(k,0)\right\| \geq d_k\right) \geq \mathbb{P}\left(\left\|\widetilde{\Phi}(k,0)\right\| = 1\right) \geq \mathbb{P}\left(\Gamma(k,0) \text{ is disconnected}\right). \tag{2.12}$$

We now focus on computing the probability of the event that $\Gamma(k,0)$ is disconnected. For any fixed $k \geq 1$, a sufficient condition that guarantees that $\Gamma(k,0)$ is disconnected is that every graph realization G_t from time 1 to time k is drawn from some disconnected collection $\mathcal{H} \in \Pi(\mathcal{G})$. More precisely, for every $\mathcal{H} \in \Pi(\mathcal{G})$ and every $k \geq 1$, it holds for all $\omega \in \Omega$:

$$G_t \in \mathcal{H}, \text{ for } 1 \leq t \leq k \;\Rightarrow\; \Gamma(k,0) \text{ is disconnected.} \tag{2.13}$$

This can be easily shown by observing that if $\{G_1, ..., G_k\} \subseteq \mathcal{H}$, then $\Gamma(k,0) = \Gamma(\{G_1, ..., G_k\})$ is a subgraph of $\Gamma(\mathcal{H})$, or, in other words, $\Gamma(k,0)$ cannot contain any additional edge beyond the ones in $\Gamma(\mathcal{H})$. Now, since $\Gamma(\mathcal{H})$ is disconnected, it must be that $\Gamma(k,0)$ is disconnected as well. Claim in (2.13) implies that for every $\mathcal{H} \in \Pi(\mathcal{G})$ and every $k \geq 1$

$$\mathbb{P}\left(\Gamma(k,0) \text{ is disconnected}\right) \geq \mathbb{P}(G_t \in \mathcal{H}, \text{ for } 1 \leq t \leq k) = p_{\mathcal{H}}^k, \tag{2.14}$$

where the last equality follows by the time independence assumption. Combining the previous bound with eq. (2.12) and optimizing the bound over $\mathcal{H} \in \Pi(\mathcal{G})$ yields

$$\mathbb{P}\left(\left\|\widetilde{\Phi}(k,0)\right\| \geq d_k\right) \geq p_{\max}^k.$$

Finally, taking the $\log$, dividing by k, and taking the $\liminf$ over $k \to \infty$, the lower bound (2.10) follows.

2.3.2 Proof of the upper bound (2.11)

To prove the upper bound, we first extend the concept of the induced graph operator to the δ-induced graph operator which accounts only for those entries that are above some given $\delta > 0$, i.e., the entries that are sufficiently important. Using the definition of the δ-induced graph, we correspondingly extend the concepts of the accumulation graph, the set of realizable graphs and the most likely disconnected collection. We explain this next.

The family of δ-induced graph sequences.

Definition 2.9 For each $\delta > 0$ we define the δ-induced graph operator $G_\delta : \mathbb{S}^N \mapsto \mathbb{G}^N \cup \{\mathsf{E}\}$ by

$$G_\delta(W) = \begin{cases} \left(V, \left\{\{i,j\} \in \binom{V}{2} : W_{ij} \geq \delta\right\}\right), & \text{if } W_{ii} \geq \delta, \ \forall\, i \\ \mathsf{E}, & \text{otherwise} \end{cases}. \tag{2.15}$$

As we can see from the definition, if a matrix has all diagonal entries above δ, then its δ-induced graph contains all the edges whose corresponding entries of the matrix are above δ. On the other hand, any matrix that has a diagonal entry below δ gets mapped by G_δ to the symbol E; note that, by doing this, we discard all the potential edges for such a matrix (no matter how large their corresponding entries are). Intuitively, δ-induced graph operator G_δ, compared to G, acts as an edge truncator that cuts off all the non-significant edges and, also, it discards all the matrices with low diagonals by mapping them to E. We will see later in the analysis that, whenever at some point in the sequence W_t, $t \geq 1$, a matrix with a small diagonal entry occurs, we cannot say much about the continuity of the "information flow" at that point. Thus, we introduce a special symbol, E, to indicate such matrices that cut (or "erase") the information flow.

We now use operators G_δ, $\delta > 0$, to construct from $\{W_t\}_{t\geq 1}$ new induced graph sequences. For every $\delta > 0$, $t \geq 1$, let $G_{t,\delta} : \Omega \mapsto \mathbb{G}^N \cup \{\mathsf{E}\}$ be defined by $G_{t,\delta}(\omega) = G_\delta(W_t(\omega))$, for $\omega \in \Omega$. Thus, for every δ, $G_{t,\delta}$ is the δ-induced graph of the matrix W_t, $t \geq 1$. Remark that, in contrast with the regular induced graph G_t, $G_{t,\delta}$ can take value E.

Each sequence $\{G_{t,\delta}\}_{t\geq 1}$ from this family indexed by δ is i.i.d., as the sequence $\{W_t\}_{t\geq 1}$ is i.i.d. For any $H \in \mathbb{G}^N$, denote by $p_{H,\delta}$ the probability that $G_{t,\delta}$ is equal to H, i.e., is $p_{H,\delta} = \mathbb{P}(G_{t,\delta} = H)$. The probability that $G_{t,\delta}$ takes value E is denoted by $p_{\mathsf{E},\delta} = \mathbb{P}(G_{t,\delta} = \mathsf{E})$. We show in Lemma 2.10 that, for each t, $G_{t,\delta}$ converges almost surely to G_t as $\delta \to 0$, thus implying the corresponding convergence in distribution. For convenience, we state the result in terms of the adjacency matrices: for any $\omega \in \Omega$, $t \geq 1$ and $\delta > 0$, we define $A_t(\omega) = A(G_t(\omega))$, $A_{t,\delta}(\omega) = A(G_{t,\delta}(\omega))$, if $G_{t,\delta}(\omega) \neq \mathsf{E}$, otherwise, we assign $A_{t,\delta}(\omega)$ to be the N by N matrix of all zeros.

Lemma 2.10 For any $t \geq 1$, almost surely $A_{t,\delta} \to A_t$, as $\delta \to 0$. Hence, for any $H \in \mathbb{G}^N$, $\lim_{\delta\to 0} p_{H,\delta} = p_H$ and also $\lim_{\delta\to 0} p_{\mathsf{E},\delta} = 0$.

Proof For any $t \geq 1$, let $\Omega_t^\star = \{[W_t]_{ii} > 0, \ \forall\, i\}$; note that, by Assumption 2.1, $\mathbb{P}(\Omega_t^\star) = 1$. Now, fix t and $\omega \in \Omega_t^\star$, and consider $W = W_t(\omega)$. Then, $W_{ii} > 0$ for all i and let $\delta_0 = \min_i W_{ii}$ (note that $\delta_0 > 0$ and also that it depends on ω). For all $\delta > \delta_0$, $G_\delta(W) = \mathsf{E}$, whereas for all $\delta \in (0, \delta_0]$, $G_\delta(W) \in \mathbb{G}^N$. Note that, to prove the claim, it is sufficient to consider only the case when $\delta \leq \delta_0$. First, for all $\{i,j\}$ such that $W_{ij} = 0$, we have $[A_t(\omega)]_{ij} = 0$ and also $[A_{t,\delta}(\omega)]_{ij} = 0$ for all $\delta \leq \delta_0$ (in fact, due to the definition

of $A_{t,\delta}$, the latter holds for all δ), showing that $[A_{t,\delta}(\omega)]_{ij}$ converges to $[A_t(\omega)]_{ij}$. On the other hand, let α be the minimum over all positive entries of W, $\alpha = \min_{\{i,j\}:\, W_{ij}>0} W_{ij}$ and note that $\alpha \leq \delta_0$ and $\alpha > 0$. Then, for all $\delta \leq \alpha$, $G_\delta(W)$ and $G(W)$ match, implying that $A_{t,\delta}(\omega) = A_t(\omega)$ for all such δ. As ω was an arbitrary point from $\Omega_t^\star$ and since $\mathbb{P}\left(\Omega_t^\star\right) = 1$, the almost sure convergence follows. The second part of the claim follows from the fact that almost sure convergence implies the convergence in distribution. □

Similarly as with the set of realizable graphs, for each $\delta > 0$, we define the set of δ-realizable graphs

$$\mathcal{G}_\delta = \left\{ H \in \mathbb{G}^N : p_{H,\delta} > 0 \right\}. \tag{2.16}$$

For a collection of graphs $\mathcal{H} \subseteq \mathbb{G}^N$, we denote by $p_{\mathcal{H},\delta}$ the probability that $G_{t,\delta}$ belongs to $\mathcal{H}$, which is equal to $p_{\mathcal{H},\delta} = \sum_{H\in\mathcal{H}} p_{H,\delta}$. Similarly as before, $\Pi(\mathcal{G}_\delta)$ denotes the set of all possible disconnected collections on $\mathcal{G}_\delta$.

For $\delta > 0$ such that $\Pi(\mathcal{G}_\delta) \neq \emptyset$, let $p_{\max,\delta} = \max_{\mathcal{H}\in\Pi(\mathcal{G}_\delta)} p_{\mathcal{H},\delta}$; that is, $p_{\max,\delta}$ is the probability that $G_{t,\delta}$ belongs to the most likely disconnected collection on $\mathcal{G}_\delta$. The following corollary of Lemma 2.10 is one of the main steps in the proof of the upper bound (2.11). We omit the proof of Corollary 2.11 noting that it uses the similar arguments as the ones in the proof of Lemma 2.10.

Corollary 2.11 If $\Pi(\mathcal{G}) \neq \emptyset$, then there must exist $\overline{\delta} > 0$ such that $\Pi(\mathcal{G}_\delta) \neq \emptyset$ for every $0 < \delta \leq \overline{\delta}$. Moreover,

$$\lim_{\delta\to 0} p_{\max,\delta} = p_{\max}.$$

Similarly as with accumulation graph $\Gamma(s,t)$ that collects all the edges of the (regular) induced graphs G_{t+1},...,G_s, for each $\delta > 0$, we define the δ-accumulation graph $\Gamma_\delta(s,t)$ to collect the edges of the δ-induced graphs $G_{t+1,\delta}$,...,$G_{s,\delta}$. In contrast with $\Gamma(s,t)$, here we have to take into account that, for some δ, realizations of the δ-induced graphs might be equal to E. To handle this, for each $\delta > 0$ and $t \geq 1$, we introduce $R_{t,\delta} : \Omega \mapsto \mathbb{N} \cup \{0\}$ which we define by

$$R_{t,\delta}(\omega) = \begin{cases} 0, & \text{if } G_{r,\delta}(\omega) \neq \mathsf{E}, \text{ for all } 1 \leq r \leq t \\ \max\{1 \leq r \leq t : G_{r,\delta}(\omega) = \mathsf{E}\}, & \text{otherwise} \end{cases}. \tag{2.17}$$

Now, for any $1 \leq t < s$ and $\delta > 0$, we define $\Gamma_\delta(s,t)$ to be

$$\Gamma_\delta(s,t) = \begin{cases} \Gamma(\{G_{s,\delta}, ..., G_{t+1,\delta}\}), & \text{if } R_{s,\delta} \leq t \\ \Gamma(\{G_{s,\delta}, ..., G_{R_{s,\delta}+1,\delta}\}), & \text{if } t < R_{s,\delta} < s \\ \mathsf{E}, & \text{if } R_{s,\delta} = s \end{cases} . \tag{2.18}$$

We now explain the intuition behind this construction of $\Gamma_\delta(s,t)$. If $R_{s,\delta} \leq t$, that is, if the interval from t until s is clear from the realization E, then we assign $\Gamma_\delta(s,t)$ to collect all the edges of all the δ-induced graph realizations that occurred from time $t+1$ until time s. If, on the other hand, it happens that, starting from time t we encounter the realization E, i.e., if $G_{r,\delta} = \mathsf{E}$ for some $r > t$, we consider this to be a bad event and we reset the number of collected edges so far to zero (formally, by assigning at time r $\Gamma_\delta(r,t) = \mathsf{E}$). We repeat this until we hit time s. Since the last occurrence of the bad realization E was at time $R_{s,\delta}$, assuming that $R_{s,\delta} < s$, the δ-accumulation graph will contain all the edges of the δ-induced graph realizations that occurred from time $R_{s,\delta}+1$ until time s.

We have seen in the proof of the lower bound in Lemma 2.8 that, if the accumulation graph $\Gamma(k,0)$ is disconnected, then the norm of the error matrix is still equal to 1 at time k. Lemma 2.12 is, in a sense, a converse to this result, as it provides a sufficient condition in terms of $\Gamma_\delta(s,t)$ for the norm of the error matrix to drop on the time interval from t until s.

Lemma 2.12 For any fixed $\omega \in \Omega$, for all $1 \leq t < s$ and $\delta \in (0,1)$ such that $\Gamma_\delta(s,t) \neq \mathsf{E}$, it holds

$$\left\| \widetilde{\Phi}(s,t) \right\|^2 \leq 1 - \lambda_F\left(\Gamma_\delta(s,t)\right) \delta^{2(s-t)}. \tag{2.19}$$

Using the fact that the Fiedler value (algebraic connectivity) of a connected graph is positive [70], if $\Gamma_\delta(s,t)$ is connected (and $\Gamma_\delta(s,t) \neq \mathsf{E}$), then the squared norm of the error matrix on this interval drops for at least $\lambda_F\left(\Gamma_\delta(s,t)\right) \delta^{2(s-t)} > 0$. To get a uniform bound for this drop (that holds for all connected realizations of $\Gamma(s,t)$), we use the Fiedler value of the path graph on N vertices. This is stated next in Corollary 2.13.

Corollary 2.13 For any fixed $\omega \in \Omega$, for all $1 \leq t < s$, $\delta > 0$ such that $\Gamma_\delta(s,t) \neq \mathsf{E}$ and $\Gamma_\delta(s,t)$ is connected

$$\left\| \widetilde{\Phi}(s,t) \right\|^2 \leq 1 - c_N \delta^{2(s-t)}, \tag{2.20}$$

where $c_N = 2(1 - \cos\frac{\pi}{N})$ is the Fiedler value of the path graph on N vertices, i.e., the minimum of $\lambda_{\mathrm{F}}(H) > 0$ over all connected graphs H on N vertices [70].

We next prove Lemma 2.12.

Proof We first prove Lemma 2.12 for all ω, δ, s, t such that $R_{s,\delta}(\omega) \leq t$. To this end, fix $\omega \in \Omega$, $\delta > 0$ and consider a fixed t, s $1 \leq t < s$, for which $R_{s,\delta}(\omega) \leq t$. Similarly to the proof of Lemma 1 a), b) in [71], it can be shown here that: 1) $[\Phi(s,t)(\omega)]_{ii} \geq \delta^{s-t}$, for all i; and 2)$[\Phi(s,t)(\omega)]_{ij} \geq \delta^{s-t}$, for all $\{i,j\} \in E$, where we let E denote the set of edges of the graph $\Gamma_\delta(s,t)(\omega)$.

Notice that $\left\|\widetilde{\Phi}(s,t)\right\|^2$ is the second largest eigenvalue of $\Phi(s,t)^\top \Phi(s,t)$, and, thus can be computed as:

$$\left\|\widetilde{\Phi}(s,t)\right\|^2 \;\; = \;\; \max_{q^\top q = 1,\, q \perp 1} q^\top \Phi(s,t)^\top \Phi(s,t) q$$

Since $\Phi(s,t)^\top \Phi(s,t)$ is a symmetric stochastic matrix, it can be shown, e.g., [15], that its quadratic form, for a fixed vector $q \in \mathbb{R}^N$, can be written as:

$$q^\top \Phi(s,t)^\top \Phi(s,t) q \;\; = \;\; q^\top q - \sum_{\{i,j\}} \left[\Phi(s,t)^\top \Phi(s,t)\right]_{ij} (q_i - q_j)^2 \tag{2.21}$$

Now, combining the two auxiliary inequalities from the beginning of the proof, we get that, for all $\{i,j\} \in E$, $[\Phi(s,t)^\top \Phi(s,t)]_{ij} \geq \delta^{2(s-t)}$, where, we recall, E is the set of edges of $\Gamma_\delta(s,t)(\omega)$. Further, since all the entries of $\Phi(s,t)$ are non-negative (for every t, every realization of W_t is a stochastic matrix, and thus has non-negative entries), we can upper bound the sum in (2.21) over all $\{i,j\}$ by the sum over $\{i,j\} \in E$ only, yielding:

$$q^\top \Phi(s,t)^\top \Phi(s,t) q \leq q^\top q - \delta^{2(s-t)} \sum_{\{i,j\} \in E} (q_i - q_j)^2 \,. \tag{2.22}$$

Finally, $\min_{q^\top q = 1,\, q \perp 1} \sum_{\{i,j\} \in E} (q_i - q_j)^2$ is equal to the Fiedler value (i.e., the second smallest eigenvalue of the Laplacian) of the graph $\Gamma_\delta(s,t)(\omega)$. This completes the proof of Lemma 2.12 for the case when ω, δ, s, t are such that $R_{s,\delta} \leq t$. The claim of Lemma 2.12 for the case when ω, δ, s, t are such that $t < R_{s,\delta}(\omega) < s$ essentially follows from the submultiplicativity of the spectral norm, the result of Lemma 2.12 for the case that we just proved (with $t' = R_{s,\delta}(\omega)$), and the fact that $\Gamma_\delta(s, R_{s,\delta}(\omega))(\omega) = \Gamma_\delta(s,t)(\omega)$. □ Lemma 2.12 and Corollary 2.13 say that, for each fixed $\delta > 0$, whenever there is an interval in which the δ-accumulation graph is connected, then the norm of the error matrix on this interval improves by some finite amount (dependent on the interval size). We next introduce, for each $\delta > 0$, the sequence of δ-stopping times that registers these times at which we are certain that the error matrix makes an improvement.

Family of the sequences of δ-stopping times. For each $\delta > 0$, we define the sequence of δ-stopping times

$T_{i,\delta} : \Omega \mapsto \mathbb{N} \cup \{+\infty\}$, $i = 1, 2, \ldots$ by:

$$T_{i,\delta}(\omega) = \min\{t \geq T_{i-1,\delta}(\omega) + 1 : \ \Gamma_\delta(t, T_{i-1}(\omega)) \text{ is connected}\}, \text{ for } i \geq 1, \tag{2.23}$$
$$T_{0,\delta} \equiv 0.$$

By its construction, the sequence $\{T_{i,\delta}\}_{i\geq 1}$ defines the times that mark the right end point of "clear" intervals, without realization of δ-induced graphs equal to E, on which Γ_δ is connected. Using the result of Lemma 2.12, we have that at times $T_{i,\delta}$ the norm of $\widetilde{\Phi}$ drops below 1 and the averaging process makes an improvement. Let further, for each $\delta > 0$ and $k \geq 1$, $M_{k,\delta} : \Omega \mapsto \mathbb{N} \cup \{0\}$ count the number of improvements with respect to the δ-stopping times until time k:

$$M_{k,\delta}(\omega) = \max\{i \geq 0 : T_{i,\delta}(\omega) \leq k\}. \tag{2.24}$$

We now explain how, at any given time k, we can use the knowledge of $M_{k,\delta}$ to bound the norm of the "error" matrix $\widetilde{\Phi}(k, 0)$. Suppose that $M_{k,\delta} = m$. If we knew the locations of all the improvements until time k, $T_{i,\delta} = t_i$, $i = 1, \ldots, m$ then, using Lemma 2.12, we could bound the norm of $\widetilde{\Phi}(k, 0)$. Intuitively, since for fixed k and fixed m the number of allocations of $T_{i,\delta}$'s is finite, there will exist the one which yields the worst bound on $\left\|\widetilde{\Phi}(k, 0)\right\|$. It turns out that the worst case allocation is the one with equidistant improvements, thus allowing for deriving a bound on $\left\|\widetilde{\Phi}(k, 0)\right\|$ only in terms of $M_{k,\delta}$. This result is given in Lemma 2.14.

Lemma 2.14 For any fixed $\omega \in \Omega$, $\delta > 0$ and $k \geq 1$:

$$\left\|\widetilde{\Phi}(k, 0)\right\| \leq \left(1 - c_N \delta^{2\frac{k}{M_{k,\delta}}}\right)^{\frac{M_{k,\delta}}{2}}. \tag{2.25}$$

Proof Fix $\omega \in \Omega$, $\delta > 0$, $k \geq 1$. If $M_{k,\delta}(\omega) = 0$, then the claim holds trivially. Thus, suppose $M_{k,\delta}(\omega) = m \geq 1$, and, suppose further $T_{1,\delta}(\omega) = t_1$, $T_{2,\delta}(\omega) = t_2$, $\ldots$, $T_{m,\delta}(\omega) = t_m \leq k$ ($T_{i,\delta}(\omega) > k$, for $i > m$, because $M_{k,\delta}(\omega) = m$). By the construction of the δ-stopping times, we know that $\Gamma_\delta(t_i, t_{i-1})(\omega)$ is connected for all $i = 1, ..., m$. Thus, we apply Lemma 2.12 on the intervals from t_{i-1} until t_i, for $i = 1, \ldots, m$, to get $\left\|\widetilde{\Phi}(t_i, t_{i-1})\right\| \leq \left(1 - c_N \delta^{2(t_i - t_{i-1})}\right)^{\frac{1}{2}}$. Combining this with

submultiplicativity of the spectral norm, yields:

$$\begin{aligned}\left\|\widetilde{\Phi}(k,0)(\omega)\right\| &= \left\|\widetilde{\Phi}(k,t_m)(\omega)\widetilde{\Phi}(t_m,t_{m-1})(\omega)\cdots\widetilde{\Phi}(t_1,0)(\omega)\right\| \\ &\leq \left\|\widetilde{\Phi}(k,t_m)(\omega)\right\|\left\|\widetilde{\Phi}(t_m,t_{m-1})(\omega)\right\|\cdots\left\|\widetilde{\Phi}(t_1,0)(\omega)\right\| \\ &\leq \prod_{i=1}^{m}\left(1-c_N\,\delta^{2(t_i-t_{i-1})}\right)^{\frac{1}{2}}. \end{aligned} \tag{2.26}$$

Denote $\Delta_i = t_i - t_{i-1}$ and note that $\sum_{i=1}^m \Delta_i \leq k$. Further, remark that $f(\Delta) = \log\left(1 - c_N\delta^{2\Delta}\right)$ is a concave function. Taking the $\log$ in (2.26) and applying Jensen's inequality [72] for equal convex multipliers $\alpha_i = \frac{1}{m}$, $i = 1, ..., m$, yields

$$\sum_{i=1}^{m}\alpha_i \log\left(1-c_N\,\delta^{2\Delta_i}\right) \leq \log\left(1-c_N\,\delta^{2(\sum_{i=1}^m \alpha_i\Delta_i)}\right) = \log\left(1-c_N\,\delta^{\frac{2}{m}\sum_{i=1}^m\Delta_i}\right).$$

Finally, since f is increasing and $\sum_{i=1}^m \Delta_i \leq k$, $\sum_{i=1}^m \frac{1}{m}\log\left(1-c_N\,\delta^{2(\Delta_i)}\right) \leq \log\left(1-c_N\,\delta^{\frac{2k}{m}}\right)$. Multiplying both sides of the last inequality with $\frac{m}{2}$, and computing the exponent yields (2.25). □

Lemma 2.14 provides a bound on the norm of the "error" matrix $\widetilde{\Phi}(k,0)$ in terms of the number of improvements $M_{k,\delta}$ up to time k. Intuitively, if $M_{k,\delta}$ is high enough relative to k, then the norm of $\widetilde{\Phi}(k,0)$ decays exponentially fast (to see this, just take $M_{k,\delta} = k$ in eq. (2.25)) and, thus, it cannot stay above d_k, which decays sub-exponentially as k increases. We show that this is indeed true for all $\omega \in \Omega$ for which $M_{k,\delta} = \alpha k$ or higher, for any choice of $\alpha \in (0,1]$; this result is stated in Lemma 2.15, part 1. On the other hand, if the number of improvements is less than αk, then there must have been long intervals on which Γ_δ was disconnected. The probability that such an interval of length t occurs is essentially determined by the probability that the sequence of δ-induced graphs is "trapped" in some disconnected collection for time $t-1$, and it equals $p_{\max,\delta}^{t-1}$. As the number of these intervals until time k is at most αk, this yields, in a crude approximation, the probability of $p_{\max,\delta}^{k-\alpha k}$ for the event $M_{k,\delta} \leq \alpha k$; this intuition is formalized in part 2 of Lemma 2.15.

Lemma 2.15 For any fixed $\delta \in (0,1)$, $\alpha \in (0,1]$:

1. there exists sufficiently large $k_0 = k_0(\delta, \alpha, \{d_k\})$ such that

$$\mathbb{P}\left(\left\|\widetilde{\Phi}(k,0)\right\| \geq d_k,\ M_{k,\delta} \geq \alpha k\right) = 0, \quad \forall k \geq k_0(\delta,\alpha,\{d_k\}); \tag{2.27}$$

2. for every $\lambda \in (0, \overline{\mathcal{J}}_\delta)$

$$\mathbb{P}\left(M_{k,\delta} < \alpha k\right) \leq \exp(-\lambda(k - \lceil \alpha k \rceil))(1 - a_\delta(\lambda))^{-\lceil \alpha k \rceil}, \tag{2.28}$$

where $a_\delta(\lambda) = \exp(\lambda - \overline{\mathcal{J}}_\delta) < 1$ and $\overline{\mathcal{J}}_\delta$ is defined as $\overline{\mathcal{J}}_\delta = |\log(p_{\max,\delta} + |\Pi(\mathcal{G}_\delta)|p_{\mathsf{E},\delta})|$, for δ such that $\Pi(\mathcal{G}_\delta) \neq \emptyset$, and $\overline{\mathcal{J}}_\delta = |\log p_{\mathsf{E},\delta}|$, otherwise.

Proof Fix $\delta \in (0,1)$, $\alpha \in (0,1]$. To prove 1, we first note that, by Lemma 2.14 we have:

$$\left\{\left\|\widetilde{\Phi}(k,0)\right\| \geq d_k\right\} \subseteq \left\{\left(1 - c_N \delta^{2\frac{k}{M_{k,\delta}}}\right)^{\frac{M_{k,\delta}}{2}} \geq d_k\right\}. \tag{2.29}$$

This gives

$$\begin{aligned}
&\mathbb{P}\left(\left\|\widetilde{\Phi}(k,0)\right\| \geq d_k,\ M_{k,\delta} \geq \alpha k\right) \leq \mathbb{P}\left(\left(1 - c_N \delta^{2\frac{k}{M_{k,\delta}}}\right)^{\frac{M_{k,\delta}}{2}} \geq d_k,\ M_{k,\delta} \geq \alpha k\right) \\
&= \sum_{m=\lceil \alpha k \rceil}^{k} \mathbb{P}\left(\left(1 - c_N \delta^{2\frac{k}{M_{k,\delta}}}\right)^{\frac{M_{k,\delta}}{2}} \geq d_k,\ M_{k,\delta} = m\right) = \sum_{m=\lceil \alpha k \rceil}^{k} \mathbb{P}\left(g(k, M_{k,\delta}) \geq \frac{\log d_k}{k},\ M_{k,\delta} = m\right),
\end{aligned} \tag{2.30}$$

where $g(k,m) := \frac{m}{2k} \log\left(1 - c_N \delta^{2\frac{k}{m}}\right)$, for $m > 0$. For fixed k, each of the probabilities in the sum above is equal to 0 for those m such that $g(k,m) < -\frac{\log d_k}{k}$. This yields:

$$\sum_{m=\lceil \alpha k \rceil}^{k} \mathbb{P}\left(g(k, M_{k,\delta}) \geq \frac{\log d_k}{k},\ M_{k,\delta} = m\right) \leq \sum_{m=\lceil \alpha k \rceil}^{k} s(k,m), \tag{2.31}$$

where $s(k,m)$ is the switch function defined by:

$$s(k,m) := \begin{cases} 0, & \text{if } g(k,m) < \frac{\log d_k}{k} \\ 1, & \text{otherwise} \end{cases}$$

Also, as $g(k,\cdot)$ is, for fixed k, decreasing in m, it follows that $s(k,m) \leq s(k,\alpha k)$ for $m \geq \alpha k$. Combining this with eqs. (2.30) and (2.31), we get:

$$\mathbb{P}\left(\left\|\widetilde{\Phi}(k,0)\right\| \geq d_k,\ M_{k,\delta} \geq \alpha k\right) \leq (k - \lceil \alpha k \rceil + 1) s(k, \alpha k).$$

We now show that $s(k, \alpha k)$ will eventually become 0, as k increases, which would yield part 1 of Lemma 2.15. To show this, we observe that g has a constant negative value at $(k, \alpha k)$:

$$g(k, \alpha k) = \frac{\alpha}{2} \log \left(1 - c_N \delta^{\frac{2}{\alpha}}\right).$$

Since $\frac{\log d_k}{k} \to 0$, as $k \to \infty$, there exists $k_0 = k_0(\delta, \alpha, \{d_k\})$ such that $g(k, \alpha k) < \frac{\log d_k}{k}$, for every $k \geq k_0$. Thus, $s(k, \alpha k) = 0$ for every $k \geq k_0$. This completes the proof of part 1.

To prove part 2, we first prove the following result which is the main argument in the proof of part 2.

Lemma 2.16 For any $\delta > 0$, $t \geq 1$

$$\mathbb{P}(T_{1,\delta} > t) \leq e^{-\overline{\mathcal{J}}_\delta t}. \tag{2.32}$$

Proof Fix $\delta > 0$, $t \geq 1$. For the case when $\Pi(\mathcal{G}_\delta) = \emptyset$ the claim easily follows by noting that $\mathbb{P}\left(T_{1,\delta} > t\right) = \mathbb{P}\left(G_{r,\delta} = \mathsf{E},\ 1 \leq r \leq t\right)$. (The latter is true because each realization of G_δ which has a positive probability of occurrence is either a connected graph or equal to E.) Suppose now that δ is such that $\Pi(\mathcal{G}_\delta) \neq \emptyset$. Define S_l, $l \geq 1$ to be the (random) locations of the realization E in the sequence of δ-induced graphs and let also Q_t be the number of such realizations until time t; for convenience, let also $S_0 \equiv 0$. By definition of Γ_δ, the event $\{T_{1,\delta} > t\}$ is equivalent to the event that Γ_δ is disconnected on each block in the sequence of $G_{r,\delta}$, $1 \leq r \leq t$ that is clear from realizations of E. Partitioning this event over all possible number of realizations of E on the interval from time 1 until time t, Q_t, and, also, over all possible locations of E, S_l, we get

$$\begin{aligned}
\mathbb{P}\left(T_{1,\delta} > t\right) &= \sum_{L=1}^{t} \sum_{1 \leq s_1 < \ldots < s_L \leq t} \mathbb{P}\left(\, Q_t = L, S_l = s_l, \Gamma_\delta(s_l - 1, s_{l-1}) \text{ is disc.}, l = 1, ..., L, \Gamma_\delta(t, s_L) \text{ is disc.}\, \right) \\
&= \sum_{L=0}^{t} p_{\mathsf{E},\delta}^{L} \sum_{1 \leq s_1 < \ldots < s_L \leq t} \mathbb{P}(\, \Gamma_\delta(t, s_L) \text{ is disc.}\,) \prod_{l=1}^{L} \mathbb{P}(\, \Gamma_\delta(s_l - 1, s_{l-1}) \text{ is disc.}\,),
\end{aligned} \tag{2.33}$$

where the last equality follows from the fact that realizations of $G_{r,\delta}$ belonging to disjoint blocks are independent, and, also, we implicitly assume that the statement $\Gamma_\delta(s_l - 1, s_{l-1})$ is disc. implies that $G_{r,\delta} \neq \mathsf{E}$ $s_{l-1} < r \leq s_l$. We now fix l, s_l, s_{l-1} and focus on computing $\mathbb{P}(\, \Gamma_\delta(s_l - 1, s_{l-1}) \text{ is disc.}\,)$. To this end, let $\Omega_\delta^\star = \cap_{t \geq 1}\{G_{t,\delta} \in \{\mathcal{G}_\delta \cup \{\mathsf{E}\}\}\}$ and note that, since each of the events in the intersection has probability 1, the event $\Omega_\delta^\star$ also has probability 1. We show that $\{\Gamma_\delta(s_l - 1, s_{l-1}) \text{ is disc.}\} \cap \Omega_\delta^\star \subseteq \cup_{\mathcal{H} \in \Pi(\mathcal{G}_\delta)} \{G_{r,\delta} \in \mathcal{H},\ s_{l-1} < r < s_l\}$, or, in words, if Γ_δ is disconnected on some interval and all the graph realizations that occurred during this interval belong to $\mathcal{G}_\delta$, then there must exist a disconnected collection

on $\mathcal{G}_\delta$ to which all the graph realizations belong to; the last claim, since $\mathbb{P}(\Omega_\delta^\star) = 1$, would yield

$$\mathbb{P}\left(\Gamma_\delta(s_l - 1, s_{l-1}) \text{ is disc.}\right) \leq \sum_{\mathcal{H} \in \Pi(\mathcal{G}_\delta)} p_{\mathcal{H},\delta}^{s_l - s_{l-1} - 1} \leq |\Pi(\mathcal{G}_\delta)| p_{\max,\delta}^{s_l - s_{l-1} - 1}. \tag{2.34}$$

To prove the claim above, consider fixed $\omega \in \Omega_\delta^\star$ such that $\Gamma_\delta(s_l - 1, s_{l-1})$ is disconnected, and let $\mathcal{H}_l = \{G_{s_{l-1}+1,\delta}(\omega), ..., G_{s_l-1,\delta}(\omega)\}$. Since $\omega \in \Omega_\delta^\star$, and we assume that $G_{r,\delta}(\omega) \neq \mathsf{E}$, then it must be that $G_{r,\delta}(\omega) \in \mathcal{G}_\delta$, for all $s_{l-1} < r < s_l$. On the other hand, since $\Gamma_\delta(s_l - 1, s_{l-1})(\omega) = \Gamma(\mathcal{H}_l)$ is disconnected, it follows that $\mathcal{H}_l$ is a disconnected collection on $\mathcal{G}_\delta$, thus proving the claim. Combining now (2.33) and (2.34) yields (2.32):

$$\begin{aligned}
\mathbb{P}\left(T_{1,\delta} > t\right) &\leq \sum_{L=0}^{t} p_{\mathsf{E},\delta}^L \sum_{1 \leq s_1 < ... < s_L \leq t} |\Pi(\mathcal{G}_\delta)| p_{\max,\delta}^{t - s_L} \prod_{l=1}^{L} |\Pi(\mathcal{G}_\delta)| p_{\max,\delta}^{s_l - s_{l-1} - 1} \\
&= \sum_{L=0}^{t} \binom{t}{L} p_{\mathsf{E},\delta}^L \left|\Pi(\mathcal{G}_\delta)\right|^{L+1} p_{\max,\delta}^{t-L} = \left(p_{\max,\delta} + |\Pi(\mathcal{G}_\delta)| p_{\mathsf{E},\delta}\right)^t.
\end{aligned}$$

□

Now, notice that we can express the event that $M_{k,\delta} < \alpha k$ through increments of δ-stopping times: $\{M_{k,\delta} < \alpha k\} = \left\{T_{\lceil \alpha k \rceil,\delta} > k\right\} = \left\{\sum_{i=1}^{\lceil \alpha k \rceil} T_{i,\delta} - T_{i-1,\delta} > k\right\}$. Applying the exponential Markov inequality [69] with parameter $\lambda > 0$

$$\mathbb{P}(M_{k,\delta} < \alpha k) \leq \exp(-\lambda k) \mathbb{E}\left[\exp\left(\sum_{i=1}^{\lceil \alpha k \rceil} \lambda(T_{i,\delta} - T_{i-1,\delta})\right)\right] = \exp(-\lambda k) \left(\mathbb{E}\left[\exp(\lambda T_{1,\delta})\right]\right)^{\lceil \alpha k \rceil}, \tag{2.35}$$

where the equality follows from the fact that the increments of δ-stopping times are i.i.d. We now focus on computing the expectation in the equation above. Using the result of Lemma 2.16

$$\begin{aligned}
\mathbb{E}\left[\exp(\lambda T_{1,\delta})\right] &= \sum_{t=1}^{\infty} \exp(\lambda t) \mathbb{P}(T_{1,\delta} = t) \leq \sum_{t=1}^{\infty} \exp(\lambda t) \mathbb{P}(T_{1,\delta} > t - 1) \\
&\leq \exp(\lambda) \sum_{t=1}^{\infty} \exp(\lambda(t-1)) (p_{\max,\delta} + |\Pi(\mathcal{G}_\delta)| p_{\mathsf{E},\delta})^{t-1}.
\end{aligned} \tag{2.36}$$

The sum in the previous equation converges for all $\lambda < \overline{\mathcal{J}}_\delta$ to $1/(1 - a_\delta(\lambda))$. Combining this with (2.35) completes the proof of part 2. □

From parts 1 and 2 of Lemma 2.15 it follows that for any fixed $\alpha \in (0,1]$, $\delta \in (0,1)$ and $\lambda \in (0, \overline{\mathcal{J}}_\delta)$:

$$\limsup_{k\to\infty} \log \frac{1}{k}\mathbb{P}\left(\left\|\widetilde{\Phi}(k,0)\right\| \geq d_k\right) \leq -\lambda(1-\alpha) - \alpha\log(1-a_\delta(\lambda)). \tag{2.37}$$

Now, taking first the infimum over α and then the infimum over λ yields:

$$\begin{aligned}\limsup_{k\to+\infty} \frac{1}{k}\log\mathbb{P}\left(\left\|\widetilde{\Phi}(k,0)\right\| \geq d_k\right) &\leq \inf_{\lambda\in(0,\overline{\mathcal{J}}_\delta)} \inf_{\alpha\in(0,1]} -\lambda(1-\alpha) - \alpha\log(1-a_\delta(\lambda)) \\ &= \inf_{\lambda\in(0,\overline{\mathcal{J}}_\delta)} -\lambda = -\overline{\mathcal{J}}_\delta. \end{aligned} \tag{2.38}$$

Finally, if $\Pi(\mathcal{G}) \neq \emptyset$, then, by Lemma 9 and Corollary 10, $\overline{\mathcal{J}}_\delta$ converges to $|\log p_{\max}|$, as $\delta \to 0$. On the other hand, if $\Pi(\mathcal{G}) = \emptyset$, it can be easily shown that $\overline{\mathcal{J}}_\delta$ goes to $+\infty$, as $\delta \to 0$. Taking the limit $\delta \to 0$ in eq. (36) yields the upper bound (10).

2.4 Computation of $p_{\max}$ via generalized min-cut

This section introduces a generalization of the minimum cut (min-cut) problem and shows that computing $p_{\max}$ is equivalent to solving an instance of the generalized min-cut. For certain types of averaging, in which the number of graphs that "cover" an edge is relatively small, we show in Subsection 2.4.1 that the generalized min-cut can be well approximated with the standard min-cut, and thus can be efficiently solved. We illustrate this with the broadcast gossip example in Subsection 2.5.1, where we find a 2-approximation for $p_{\max}$ by solving two instances of the standard min-cut.

Generalization of the min-cut. Let $G = (V, E)$ be a given undirected graph, with the set of nodes V and the set of edges E. The generalization of the min-cut problem that is of interest to us assigns a cost to each set of edges $F \subseteq E$. This is different than the standard min-cut, as with the standard min-cut the costs are assigned to each edge of E and, thus, where the cost of F is simply the sum of the individual costs of edges in F. Similarly as with the standard min-cut, the goal is to find F that disconnects G with minimal cost. More formally, let the function $\mathcal{C} : 2^E \mapsto \mathbb{R}_+$ assign costs to subsets of E, i.e., the cost of F is $\mathcal{C}(F)$, for $F \subseteq E$. Then, the generalized min-cut problem is

$$\begin{array}{ll} \text{minimize} & \mathcal{C}(F) \\ \text{subject to} & F \subseteq E : (V, E\setminus F) \text{ is disconnected} \end{array}. \tag{2.39}$$

We denote by $\mathrm{gmc}(G, \mathcal{C})$ the optimal value of (2.39). We remark that, when the cost $\mathcal{C}(F)$ is decomposable

over the edges of F, i.e., when for all $F \subseteq E$, $\mathcal{C}(F) = \sum_{e \in F} c(e)$, for some function $c : E \mapsto \mathbb{R}_+$, then the generalized min-cut simplifies to the standard min-cut. For this case, we denote the optimal value of (2.39) by $\mathrm{mc}(G, c)$.

Consider now a general averaging model on the set of nodes V and with the collection of realizable graphs $\mathcal{G}$. Let $G = \Gamma(\mathcal{G})$, where $G = (V, E)$ and E collects all the edges that appear with positive probability. The following lemma shows that the rate $\mathcal{J}$ for the general averaging model can be computed by solving an instance of the generalized min-cut problem.

Lemma 2.17 Let the cost function $\mathcal{C} : 2^E \mapsto \mathbb{R}_+$ be defined by $\mathcal{C}(F) = \mathbb{P}(\cup_{e \in F} \{e \in E(G_t)\})$, for $F \subset E$. Then,

$$\mathcal{J} = -\log\left(1 - \mathrm{gmc}(G, \mathcal{C})\right) \tag{2.40}$$

Proof For each $F \subseteq E$ such that $(V, E \setminus F)$ is disconnected, define S_F by: $S_F = \{\mathcal{H} \in \Pi(\mathcal{G}) : E(\Gamma(\mathcal{H})) \subseteq E \setminus F\}$. Note that $S_F \subseteq \Pi(\mathcal{G})$, for each F. We show that sets S_F cover $\Pi(\mathcal{G})$, i.e., that $\cup_{F \subseteq E:(V,E\setminus F) \text{ is disc.}} S_F = \Pi(\mathcal{G})$. To this end, pick an arbitrary $\mathcal{H} \in \Pi(\mathcal{G})$ and let $F^\star := E \setminus E(\Gamma(\mathcal{H}))$. Then, because supergraph $\Gamma(\mathcal{H})$ is disconnected, $F^\star$ must be a set of edges that disconnects G; if we now take the set $S_{F^\star}$ that is associated with $F^\star$, we have that $\mathcal{H}$ belongs to $S_{F^\star}$ proving the claim above. Since we established that $\cup_{F \subseteq E:(V,E\setminus F) \text{ is disc.}} S_F = \Pi(\mathcal{G})$, in order to find $p_{\max}$, we can branch the search over the S_F sets:

$$p_{\max} = \max_{\mathcal{H} \in \Pi(\mathcal{G})} p_{\mathcal{H}} = \max_{F \subseteq E:(V,E\setminus F) \text{ is disc.}} \; \max_{\mathcal{H} \in S_F} p_{\mathcal{H}} \tag{2.41}$$

(where, for every empty S_F, we define its corresponding value $\max_{\mathcal{H} \in S_F} p_{\mathcal{H}}$ to be 0). Next, pick a fixed set F for which S_F is nonempty and define $\mathcal{H}_F$ by:

$$\mathcal{H}_F = \{H \in \mathcal{G} : E(H) \subseteq E \setminus F\}; \tag{2.42}$$

that is, $\mathcal{H}_F$ collects all the realizable graphs whose edges do not intersect with F. Note that, by construction of $\mathcal{H}_F$, $E(\Gamma(\mathcal{H}_F)) \subseteq E \setminus F$, proving that $\mathcal{H}_F \in S_F$. Now, for an arbitrary fixed collection $\mathcal{H} \in S_F$, since any graph H that belongs to $\mathcal{H}$ must satisfy the property in (2.42), we have that $\mathcal{H} \subseteq \mathcal{H}_F$ and, consequently, $p_{\mathcal{H}} \leq p_{\mathcal{H}_F}$. This proves that, for every fixed non-empty S_F the maximum $\max_{\mathcal{H} \in S_F} p_{\mathcal{H}}$ is attained at $\mathcal{H}_F$

and equals $p_{\mathcal{H}_F} = \mathbb{P}(E(G_t) \subseteq E \setminus F)$. Combining the last remark with (2.41), yields:

$$p_{\max} = \max_{F \subseteq E:(V,E\setminus F) \text{ is disc.}} \mathbb{P}(E(G_t) \subseteq E \setminus F). \tag{2.43}$$

Finally, noting that $\mathbb{P}(E(G_t) \subseteq E \setminus F) = 1 - \mathbb{P}\left(\cup_{e\in F} \{e \in E(G_t)\}\right)$ completes the proof of Lemma 2.17. □

Rate $\mathcal{J}$ for algorithms running on a tree. When the graph that collects all the links that appear with positive probability is a tree, we obtain a particularly simple solution for $\mathcal{J}$ using formula (2.40). To this end, let $T = (V, E)$ be the supergraph of all the realizable graphs and suppose that T is a tree. Then, removal of any edge from E disconnects T. This implies that, to find the rate, we can shrink the search space of the generalized min-cut problem in (2.40) (see also eq. (2.39)) to the set of edges of the tree:

$$\min_{F \subseteq E:(V,E\setminus F) \text{ is disc.}} \mathcal{C}(F) = \min_{e\in E} \mathcal{C}(e).$$

Now, $\mathcal{C}(e) = \mathbb{P}(e \in E(G_t))$ can be computed by summing up the probabilities of all graphs that cover e, i.e., $\mathcal{C}(e) = \sum_{H\in\mathcal{G}:e\in E(H)} p_H$. The minimum of $\mathcal{C}(e)$ is then achieved at the link that has the smallest probability of occurrence $p_{\text{rare}} = \min_{e\in E} \sum_{H\in\mathcal{G}:e\in E(H)} p_H$. Thus, the rate $\mathcal{J}$ is determined by the probability of the "weakest" link in the tree, i.e., the link that is most rarely online and

$$\mathcal{J}^{\text{Tree}} = -\log\left(1 - p_{\text{rare}}\right). \tag{2.44}$$

2.4.1 Approximation of $\mathcal{J}$ by min-cut based bounds

We now explain how we can compute the rate $\mathcal{J}$ by approximately solving the instance of the generalized min-cut in (2.40) via two instances of the standard min-cut. Our strategy to do this is to "sandwich" each cost $\mathcal{C}(F)$, $F \subseteq E$, by two functions, which are decomposable over the edges of F. To this end, fix F and observe that

$$\mathcal{C}(F) = \mathbb{P}\left(\bigcup_{e\in F} \{e \in E(G_t)\}\right) = \sum_{H\in\mathcal{G}:e\in E(H), \text{ for } e\in F} p_H. \tag{2.45}$$

By the union bound, $\mathcal{C}(F)$ is upper bounded as follows:

$$\mathcal{C}(F) \leq \sum_{e\in F} \mathbb{P}(e \in E(G_t)) = \sum_{e\in F} \sum_{H\in\mathcal{G}:e\in E(H)} p_H. \tag{2.46}$$

We next assume that for every set of m edges we can find m distinct graphs, say $H_1, \ldots, H_m \in \mathcal{G}$, such that H_i covers e_i, $i = 1, \ldots, m$[4]. Then, for each $e \in F$, we can pick a different graph in the sum in (2.45), say H_e, such that $e \in E(H_e)$, until all the edges in F have its associated graph H_e. The sum of the probabilities of the chosen graphs $\sum_{e \in F} p_{H_e}$ is then smaller than $\mathcal{C}(F)$. Finally, we can bound each p_{H_e} by the probability of the least likely graph that covers e, thus yielding:

$$\sum_{e \in F} \min_{H \in \mathcal{G}: e \in E(H)} p_H \leq \mathcal{C}(F).$$

Motivated by the previous observations, we introduce $\overline{c}, \underline{c} : E \mapsto \mathbb{R}_+$ defined by

$$\overline{c}(e) = \sum_{H \in \mathcal{G}: e \in E(H)} p_H, \;\; \underline{c}(e) = \min_{H \in \mathcal{G}: e \in E(H)} p_H. \tag{2.47}$$

Then, for each $F \subseteq E$, we have:

$$\sum_{e \in F} \underline{c}(e) \leq \mathcal{C}(F) \leq \sum_{e \in F} \overline{c}(e).$$

Because the inequality above holds for all $F \subseteq E$, we have that:

$$\mathrm{mc}(G, \underline{c}) \leq \mathrm{gmc}(G, \mathcal{C}) \leq \mathrm{mc}(G, \overline{c}). \tag{2.48}$$

Therefore, we can efficiently approximate the rate $\mathcal{J}$ by solving two instances of the standard min-cut problem, with the respective costs $\underline{c}$ and $\overline{c}$. To further simplify the computation of $\mathcal{J}$, we introduce D– the maximal number of graphs that "covers" an edge e, where the maximum is over all edges $e \in E$. We also introduce $\overline{p}$ and $\underline{p}$ to denote the probabilities of the most likely and least likely graph, respectively, i.e., $\overline{p} = \max_{H \in \mathcal{G}} p_H$ and $\underline{p} = \min_{H \in \mathcal{G}} p_H$. Then, the function $\overline{c}$ can be uniformly bounded by $D\overline{p}$ and, similarly, function $\underline{c}$ can be uniformly bounded by $\underline{p}$, which combined with (2.48) yields[5]:

$$\underline{p}\,\mathrm{mc}(G, 1) \leq \mathrm{gmc}(G, \mathcal{C}) \leq D\,\overline{p}\,\mathrm{mc}(G, 1); \tag{2.49}$$

The expression in (2.49) gives a $D\,\overline{p}/\underline{p}$-approximation for $\mathrm{gmc}(G, \mathcal{C})$, and it requires solving only one

[4]The case when this is not true can be handled by splitting the probability p_H of a graph H into d equal parts, where d is the number of edges covered by H. The approximation bounds (that are derived further ahead) would then depend on d; we omit the details here due to lack of space

[5]We are using here the property of the min-cut with uniform positive costs by which $\mathrm{mc}(G, \alpha 1) = \alpha \mathrm{mc}(G, 1)$, for $\alpha \geq 0$ [73], where 1 denotes the cost function that has value 1 at each edge

instance of the standard min-cut, with uniform (equal to 1) costs.

2.5 Examples: randomized gossip and fading model

This section computes the rate $\mathcal{J}$ for the commonly used averaging models: randomized gossip and link failure. Subsection 2.5.1 studies two types of the randomized gossip algorithm, namely pairwise gossip and symmetrized broadcast gossip and it shows that, for the pairwise gossip on a generic graph $G = (V, E)$, the corresponding rate can be computed by solving an instance of the standard min-cut; for broadcast gossip, we exploit the bounds derived in Subsection 2.4.1 to arrive at a tight approximation for its corresponding rate. Subsection 2.5.2 studies the network with fading links for the cases when 1) all the links at a time experience the same fading (correlated fading), and 2) the fading is independent across different links (uncorrelated fading). Similarly as with the pairwise gossip, the rate for the uncorrelated fading can be computed by solving an instance of a min-cut problem. With the correlated fading, there exists a threshold on the fading coefficients, which induces two regimes of the network operation, such that if at a time t the fading coefficient is above the threshold, the network realization at time t is connected. We show that the rate is determined by the probability of the "critical" link that marks the transition between these two regimes.

2.5.1 Pairwise and broadcast gossip

Min-cut solution for pairwise gossip. Let $G = (V, E)$ be an arbitrary connected graph on N vertices. With pairwise gossip on graph G, at each averaging time, only one link from E can be active. Therefore, the set of realizable graphs $\mathcal{G}^{\text{Gossip}}$ is the set of all one link graphs on G:

$$\mathcal{G}^{\text{Gossip}} = \{(V, e) : e \in E\}.$$

Now, consider the probability $\mathbb{P}(\cup_{e\in F} \{e \in E(G_t)\})$, for a fixed subset of edges $F \subseteq E$. Because each realization of G_t can contain only one link, the events under the union are disjoint. Thus, the probability of the union equals the sum of the probabilities of individual events, yielding that the cost $\mathcal{C}(F)$ is decomposable for gossip, i.e.,

$$\mathcal{C}(F) = \sum_{e\in F} p_{(V,e)}$$

Therefore, the rate for gossip is given by

$$\mathcal{J}^{\text{Gossip}} = -\log\left(1 - \text{mc}(G, c^{\text{Gossip}})\right), \tag{2.50}$$

where $c^{\text{Gossip}}(e) = p_{(V,e)}$. We remark here that, for pairwise gossip, functions $\underline{c}, \overline{c}$ in (2.47) are identical (each link e has exactly one graph (V, e) that covers it, hence $\underline{c}(e) = \overline{c}(e) = p_{(V,e)}$), which proves that bounds in (2.48) are touched for this problem instance. For the case when all links have the same activation probability equal to $1/|E|$, the edge costs $c^{\text{Gossip}}(e)$ are uniform and equal to $1/|E|$, for all $e \in E$ and (2.50) yields the following simple formula for the rate for uniform gossip:

$$\mathcal{J}^{\text{Gossip}} = -\log\left(1 - 1/|E|\,\text{mc}(G,1)\right). \tag{2.51}$$

Gossip on a regular network. Consider the special case when the gossip algorithm runs on a connected regular graph of degree d, $d = 2, ..., N-1$, and the link occurrence probabilities are all equal, $p := p_{ij} = \frac{2}{Nd}$. It can be easily seen that the value of the min-cut is p times the minimal number of edges that disconnects the graph, which equals $pd = 2/N$; this corresponds to cutting all the edges of a fixed node, i.e., isolating a fixed node. Hence,

$$p_{\max} = \mathbb{P}\left(\text{node } i \text{ is isolated}\right) = 1 - 2/N, \;\; \mathcal{J} = -\log(1 - 2/N).$$

Note that the rate $\mathcal{J}$ is determined by the probability that a fixed node is isolated, and, also, the rate $\mathcal{J}$ does not depend on the degree d.

2-approximation for broadcast gossip. With bidirectional broadcast gossip on an arbitrary connected graph $G = (V, E)$, at each time a node $v \in V$ is chosen at random and the averaging is performed across the neighborhood of v. Thus, at each time t, the set of active edges is the set of all edges adjacent to the vertex that is chosen at time t; hence, the set of realizable graphs $\mathcal{G}^{\text{B-Gossip}}$ is

$$\mathcal{G}^{\text{B-Gossip}} = \left\{(V, \{\{u, v\} : \{u, v\} \in E\} : v \in V\right\}.$$

We can see that each edge $e = \{u, v\}$ can become active in two ways, when either node u or node v is active. In other words, each edge is covered by exactly two graphs. This gives $D = 2$ and using (2.48) we get the following approximation:

$$\underline{p}\,\text{mc}(G,1) \le \text{gmc}(G,\mathcal{C}) \le 2\overline{p}\,\text{mc}(G,1),$$

where $\underline{p}$ and $\overline{p}$ are the probabilities of the least, resp., most, active node. For the case when all the nodes

have the same activation probability equal to $1/N$, using (2.49) we get a 2-approximation:

$$\frac{1}{N}\mathrm{mc}(G,1) \leq \mathrm{mc}(G,\mathcal{C}) \leq \frac{2}{N}\mathrm{mc}(G,1).$$

Thus, the rate $\mathcal{J}$ for the broadcast gossip with uniform node activation probability satisfies:

$$\mathcal{J}^{\text{B-Gossip}} \in [-\log(1-\frac{1}{N}\mathrm{mc}(G,1)), -\log(1-\frac{2}{N}\mathrm{mc}(G,1))]. \tag{2.52}$$

We now compare the rates for the uniform pairwise and uniform broadcast gossip when both algorithms are running on the same (connected) graph $G = (V, E)$. Consider first the case when G is a tree. Then, $E = N - 1$ and, since all the links have the same occurrence probability $1/(N-1)$, the formula for gossip gives $\mathcal{J}^{\text{Gossip}} = -\log(1-1/(N-1))$. To obtain the exact rate for the broadcast gossip, we recall formula (2.44). As each link in the tree is covered by exactly two graphs, and the probability of a graph is $1/N$, we have that $p^{\text{rare}} = 2/N$. Therefore, the rate for broadcast gossip on a tree is $\mathcal{J}^{\text{Gossip}} = -\log(1-2/N)$, which is higher than $\mathcal{J}^{\text{Gossip}} = -\log(1-1/(N-1))$. Consider now the case when G is not a tree. Then, the number of edges $|E|$ in G is at least N and we have $\mathcal{J}^{\text{Gossip}} = -\log(1-1/|E|\mathrm{mc}(G,1)) \leq -\log(1-1/N\mathrm{mc}(G,1))$. On the other hand, by (2.52), $\mathcal{J}^{\text{B-Gossip}} \geq -\log(1-1/N\mathrm{mc}(G,1))$. Combining the last two observations yields that the rate of broadcast gossip is always higher than the rate of pairwise gossip running on the same graph. This is in accordance with the intuition, as with broadcast gossip more links are active at a time, and, thus, we would expect that it performs the averaging faster.

2.5.2 Link failure: fading channels

Consider a network of N sensors described by graph $G = (V, E)$, where the set of edges E collects all the links $\{i, j\}$ that appear with positive probability, $i, j \in V$. To model the link failures, we adopt a symmetric fading channel model, a model similar to the one proposed in [74] (reference [74] assumes asymmetric channels). At time k, sensor j receives from sensor i

$$y_{ij,k} = g_{ij,k}\sqrt{\frac{S_{ij}}{d_{ij}^{\alpha}}}x_{i,k} + n_{ij,k},$$

where S_{ij} is the transmission power that sensor i uses for transmission to sensor j, $g_{ij,k}$ is the channel fading coefficient, $n_{ij,k}$ is the zero mean additive Gaussian noise with variance σ_n^2, d_{ij} is the inter-sensor distance, and α is the path loss coefficient. We assume that $g_{ij,k}$, $k \geq 1$, are i.i.d. in time and that $g_{ij,t}$ and $g_{lm,s}$ are mutually independent for all $t \neq s$; also, the channels (i, j) and (j, i) at time k experience the same fade, i.e.,

$g_{ij,k} = g_{ji,k}$. We adopt the following link failure model. Sensor j successfully decodes the message from sensor i (link (i,j) is online) if the signal to noise ratio exceeds a threshold, i.e., if: $\text{SNR} = \frac{S_{ij} g_{ij,k}^2}{\sigma_n^2 d_{ij}^\alpha} > \tau$, or, equivalently, if $g_{ij,k}^2 > \frac{\tau \sigma_n^2 d_{ij}^\alpha}{S_{ij}} := \gamma_{ij}$. Since link occurrences are "controlled" by the realizations of the fading coefficients, the set of realizable graphs in the link failure model depends on the joint distribution of $\{g_{ij,k}\}_{\{i,j\}\in E}$. In the sequel, we study the cases when the fading coefficients at some time k are either fully correlated or uncorrelated, and we compute the rate $\mathcal{J}$ for each of these cases.

Uncorrelated fading. With uncorrelated fading, $g_{ij,k}$ are independent across different links for all k. Therefore, in this model, the indicators of link occurrences are independent Bernoulli random variables, such that the indicator of link $\{i,j\}$ being online is 1 if the fading coefficient at link $\{i,j\}$ exceeds the communication threshold of $\{i,j\}$, i.e., if $g_{ij,k}^2 > \gamma_{ij}$, and is zero otherwise. Due to the independence, each subgraph $H = (V, E')$ of G, $E' \subseteq E$, is a realizable graph in this model, hence,

$$\mathcal{G}^{\text{Fail-uncorr}} = \left\{ H = (V, E') : E' \subseteq E \right\}.$$

Also, the probability of occurrence of $H = (V, E') \in \mathcal{G}^{\text{Fail-uncorr}}$ is

$$p_H = \prod_{\{i,j\}\in E'} P_{ij} \prod_{\{l,m\}\in E\setminus E'} (1 - P_{lm}),$$

Denote with $P_{ij} = \mathbb{P}(g_{ij,k}^2 > \gamma_{ij})$ the probability that link $\{i,j\}$ is online. We compute the rate $\mathcal{J}^{\text{Fail-uncorr}}$ for the uncorrelated link failure using the result of Lemma 2.17. To this end, let F be a fixed subset of E and consider the probability that defines $\mathcal{C}(F)$. Then,

$$\begin{aligned}
\mathcal{C}(F) &= \mathbb{P}\left(\cup_{\{i,j\}\in F} \{\{i,j\} \in E(G_t)\} \right) \\
&= 1 - \mathbb{P}\left(\cap_{\{i,j\}\in F} \{\{i,j\} \notin E(G_t)\} \right) \\
&= 1 - \prod_{\{i,j\}\in F} (1 - P_{ij}),
\end{aligned}$$

where the last equality follows by the independence of the link failures. To compute the rate, we follow

formula (2.40):

$$1 - \min_{F \subseteq E:(V,E\setminus F) \text{ is disc.}} \mathcal{C}(F) \tag{2.53}$$

$$= \max_{F \subseteq E:(V,E\setminus F) \text{ is disc.}} \prod_{\{i,j\} \in F} (1 - P_{ij})$$

$$= \exp(- \min_{F \subseteq E:(V,E\setminus F) \text{ is disc.}} \sum_{\{i,j\} \in F} - \log(1 - P_{ij})). \tag{2.54}$$

The optimization problem in the exponent is an instance of the standard min-cut problem with edge costs

$$c^{\text{Fail-uncorr}}(\{i,j\}) = -\log(1 - P_{ij}). \tag{2.55}$$

By formula (2.40), the rate is obtained from the expression in line (2.53) by taking the $-\log$, which finally yields:

$$\mathcal{J}^{\text{Fail-uncorr}} = \text{mc}(G, c^{\text{Fail-uncorr}}). \tag{2.56}$$

Regular graph and uniform link failures. Consider now the special case when the underlying graph is a connected regular graph with degree d, $d = 2, ..., N-1$, and the uniform link occurrence probabilities $p_{ij} = p$. It is easy to see that $p_{\max}$ and $\mathcal{J}$ simplify to:

$$p_{\max} = \mathbb{P}\left(\text{node } i \text{ is isolated}\right) = (1-p)^d,$$
$$\mathcal{J} = -d \log(1-p).$$

Correlated fading. With the correlated fading, at any time k each link experiences the same fading, i.e., $g_{ij,k} = g_k$ for all $\{i,j\} \in E$ and so the realization of the common fading g_k sets all the link occurrences at time k. For instance, if $g_k^2 = \bar{g}^2$, then all the links $\{i,j\}$ with $\gamma_{ij} < \bar{g}^2$ are online, and the rest of the links are offline. Therefore, the graph realization corresponding to the fading realization $\bar{g}^2$ is (V, E'), where $E' = \left\{\{i,j\} \in E : \gamma_{ij} < \bar{g}^2\right\}$. We can see that the higher the $\bar{g}^2$ is, the more links are online. Also, if we slide $\bar{g}^2$ from zero to $+\infty$, the corresponding graph realization gradually increases in size by adding one more link whenever $\bar{g}^2$ crosses some threshold γ_{ij} – starting from the empty graph ($\bar{g}^2 = 0$), until the full graph (V, E) is achieved, which occurs when $\bar{g}^2$ crosses the highest threshold. Therefore, if we order the links in the increasing order with respect to their thresholds γ_{ij}, such that $e_1 \in E$ has the lowest threshold, $\gamma_{e_1} = \min_{\{i,j\} \in E} \gamma_{ij}$, and $e_{|E|} \in E$ has the highest threshold, $\gamma_{e_{|E|}} = \max_{\{i,j\} \in E} \gamma_{ij}$, then the set of all

realizable graphs with the correlated fading model is

$$\mathcal{G}^{\text{Fail-corr}} = \{H_l = (V, \{e_1, e_2, \ldots, e_l\}) : 0 \leq l \leq |E|\},$$

where the graph realization corresponding to $l = 0$ is the empty graph $(V, \emptyset)$. For fixed l, $0 \leq l \leq |E|$, let $p_l = \mathbb{P}(g_k^2 > \gamma_l)$ denote the probability that link e_l is online. Then, the probability p_{H_l} of graph realization H_l is $p_{H_l} = p_l - p_{l+1} = \mathbb{P}(\gamma_l < g_k^2 \leq \gamma_{l+1})$. Let l^{c} be the index corresponding to the link that marks the connectedness transition of graphs H_l, such that H_l is disconnected for all $l < l^{\text{c}}$, and H_l is disconnected, for all $l \geq l^{\text{c}}$. Then, any disconnected collection on $\mathcal{G}^{\text{Fail-corr}}$ is of the form $\{H_1, H_2, \ldots, H_l\}$, where $l < l^{\text{c}}$. The most likely one is $\{H_1, H_2, \ldots, H_{l^{\text{c}}-1}\}$, and its probability is $p_{H_1} + p_{H_2} + \ldots + p_{H_{l^{\text{c}}-1}} = 1 - p^{\text{crit}}$, where we use p^{crit} to denote the probability of the "critical" link $e_{l^{\text{c}}}$ (i.e., $p^{\text{crit}} = p_{l^{\text{c}}}$). Therefore, $p_{\max} = 1 - p^{\text{crit}}$ and the rate for the correlated fading model is:

$$\mathcal{J}^{\text{Fail-corr}} = -\log\left(1 - p^{\text{crit}}\right).$$

Chapter 3

Products of Random Stochastic Matrices: Temporal Dependencies and Directed Networks

3.1 Introduction

We have seen in Chapter 2 how to find and compute the large deviation rate $\mathcal{J}$ for products of i.i.d. stochastic symmetric matrices. In this chapter we go beyond the results in Chapter 2 in the following two ways. First, we generalize Theorem 2.7 for the sequence of *temporally dependent random matrices*. More specifically, we associate a state of a Markov chain to each of the topology realizations G_t. The distribution of the topologies G_t, $t \geq 1$, is then determined by a specified $\mathcal{M} \times \mathcal{M}$ transition probability matrix P, where $\mathcal{M}$ is the number of possible realizations of G_t. This model subsumes, e.g., the token-based protocols similar to [17], or temporally dependent link failure models, where the on/off state of each link follows a Markov chain. The model that we study is also very similar to the one proposed in [75]. Reference [75] considers random consensus algorithms where the sequence of switching topologies follows a Markov chain and derives conditions for almost sure convergence to consensus. Besides the difference in the problems themselves, we note that the model that we study is more general than the one in [75]. In [75], the consensus matrices W_t are assigned deterministically once G_t is given: $W_t = I - \alpha L(G_t)$, where $L(G_t)$ is the Laplacian matrix of the topology G_t and α is a sufficiently small constant. Contrary to [75], in our model we allow each W_t to be chosen randomly from the set of matrices with the sparsity pattern defined by G_t and, furthermore, the conditional distributions of W_t (conditioned on the realization of G_t) can differ across

time t.

We characterize the rate $\mathcal{J}$ as a function of the transition probability matrix P. We refer to Theorem 3.3 for details, but here we briefly convey the general idea. Namely, the rate $\mathcal{J}$ will be determined by the most likely way the Markov chain stays "trapped" in some subset of states (graphs) whose union is disconnected. The probability of this event is determined by the spectral radius of the block in the transition matrix P that corresponds to this most likely subset of states, and the value of this spectral radius will thus determined the rate $\mathcal{J}$. We illustrate the results on two examples, namely gossip with Markov dependencies and temporally correlated link failures. The example with temporally correlated link failures shows that "negative temporal correlations" of the links' states (being ON or OFF) increase (improve) the rate $\mathcal{J}$ when compared with the uncorrelated case, while positive correlations decrease (degrade) the rate. This result is in accordance with standard large deviations results on temporally correlated sequences, see, e.g., [[18], exercise V.12, page 59.]

In our second generalization of Theorem 2.7 we remove the assumption that the matrices W_t need to be symmetric. This is of special importance for distributed algorithms that run on networks in which the physical communication links can be asymmetric (e.g., at some time, i successfully sends a packet to j, but the packet that j sends to i drops). When W_t's are stochastic and with positive diagonal, it is known that the product $W_k \cdots W_1$ converges almost surely (a.s.) to a random, rank one matrix $1v^\top$ (the vector v is random) [27], under the condition that $|\lambda_2(\mathbb{E}[W_k])|$ is strictly less than one[1]. Further, the path-wise convergence of $W_k \cdots W_1$ to $1v^\top$ is equivalent to the path-wise convergence of $|\lambda_2(W_k \cdots W_1)|$ to zero. Thus, as a measure of how far the product at time k is from its limit, we naturally adopt $|\lambda_2(W_k \cdots W_1)|$. Similarly as in Chapter 2, we are interested in characterizing the probability that the convergence of $W_k \cdots W_1$ to a (random) limit $1v^\top$ is subexponential. More precisely, let d_k, $k \geq 1$, $d_k \in (0, 1]$ be a sequence with a decay rate slower than exponential, i.e., $\log d_k = o(k)$. Then, adopting $|\lambda_2(\cdot)|$ as the metric, we study the probability of the event that, at some time k, the product $\Phi(k, 0) = W_k \cdots W_1$ is still d_k far away from its limit:

$$\mathbb{P}\left(|\lambda_2(W_k \cdots W_1)| \geq d_k\right), \quad k = 1, 2, \ldots. \tag{3.1}$$

We show that the sequence of probabilities in (3.1) decays exponentially fast with k. More precisely, for any sequence $d_k \in (0, 1]$, $k \geq 1$, such that $\log d_k = o(k)$, we show that the following large deviation limit exists:

$$\mathcal{J} = \lim_{k \to +\infty} -\frac{1}{k} \log \mathbb{P}\left(|\lambda_2(W_k \cdots W_1)| \geq d_k\right). \tag{3.2}$$

[1]Note that this condition is equivalent to the condition that the topology of the expected matrix $\mathbb{E}[W_k]$ contains a directed spanning tree.

We fully characterize the limit $\mathcal{J}$ and show that it depends on the distribution of matrices only through their support graphs. More precisely, $\mathcal{J}$ is determined by the probability of the most likely set of support graphs whose union fails to form a directed spanning tree. Thus, the characterization of $\mathcal{J}$ that we discover exhibits full consistency with the result for symmetric matrices in Chapter 2: for undirected networks a collection of topologies is jointly tree-free if and only if it is disconnected, and thus when the matrices are symmetric the two rates $\mathcal{J}$ in (3.2) and in (2.2) coincide. Finally, to illustrate our results we consider a commonly used broadcast gossip protocol [19] in sensor networks, where (only one) node u activates at a time with probability p_u, and broadcasts its state to all single-hop neighbors. For this model, the rate $\mathcal{J} = |\log 1 - p_{\min}|$, where $p_{\min}$ is the probability of the most rarely active node.

Portions of this chapter have been published in conference proceedings [76] and [77]. The results from this chapter are to be submitted for a journal publication.

Chapter organization. The next paragraph introduces notation that we use throughout the chapter. Section 3.2 studies the model with temporal dependencies, and Section 3.3 studies the model with directed networks.

Notation. We denote by: A_{ij} or $[A]_{ij}$ the entry in ith row and jth column of a matrix A; A_l and A^l the l-th row and column, respectively; $\rho(A)$ the spectral radius of A; I and $J := (1/N)11^\top$ the identity matrix, and the ideal consensus matrix, respectively; 1 and e_i the vector with unit entries, and ith canonical vector (with ith entry equal to 1 and the rest being zeros), respectively. Further, for a vector a, the inequality $a > 0$ is understood component wise. Given a selection $S \subseteq \{1, ..., N\}$ of rows and columns of a matrix A: $\{A_l : l \in S\}$ and $\{A^l : l \in S\}$, we denote by A_S the submatrix of A corresponding to the selection S. Similarly, if S is a selection of rows, we denote by A_{Sl} the part of A^l that corresponds to the selection S. Likewise, for the selection of columns S, we denote by A_{lS} the part of A_l that corresponds to S.

3.2 Matrices with Markov dependencies

In Subsection 3.2.1 we explain the random model with temporal dependencies that we address and state the main result on the large deviation rate $\mathcal{J}$ in Theorem 3.3. This result is proved in Subsections 3.2.2 and 3.2.3: in Subsection 3.2.2 we prove the large deviation lower bound, and in Subsection 3.2.3 we prove the large deviation upper bound. Finally, Subsection 3.2.4 gives examples for the studied model with temporal dependencies and it illustrates through the examples what is the effect of correlations on the rate $\mathcal{J}$.

3.2.1 Random model and the main result

Graph temporal dependence. Let G_t, $t \geq 1$, be a sequence of random graphs that takes realizations in $\mathbb{G}^N$ - the set of all undirected graphs on the set of vertices $\{1, ..., N\}$. We assume that G_t is defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ such that $\{G_t = H\} \in \mathcal{F}$ for each $H \in \mathbb{G}^N$. Similarly as in Chapter 2, let $\mathcal{G}$ denote the set of all possible graph realizations; that is, $\mathcal{G} = \{H_1, ..., H_{\mathcal{M}}\}$, where, for each l, $\mathbb{P}(G_t = H_l) > 0$ for some $t \geq 1$ and for every t $\mathbb{P}(G_t \in \mathcal{G}) = 1$.

We assume that the sequence of random graphs G_t follows a Markov chain.

Assumption 3.1 (Markov chain of graphs G_t) There exist a nonnegative matrix $P \in \mathbb{R}^{\mathcal{M}\times\mathcal{M}}$ and a nonnegative vector $v \in \mathbb{R}^{\mathcal{M}}$ satisfying $\sum_{m=1}^{\mathcal{M}} P_{lm} = 1$ for all $l = 1, \ldots, \mathcal{M}$ and $\sum_{l=1}^{\mathcal{M}} v_l = 1$, such that for all t and all $l_1, l_2, \ldots, l_t \in \{1, \ldots, \mathcal{M}\}$

$$\mathbb{P}(G_1 = H_{l_1}, G_2 = H_{l_2}, \ldots, G_t = H_{l_t}) = v_{l_1} P_{l_1 l_2} \cdots P_{l_{t-1} l_t}.$$

Thus, each state in this Markov chain corresponds to one realizable graph H_l, and the chain of graph realizations evolves according to the transition matrix P: assuming that H_l is the current topology, the probability to switch to topology H_m in the next time is given by the entry l, m of P, P_{lm}.

Suppose that W_t, $t \geq 1$, is a sequence of random matrices defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$ on which the sequence of graphs G_t is defined, such that, for each t, the corresponding W_t takes realizations in $\mathbb{S}^N$ – the set of all symmetric stochastic N by N matrices, and is $(\mathcal{F}, \mathcal{B}(\mathbb{R}^{N\times N}) \cap \mathbb{S}^N)$-measurable. We will further, for each t, assume that W_t has positive diagonal entries a.s. (similarly as in Chapter 2), and also that there exists a small positive number δ such that all positive entries of W_t are a.s. lower bounded by δ. Now, in terms of the joint distribution of the W_t's, the only assumption that we make is that at each time t, W_t respects the sparsity pattern of G_t.

Assumption 3.2 (Matrices W_t)

1. $G(W_t(\omega)) = G_t(\omega)$, for every t and every ω;
2. There exists $\delta > 0$ such that the following two conditions are satisfied:
   - diagonal entries of W_t are a.s. greater or equal to δ; that is, for every t, with probability 1 $[W_t]_{ii} \geq \delta$ for all i;
   - Whenever positive, the entries $[W_t(\omega)]_{ij}$ are a.s. greater or equal to δ; that is, for every t, if $\mathbb{P}(\{i, j\} \in G_t) > 0$ for some i, j, then, conditioned on $\{\{i, j\} \in G_t\}$, $[W_t]_{ij} \geq \delta$ with probability 1.

Thus, we see that W_t's will be intrinsically dependent between themselves as a result of correlations between their support graphs: by construction, the sequence W_t, $t \geq 1$, is such that the induced graphs $G(W_t)$, $t \geq 1$, follow a Markov chain. However, note that this puts no constraints on the value of the positive elements of the matrices: once the sparsity pattern of each W_t is set by the sequence G_t, the joint distribution of the positive entries of $\{W_t : t \geq 1\}$ can be arbitrary. For instance, for any t, given the realization of G_t, $t \geq 1$, the positive entries of W_t can be picked as a function of the arbitrary subset of elements of $\{W_t : t \geq 1\}$. Therefore, the model that we assume for the sequence W_t allows for much stronger and longer ranging correlations than those of the Markov chain of their induced graphs G_t.

We next illustrate the model with two simple examples. First, the obvious choice for W_t is when W_t is itself a Markov chain. For example, suppose that W_t is a Markov chain of matrices on the set of states $\{A_1, ..., A_{\mathcal{M}}\}$, such that each A_m has positive diagonal entries. Suppose that each A_m has a different support graph [2]. Then, defining $G_t = G(W_t)$, and noting that the minimal positive entry among all the matrices A_m is strictly greater than zero (due to finiteness of the set of states) we see that W_t falls in the class of models that satisfy Assumption 3.2. Note that in this example we have a one-to-one correspondence between W_t and G_t, for each t.

On the other extreme, we could create a sequence of matrices satisfying Assumption 3.2 in which, at every time t, positive entries of W_t are completely independent of G_t. To show how this can be done, define for every $H \in \mathbb{G}^N$ set $\mathbb{S}^H = \left\{S \in \mathbb{S}^N : G(S) = H,\ S_{ii} \geq \delta \text{ for } i = 1, ..., N, \text{ and } S_{ij} \geq \delta \text{ for } \{i, j\} \in H\right\}$, where δ is some small positive number; that is, $\mathbb{S}^H$ is the set of all stochastic symmetric matrices with the sparsity pattern given by H, and whose diagonal and positive entries are lower bounded by $\delta > 0$. Let now the sequence W_t be defined as follows: for any time t, given $G_t = H$, W_t is picked uniformly at random from $\mathbb{S}^H$. It is easy to check that W_t satisfies Assumption 3.2. Also, we can see that the distribution of W_t given G_t depends on G_t only through its sparsity pattern, which we wanted to show. Remark finally that, instead of the uniform distribution, we could have chosen for each H an arbitrary distribution on $\mathbb{S}^H$ to generate W_t given H (e.g., uniform distribution on $\mathbb{S}^{H_1}$, for the realization $G_t = H_1$, and, say, discrete on $\mathbb{S}^{H_2}$, for realization $G_t = H_2 \neq H_1$), and the same conclusions would hold.

Further models that satisfy Assumption 3.2 are given in Subsection 3.2.4.

We assume in the sequel that $v > 0$.

[2] Note that we need this assumption because we assumed that each graph in the set of states $\mathcal{G}$, of the graph Markov chain, is different. In order to address the case when a Markov chain of matrices has matrices (states) with the same sparsity patterns, we simply modify the model by creating (where necessary) multiple states for the same topology. The rest of the analysis would then proceed the same.

Theorem 3.3 Let d_k be a sequence of real numbers such that $d_k \in (0,1]$ and $\log d_k = o(k)$. Then:

$$\lim_{k\to\infty} \frac{1}{k} \log \mathbb{P}\left(\left\| \widetilde{\Phi}(k,0) \right\| \geq d_k \right) = -\mathcal{J},$$

where

$$\mathcal{J} = \begin{cases} |\log \rho_{\max}|, & \text{if } \Pi(\mathcal{G}) \neq \emptyset \\ +\infty, & \text{otherwise} \end{cases}, \tag{3.3}$$

and $\rho_{\max} = \max_{\mathcal{H}\in\Pi(\mathcal{G})} \rho\left(P_{\mathcal{H}}\right)$.

To prove Theorem 3.3, we consider first the case when $\Pi(\mathcal{G}) = \emptyset$. In this case each realization of G_t is connected (otherwise, $\Pi(\mathcal{G})$ would contain at least this disconnected realization). Applying Corollary 2.13 from Chapter 2 to successive graph realizations (i.e., for $s = t+1$) we get that

$$\left\| \widetilde{\Phi}(k,0) \right\| \leq \left(1 - c_N \delta^2\right)^{\frac{k}{2}}. \tag{3.4}$$

Now, for any given sequence $d_k \in (0,1]$ satisfying $\log d_k = o(k)$, for any $\epsilon > 0$, there exists $k_1 = k_1(\epsilon)$ such that $\frac{\log d_k}{k} > -\epsilon$ for all $k \geq k_1$. Taking ϵ to be the absolute value of the logarithm of the left hand side of (3.4), we have that, pointwise (and thus with probability 1), $\left\| \widetilde{\Phi}(k,0) \right\| < d_k$, for all $k \geq k_1$. Therefore, the probability from Theorem 3.3 is equal to zero for all $k \geq k_1$, yielding the rate $I = \infty$. This completes the proof of Theorem 3.3 for the case when $\Pi^\star(\mathcal{G}) = \emptyset$.

We prove Theorem 3.3 for the case when $\Pi(\mathcal{G}) \neq \emptyset$ by showing the upper and the lower large deviation bound:

$$\liminf_{k\to\infty} \frac{1}{k} \log \mathbb{P}\left(\left\| \widetilde{\Phi}(k,0) \right\| \geq d_k \right) \quad \geq \quad \log \rho_{\max} \tag{3.5}$$

$$\limsup_{k\to\infty} \frac{1}{k} \log \mathbb{P}\left(\left\| \widetilde{\Phi}(k,0) \right\| \geq d_k \right) \quad \leq \quad \log \rho_{\max}. \tag{3.6}$$

Subsection 3.2.2 states some important preliminary results needed later in the proofs of both the upper and the lower bound, and then it proceeds to prove the lower bound. The proof of the upper bound is because of its complexity given separately in Subsection 3.2.3.

3.2.2 Preliminary results and the proof of the lower bound (3.5)

Lemma 3.4 is a simple result for Markov chains: if we start from the state H_l at time t, end up in the state H_m at time $s+1$, and we restrict the trajectory $(G_r,\ t+1 \leq r \leq s)$ to belong to a subset of states $\mathcal{S}$, the

probability of this event occurring is determined by the submatrix $P_{\mathcal{S}}$.

Lemma 3.4 Let H_m and H_l, $1 \leq l, m \leq \mathcal{M}$ be two given states and $\mathcal{S} \subseteq \mathcal{G}$ a given subset of states. Then, for any $1 \leq t < s$ (assuming $P(G_t = H_l) > 0$):

$$\mathbb{P}(G_r \in \mathcal{S},\ t+1 \leq r \leq s,\ G_{s+1} = H_m \mid G_t = H_l) = P_{l\mathcal{S}}\, P_{\mathcal{S}}^{s-t-1}\, P_{\mathcal{S}m}. \tag{3.7}$$

Proof We prove Lemma 3.4 by induction on $s - t$. Fix $l, m, \mathcal{S}$ and s, t such that $s - t = 1$. We have

$$\begin{aligned}
\mathbb{P}(G_{t+1} \in \mathcal{S}, G_{s+1} = H_m | G_t = H_l) &= \sum_{n:H_n \in \mathcal{S}} \mathbb{P}(G_{t+1} = H_n, G_{s+1} = H_m | G_t = H_l) \\
&= \sum_{n:H_n \in \mathcal{S}} P_{ln} P_{nm} \\
&= P_{lS} I P_{Sm},
\end{aligned}$$

which proves that the formula in (3.7) is correct for $s - t = 1$. Assume now that (3.7) is true for all $l, m, \mathcal{S}$ and all s, t such that $s - t = r \geq 1$. Consider now the probability in (3.7) for fixed $l, m, \mathcal{S}$ and s', t' such that $s' - t' = r + 1$. Summing out over all realizations of $G_{s'}$ that belong to $\mathcal{S}$, we get

$$\begin{aligned}
&\mathbb{P}\left(G_{t'+1} \in \mathcal{S}, \ldots, G_{s'} \in \mathcal{S}, G_{s'+1} = H_m | G'_t = H_l\right) \\
&= \sum_{n:H_n \in \mathcal{S}} \mathbb{P}\left(G_{t'+1} \in \mathcal{S}, \ldots, G_{s'} = H_n, G_{s'+1} = H_m | G'_t = H_l\right) \\
&= \sum_{n:H_n \in \mathcal{S}} \mathbb{P}\left(G_{t'+1} \in \mathcal{S}, \ldots, G_{s'} = H_n | G'_t = H_l\right) \mathbb{P}(G_{s'+1} = H_m | G_{s'} = H_n) \\
&= \sum_{n:H_n \in \mathcal{S}} P_{l\mathcal{S}} P_{\mathcal{S}}^{s'-1-t'-1} P_{\mathcal{S}n} P_{nm}
\end{aligned}$$

where in the second equality we use the Bayes formula together with the fact that the graph sequence is Markov, and in the third equality we use the induction hypothesis. Finally, observing that $\sum_{n:H_n \in \mathcal{S}} P_{\mathcal{S}n} P_{nm} = P_{\mathcal{S}m}$ proves the claim. □

Spectral radius of $P_{\mathcal{H}}$. We can see from Lemma 3.4, that as we let the time interval $s - t$ increase, the probability to move in a restricted class $\mathcal{S} \subseteq \mathcal{G}$ is essentially determined by the spectral radius of the submatrix of P corresponding to this restricted class, P_S[3]. This observation is formalized in Lemma 3.5.

[3]Reader familiar with the results from Chapter 2 could now guess that the restricted set $\mathcal{S}$ that we have in mind here is some disconnected collection $\mathcal{H} \in \Pi(\mathcal{G})$.

Lemma 3.5 Let $A \in \mathbb{R}^{N\times N}$ be a nonnegative matrix. For every $\varsigma > 0$, there exists C_ς such that for all $t \geq 1$

$$\rho(A)^t \leq 1^\top A^t\, 1 \leq C_\varsigma \left(\rho(A) + \varsigma\right)^t. \tag{3.8}$$

Proof We first observe the following

$$\|A^t\|_1 \leq 1^\top A^t\, 1 \leq N\|A^t\|_1,$$

where $\|\cdot\|_1$ denotes the 1 norm (for a nonnegative matrix equal to the maximum row sum of the matrix). The left hand side of the inequality (3.8) now easily follows from the fact that spectral radius is a lower bound for every matrix norm, and for the 1-norm, in particular:

$$\rho(A)^t = \rho(A^t) \leq \|A^t\|_1.$$

To prove the right hand side of (3.8), we recall Gelfand's formula, e.g., [78], which applied to the 1-norm states that

$$\lim_{t\to\infty} \|A^t\|_1^{\frac{1}{t}} = \rho(A).$$

The previous equation implies that for every $\varsigma > 0$ there exists $t_0 = t_0(\varsigma)$ such that $\|A^t\|_1 \leq \left(\rho(A) + \varsigma\right)^t$ for all $t \geq t_0$. Choosing $C_\varsigma = N \max\{1, \max_{1\leq t\leq t_0} \|A^t\|_1\}$, proves the right hand side of (3.8) and completes the proof of Lemma 3.5. □

Focusing now on the subsets of states $\mathcal{S}$ that are disconnected in union, i.e., on $\mathcal{S} \in \Pi(\mathcal{G})$, from the upper bound of Lemma 3.5 we derive the following Corollary.

Corollary 3.6 For each $\varsigma > 0$, there exists $\overline{C}_\varsigma > 0$ such that the following holds:

$$\sum_{\mathcal{H}\in\Pi(\mathcal{G})} 1^\top P_{\mathcal{H}}^t\, 1 \leq \overline{C}_\varsigma \left(\rho_{\max} + \varsigma\right)^t. \tag{3.9}$$

Proof Fix $\varsigma > 0$. By lemma 3.5, we know that for each $\mathcal{H} \in \Pi(\mathcal{G})$ there exists $C_{\mathcal{H},\varsigma}$ such that $1^\top P_{\mathcal{H}}^t \leq C_{\mathcal{H},\varsigma} \left(\rho\left(P_{\mathcal{H}}\right) + \varsigma\right)^t$. Thus, we have:

$$\begin{aligned}\sum_{\mathcal{H}\in\Pi(\mathcal{G})} 1^\top P_{\mathcal{H}}^t \, 1 &\leq \sum_{\mathcal{H}\in\Pi(\mathcal{G})} C_{\mathcal{H},\varsigma} \left(\rho\left(P_{\mathcal{H}}\right) + \varsigma\right)^t \\ &\leq |\Pi(\mathcal{G})| \max_{\mathcal{H}\in\Pi(\mathcal{G})} C_{\mathcal{H},\varsigma} \left(\max_{\mathcal{H}\in\Pi(\mathcal{G})} \rho\left(P_{\mathcal{H}}\right) + \varsigma\right)^t,\end{aligned}$$

and we see that equation (3.9) is satisfied with the constant $\overline{C}_\varsigma = |\Pi(\mathcal{G})| \max_{\mathcal{H}\in\Pi(\mathcal{G})} C_{\mathcal{H},\varsigma}$. □

Proof of the lower bound (3.5)**.**

We start from the fact, shown in Chapter 2 (see Lemma 2.8 and eq. (2.12))[4], that, if $\Gamma(k,0)$ is disconnected, then $\left\|\widetilde{\Phi}(k,0)\right\| \geq 1$, implying further $\left\|\widetilde{\Phi}(k,0)\right\| \geq d_k$. Thus, $\{\Gamma(k,0) \text{ is disconnected}\} \subseteq \left\{\left\|\widetilde{\Phi}(k,0)\right\| \geq d_k\right\}$, and passing to the probabilities,

$$\mathbb{P}\left(\left\|\widetilde{\Phi}(k,0)\right\| \geq d_k\right) \geq \mathbb{P}\left(\Gamma(k,0) \text{ is disconnected}\right).$$

From the claim in eq. (2.13) from Chapter 2, we further have that, for any fixed disconnected collection $\mathcal{H} \in \Pi(\mathcal{G})$,

$$\mathbb{P}\left(\Gamma(k,0) \text{ is disconnected}\right) \geq \mathbb{P}\left(G_t \in \mathcal{H},\ 1 \leq t \leq s\right).$$

Computing the probability in the right hand side by partitioning over all possible realizations of the initial and the final graph G_1 and G_k which belong to $\mathcal{H}$, and applying Lemma 3.4:

$$\begin{aligned}\mathbb{P}\left(G_t \in \mathcal{H},\ 1 \leq t \leq k\right) &= \sum_{l:H_l\in\mathcal{H}} \sum_{m:H_m\in\mathcal{H}} \mathbb{P}\left(G_t \in \mathcal{H},\ 2 \leq t \leq k-1,\ G_k = H_m \,|\, G_1 = H_l\right) \mathbb{P}\left(G_1 = H_l\right) \\ &= \sum_{l:H_l\in\mathcal{H}} \sum_{m:H_m\in\mathcal{H}} v_l P_{l\mathcal{H}} P_{\mathcal{H}}^{k-2} P_{\mathcal{H}m} \\ &= \left(\sum_{l:H_l\in\mathcal{H}} v_l P_{l\mathcal{H}}\right) P_{\mathcal{H}}^{k-2} \left(\sum_{m:H_m\in\mathcal{H}} P_{\mathcal{H}m}\right) \geq v_{\min} 1^\top P_{\mathcal{H}}^k 1,\end{aligned}$$

where in the last inequality we used that $\left(\sum_{m:H_m\in\mathcal{H}} P_{\mathcal{H}m}\right) = P_{\mathcal{H}}1$ and $\sum_{l:H_l\in\mathcal{H}} v_l P_{l\mathcal{H}} \geq v_{\min} 1^\top P_{\mathcal{H}}$. Combining the previous findings, and applying Lemma 3.5 to the matrix $P_{\mathcal{H}}$, yields

$$\mathbb{P}\left(\left\|\widetilde{\Phi}(k,0)\right\| \geq d_k\right) \geq v_{\min} \rho\left(P_{\mathcal{H}}\right)^k. \tag{3.10}$$

[4]Note that the claim in Lemma 2.8 holds in the point-wise sense for arbitrary sequence of realizations W_t, $t \geq 1$ (irrespective of the distribution of W_t's), hence it applies here.

Computing the logarithm, dividing by k, and passing to the limit:

$$\liminf_{k\to+\infty} \frac{1}{k} \log \mathbb{P}\left(\left\|\widetilde{\Phi}(k,0)\right\| \geq d_k\right) \geq \rho\left(P_{\mathcal{H}}\right). \tag{3.11}$$

Noting that the preceding inequality holds for arbitrary $\mathcal{H} \in \Pi(\mathcal{G})$, and thus for $\mathcal{H}^\star$ such that $\rho\left(P_{\mathcal{H}^\star}\right) = \rho_{\max}$, completes the proof.

3.2.3 Upper bound

Similarly as with the i.i.d. case from Chapter 2, the main tools in proving the upper bound will be the sequence of stopping times $T_{i,\delta}$ and the (random) number of improvements until time k, $M_{k,\delta}$. However, note that we have a slightly easier case here (modulo the complexity of the Markov chain setup), compared to the model from Chapter 2: here we assume that positive entries of all the realizations of W_t are bounded away from zero by some small number $\delta = \delta_0$, see part 2 of Assumption 3.2. As a result of this simplification, the proof of the upper bound here will escape from the technicalities from the proof of the corresponding upper bound in Chapter 2. Here, to prove the upper bound, we need only one member of the family of the δ-stopping times, $T_{i,\delta}$ and only one member of the family of δ-number of improvements, $M_{k,\delta}$: we take T_{i,δ_0}, $i \geq 1$, and M_{k,δ_0}. For simplicity, we will drop the index δ_0 in the sequel and use $T_i = T_{i,\delta_0}$, for $i = 1, ..., N$, and $M_k = M_{k,\delta_0}$, for $k \geq 1$.

Following the arguments from Chapter 2, it can be easily checked that part 1 of Lemma 2.15 holds as well for the model of W_t that we study[5]. Therefore, to prove the upper bound (3.6), it suffices to show that

$$\limsup_{k\to+\infty} \frac{1}{k} \log \mathbb{P}\left(M_k < \alpha k\right) \leq -|\rho_{\max}|. \tag{3.12}$$

We start from the following result which is the counterpart of part 2 of Lemma 2.15 from Chapter 2.

Lemma 3.7 For every $\varsigma > 0$ and $\lambda > 0$ such that $\lambda < |\log(\rho_{\max} + \varsigma)|$,

$$\mathbb{P}\left(M_k < \alpha k\right) \leq |\mathcal{G}|^{\lceil \alpha k \rceil} |\Pi(\mathcal{G})|^{\lceil \alpha k \rceil} \overline{C}_\varsigma^{\lceil \alpha k \rceil} e^{-\lambda k} \left(\frac{e^{(\lambda - |\log(\rho_{\max}+\varsigma)|)}}{1 - e^{(\lambda - |\log(\rho_{\max}+\varsigma)|)}}\right)^{\lceil \alpha k \rceil}, \tag{3.13}$$

where the constant $\overline{C}_\varsigma$ is the constant that verifies Corollary 3.6.

[5] In fact, part 1 of Lemma 2.15 holds for arbitrary deterministic sequence of stochastic symmetric matrices with positive diagonal. The reason why this is true is that, for sufficiently large k, there is no sequence of realizations W_1,..., W_k that satisfies both $\left\|\widetilde{\Phi}(k,0)\right\| \geq d_k$ and $M_k \geq \alpha k$, where $\alpha \in (0,1)$ is fixed – i.e., for sufficiently large k, $\left\{\left\|\widetilde{\Phi}(k,0)\right\| \geq d_k\right\} \cap \{M_k \geq \alpha k\} = \emptyset$.

Proof The key property which allows to derive the result in Lemma 3.7 is the following exponential bound for the joint probability distribution of the first M stopping times.

Lemma 3.8 For $M \geq 1$, $1 \leq t_1 < ... < t_M$ and for every ς

$$\mathbb{P}\left(T_1 = t_1, ..., T_M = t_M\right) \leq |\mathcal{G}|^M |\Pi(\mathcal{G})|^M \overline{C}_{\varsigma}^{M} \left(\rho_{\max} + \varsigma\right)^{t_M - 2M}. \tag{3.14}$$

Proof Using the definition of the stopping times T_m (see (2.23) in Chapter 2) we have that for every m the accumulation graph $\Gamma\left(T_{m-1} - 1, T_m\right)$ must be disconnected. Therefore,

$$\begin{aligned}\mathbb{P}\left(T_1 = t_1, \ldots, T_M = t_M\right) &\leq \mathbb{P}\left(\Gamma(t_m - 1, t_{m-1}) \text{ is disconnected, } m = 1, \ldots, M\right) \\ &\leq \sum_{\mathcal{H}_m \in \Pi(\mathcal{G}), m=1,\ldots,M} \mathbb{P}\left(G_{t_{m-1}+1} \in \mathcal{H}_m, \ldots, G_{t_m - 1} \in \mathcal{H}_m, \, m = 1, \ldots, M\right).\end{aligned}$$

Now, fix $\mathcal{H}_1, ..., \mathcal{H}_M \in \Pi(\mathcal{G})$ and consider the probability in the summation above corresponding to these fixed collections. Summing out all possible realizations of graphs at times $0, t_1, ..., t_M$ we get

$$\begin{aligned}&\mathbb{P}\left(G_{t_{m-1}+1} \in \mathcal{H}_m, \ldots, G_{t_m - 1} \in \mathcal{H}_m, \, m = 1, \ldots, M\right) \\ &= \sum_{H_{l_m} \in \mathcal{G}, \, m=0,1,\ldots,M} \mathbb{P}\left(G_0 = H_{l_0}, G_{t_{m-1}+1} \in \mathcal{H}_m, \ldots, G_{t_m - 1} \in \mathcal{H}_m, G_{t_m} = H_{l_m}, \, m = 1, \ldots, M\right)\end{aligned}$$

Applying the chain rule for probabilities on a fixed term in the previous summation:

$$\begin{aligned}&\mathbb{P}\left(G_0 = H_{l_0}, G_{t_{m-1}+1} \in \mathcal{H}_m, \ldots, G_{t_m - 1} \in \mathcal{H}_m, G_{t_m} = H_{l_m}, \, m = 1, \ldots, M\right) \\ &= \mathbb{P}\left(G_0 = H_{l_0}\right) \prod_{m=1}^{M} \mathbb{P}\left(G_{t_{m-1}+1} \in \mathcal{H}_m, \ldots, G_{t_m - 1} \in \mathcal{H}_m, G_{t_m} = H_{l_m} | \right. \\ &\qquad\qquad \left. G_{t_{l_{m-1}}} = H_{l_{m-1}}, G_{t_{m-1}-1} \in \mathcal{H}_{m-1}, ..., G_0 = H_{l_0}\right) \\ &= \mathbb{P}\left(G_0 = H_{l_0}\right) \prod_{m=1}^{M} \mathbb{P}\left(G_{t_{m-1}+1} \in \mathcal{H}_m, \ldots, G_{t_m - 1} \in \mathcal{H}_m, G_{t_m} = H_{l_m} | G_{t_{l_{m-1}}} = H_{l_{m-1}}\right),\end{aligned}$$

where the last inequality follows from the Markov property of the graph sequence. Applying the result of

Lemma 3.4, we get:

$$\begin{aligned}
&\mathbb{P}\left(G_0 = H_{l_0}, G_{t_{m-1}+1} \in \mathcal{H}_m, \ldots, G_{t_m-1} \in \mathcal{H}_m, G_{t_m} = H_{l_m},\, m = 1, \ldots, M\right) \\
&= v_{l_0} \cdot\ P_{l_0\mathcal{H}_1} P_{\mathcal{H}_1}^{t_1-2} P_{\mathcal{H}_1 l_1} \cdot\ P_{l_1\mathcal{H}_2} P_{\mathcal{H}_2}^{t_2-t_1-2} P_{\mathcal{H}_2 l_2} \cdot\cdot\cdot\ P_{l_M\mathcal{H}_{M+1}} P_{\mathcal{H}_M}^{t_M-t_{M-1}-2} P_{\mathcal{H}_M l_M} \\
&\leq 1^\top P_{\mathcal{H}_1}^{t_1-2} 1 \cdot\ 1^\top P_{\mathcal{H}_2}^{t_2-t_1-2} 1 \cdot\cdot\cdot\ 1^\top P_{\mathcal{H}_M}^{t_M-t_{M-1}-2} 1.
\end{aligned}$$

Finally, summing over all possible $H_{l_m}, \mathcal{H}_m$, $m = 1, ..., M$

$$\mathbb{P}\left(T_1 = t_1, \ldots, T_M = t_M\right) \leq |\mathcal{G}|^M \left(\sum_{\mathcal{H}_1 \in \Pi(\mathcal{G})} 1^\top P_{\mathcal{H}_1}^{t_1-2} 1\right) \cdots \left(\sum_{\mathcal{H}_M \in \Pi(\mathcal{G})} 1^\top P_{\mathcal{H}_M}^{t_M-t_{M-1}-2} 1\right),$$

and applying Corollary 3.6 to each of the terms in the product above yields the claim. □

To continue with the proof of Lemma 3.7, note first that $\{M_k < \alpha k\} = \left\{T_{\lceil \alpha k \rceil} > k\right\}$. Hence, from now on we focus on the probability $\mathbb{P}\left(T_{\lceil \alpha k \rceil} > k\right)$. Exponential Markov inequality with parameter $\lambda \geq 0$ applied to the stopping time $T_{\lceil \alpha k \rceil}$ yields:

$$\mathbb{P}\left(T_{\lceil \alpha k \rceil} > k\right) \leq \mathbb{E}\left[e^{\lambda T_{\lceil \alpha k \rceil}}\right] e^{-\lambda k} \tag{3.15}$$

To compute the expected value in (3.15), we consider all different increments $\Delta_i \geq 1$ of the first $\lceil \alpha k \rceil$ stopping times, $T_{i+1} - T_i = \Delta_i$, $i = 1, ..., \lceil \alpha k \rceil$. Note that for any fixed realization of increments Δ_i, $i = 1, ..., \lceil \alpha k \rceil$, $T_{\lceil \alpha k \rceil} = \Delta_1 + ... + \Delta_{\lceil \alpha k \rceil}$. Thus,

$$\mathbb{E}\left[e^{\lambda T_{\lceil \alpha k \rceil}}\right] = \sum_{\Delta_1 \geq 1, \ldots, \Delta_{\lceil \alpha k \rceil} \geq 1} e^{\lambda\left(\Delta_1 + \ldots + \Delta_{\lceil \alpha k \rceil}\right)} \mathbb{P}\left(T_1 = \Delta_1, T_2 - T_1 = \Delta_2, ..., T_{\lceil \alpha k \rceil} - T_{\lceil \alpha k \rceil - 1} = \Delta_{\lceil \alpha k \rceil}\right).$$

Applying then the bound from Lemma 3.8 to each probability in the sum above, where we note that M from Lemma 3.8 now equals $\lceil \alpha k \rceil$ and for any fixed Δ_i, $i = 1, \ldots, \lceil \alpha k \rceil$, t_M equals $\Delta_1 + \ldots + \Delta_{\lceil \alpha k \rceil}$, yields

$$\begin{aligned}
\mathbb{E}\left[e^{\lambda T_{\lceil \alpha k \rceil}}\right] &\leq \sum_{\Delta_1 \geq 1, \ldots, \Delta_{\lceil \alpha k \rceil} \geq 1} e^{\lambda\left(\Delta_1 + \ldots + \Delta_{\lceil \alpha k \rceil}\right)} |\mathcal{G}|^{\lceil \alpha k \rceil} |\Pi(\mathcal{G})|^{\lceil \alpha k \rceil} \overline{C}_\varsigma^{\lceil \alpha k \rceil} \left(\rho_{\max} + \varsigma\right)^{\Delta_1 + \ldots + \Delta_{\lceil \alpha k \rceil} - 2\lceil \alpha k \rceil} \\
&= |\mathcal{G}|^{\lceil \alpha k \rceil} |\Pi(\mathcal{G})|^{\lceil \alpha k \rceil} \overline{C}_\varsigma^{\lceil \alpha k \rceil} \sum_{\Delta_1 \geq 1, \ldots, \Delta_{\lceil \alpha k \rceil} \geq 1} e^{\sum_{i=1}^{\lceil \alpha k \rceil} (\lambda - |\log(\rho_{\max} + \varsigma)|) \Delta_i}.
\end{aligned}$$

Observe that the last sum can be represented as a product of equal terms, i.e.,

$$\sum_{\Delta_1\geq 1,\ldots,\Delta_{\lceil\alpha k\rceil}\geq 1} e^{\sum_{i=1}^{\lceil\alpha k\rceil}(\lambda-|\log(\rho_{\max}+\varsigma)|)\Delta_i} = \prod_{i=1}^{\lceil\alpha k\rceil}\left(\sum_{\Delta_i\geq 1} e^{(\lambda-|\log(\rho_{\max}+\varsigma)|)\Delta_i}\right).$$

Hence, we obtain

$$\mathbb{E}\left[e^{\lambda T_{\lceil\alpha k\rceil}}\right] \leq |\mathcal{G}|^{\lceil\alpha k\rceil}|\Pi(\mathcal{G})|^{\lceil\alpha k\rceil}\overline{C}_{\varsigma}^{\lceil\alpha k\rceil}\left(\sum_{\Delta_i\geq 1} e^{(\lambda-|\log(\rho_{\max}+\varsigma)|)\Delta_i}\right)^{\lceil\alpha k\rceil}.$$

Finally, recognizing that for every fixed λ and ς such that $\lambda < |\log(\rho_{\max}+\varsigma)|$

$$\sum_{\Delta_i\geq 1} e^{(\lambda-|\log(\rho_{\max}+\varsigma)|)\Delta_i} = \frac{e^{(\lambda-|\log(\rho_{\max}+\varsigma)|)}}{1-e^{(\lambda-|\log(\rho_{\max}+\varsigma)|)}},$$

proves (3.13). □

Having the result of Lemma 3.7, the proof of the upper bound (3.6) is now easy to complete. Computing the logarithm and dividing by k in both sides of (3.13)

$$\frac{1}{k}\log\mathbb{P}\left(M_k < \alpha k\right) \leq \frac{\lceil\alpha k\rceil}{k}\left(\log|\mathcal{G}| + \log\overline{C}_{\epsilon} - 2\left|\log\left(\rho_{\max}+\varsigma\right)\right| + \log\frac{e^{-(|\log(\rho_{\max}+\varsigma)|-\lambda)}}{1-e^{-(|\log(\rho_{\max}+\varsigma)|-\lambda)}}\right) - \lambda.$$

Taking first the limit as $k\to+\infty$, and then the infimum over $\alpha>0$ yields

$$\lim_{k\to+\infty}\frac{1}{k}\log\mathbb{P}\left(M_k\right) \leq -\lambda.$$

Since the last inequality holds for every $\lambda\geq 0$ and $\varsigma>0$ such that $\lambda<|\log\left(\rho_{\max}+\varsigma\right)|$, we have

$$\lim_{k\to+\infty}\frac{1}{k}\log\mathbb{P}\left(M_k\right) \leq \inf_{\varsigma>0}\inf_{0\leq\lambda<|\log(\rho_{\max}+\varsigma)|} -\lambda = -\left|\log\rho_{\max}\right|.$$

This proves (3.12) and thus completes the proof of the upper bound (3.6).

3.2.4 Examples

We now give two instances for the assumed Markov chain model and we compute the rate $\mathcal{J}$ for each of the given examples. The first example is a gossip-type averaging protocol with Markov dependencies, similar to the protocol in [17] (except that protocol in [17] corresponds to directed graphs). One particular instance

of this protocol is a random walk of a token along the edges of a given graph, according to a given transition probability matrix. In the second example, we consider a network with temporal correlations of the link failures, where we model the correlations by a Markov chain.

Gossip with Markov dependencies. Let $G = (V, E)$ be a connected graph on N vertices. We assume that at each time $t \geq 0$ only one link of G can be active; if $e = \{u, v\} \in E$ is active at time t, then $W_t = I_N - \frac{1}{2}(e_u - e_v)(e_u - e_v)^\top$. Thus, at each time t, the topology realization, G_t, is a one-link graph. The sequence of one-link graphs, G_t, $t \geq 0$, is generated according to a Markov chain:

$$\mathbb{P}(G_0 = (V, e)) = v_e, \text{ for } e \in E$$
$$\mathbb{P}(G_{t+1} = (V, f) | G_t = (V, e)) = P_{ef}, \text{ for } e, f \in E,$$

where $v_e > 0, P_{ef} \geq 0, \sum_{f \in E} P_{ef} = 1$, for each $e \in E$, and $\sum_{e \in E} v_e = 1$. The set of states of the Markov chain is therefore

$$\mathcal{G}^{\text{Gossip}} = \{(V, e) : e \in E\}$$

and there are $M = |E|$ states. A disconnected collection on $\mathcal{G}^{\text{Gossip}}$ is of the form $\{(V, e) : e \in E \setminus F\}$, for some set of edges F that disconnects G. Thus, the set of all disconnected collections on $\mathcal{G}^{\text{Gossip}}$ is

$$\Pi(\mathcal{G}^{\text{Gossip}}) = \{\mathcal{H}_F : F \text{ disconnects } G\}.$$

where $\mathcal{H}_F =:= \{(V, e) : e \in E \setminus F\}$, for $F \subseteq E$. By Theorem 3.3, we get the formula for $\rho_{\max}$:

$$\rho_{\max} = \max_{F \subseteq E:\ F \text{ disconnects } G} \rho(P_{\mathcal{H}_F}).$$

Computing $\rho_{\max}$ for this model is difficult in general, as it involves computing the spectral radius for all submatrices $P_{\mathcal{H}_F}$ of the transition matrix P associated with disconnected collections $\mathcal{H}_F$. A simple approximation for $\rho_{\max}$ can be obtained using the row-sum based lower bound for the spectral radius. We explain this next. For any fixed disconnected collection $\mathcal{H}_F$, we denote by $\underline{c}(P_{\mathcal{H}_F})$ the minimal row sum of its associated submatrix $P_{\mathcal{H}_F}$: $\underline{c}(P_{\mathcal{H}_F}) = \min_{i=1,\ldots,|\mathcal{H}_F|} \sum_{j=1}^{|\mathcal{H}_F|} [P_{\mathcal{H}_F}]_{ij}$ We then have, for any $\mathcal{H}_F$ [78]: $\underline{c}(P_{\mathcal{H}_F}) \leq \rho(P_{\mathcal{H}_F})$, implying

$$\max_{F \subseteq E:\ F \text{ disconnects } G} \underline{c}(P_{\mathcal{H}_F}) \leq \rho_{\max}. \tag{3.16}$$

In particular, for gossip on a tree, we get a very simple lower bound on $\rho_{\max}$ that involves no computations (it involves only $O(M^2)$ comparisons of certain entries of the matrix P.) When $G = (V, E)$ is a tree,

removal of any edge $f \in E$ disconnects G. Also, for any $F' \subseteq F \subseteq E$, the matrix $P_{\mathcal{H}_F}$ is a submatrix of $P_{\mathcal{H}_{F'}}$, and so $\underline{c}(P_{\mathcal{H}_F}) \leq \underline{c}(P_{\mathcal{H}_{F'}})$, i.e., the minimal row sum can only grow as the edges are removed from F. This implies that we can decrease the space of search in (3.16) to the set of edges of G:

$$\max_{F \subseteq E:\ F \text{ disconnects } G} \underline{c}(P_{\mathcal{H}_F}) = \max_{f \in E} \underline{c}(P_{\mathcal{H}_f}) \leq \rho_{\max}. \tag{3.17}$$

Now, for any fixed $f \in E$, since P is stochastic, it holds that $\underline{c}(P_{\mathcal{H}_f}) = 1 - \max_{e \in E \setminus f} P_{ef}$; that is, to compute the minimal row sum of $P_{\mathcal{H}_f}$, we only have to find the maximal entry of the column P^f, with entry P_{ff} excluded. This finally implies:

$$\rho_{\max} \geq \max_{f \in E} 1 - \max_{e \in E \setminus f} P_{ef} = 1 - \min_{f \in E} \max_{e \in E \setminus f} P_{ef}. \tag{3.18}$$

We can see an interesting phenomenon in the lower bound on $\rho_{\max}$ in eq. (3.18): when $\max_{e \in E \setminus f} P_{ef}$ is high for every link e, that is, when the gossip token is more likely to jump to a different link $f \neq e$, rather than to stay on the same link e ($P_{ef} >> P_{ee}$, for some $f \neq e$), the bound in eq. (3.18) has a small value. Assuming that $\rho_{\max}$ follows the tendency of its lower bound, we obtain a high rate $\mathcal{J}$ for this case of "negative correlations". This is in accordance with the intuition: if every link has a low probability P_{ee} to be repeated (repeating a link is a wasteful transmission in gossip), the convergence of gossip is faster and thus the rate $\mathcal{J}$ is higher.

Link failures with temporal correlations. Let $G = (V, E)$ be a connected graph on N vertices. For each $e \in E$ and $t \geq 0$, let $Y_{e,t} \in \{0, 1\}$ be a random variable that models the occurrence of the link e at time t: if $Y_{e,t} = 1$ then e is online at time t, and e is offline otherwise. For each link e, we assume that the failures of e occur in time according to a Markov chain. Also, the failures of different links are independent. More precisely, we assume that $Y_{e,t}$ and $Y_{f,s}$ are independent for all $t, s \geq 0$ if $e \neq f$, and, for $e \in E$ and $t \geq 1$:

$$\mathbb{P}(Y_{e,t+1} = 1 | Y_{e,t} = 1) = p_e,$$
$$\mathbb{P}(Y_{e,t+1} = 0 | Y_{e,t} = 0) = q_e,$$

$\mathbb{P}(Y_{e,0} = 1) = v_e$, for some $p_e, q_e, v_e \in (0, 1)$. In other words, the joint state of all the links in the network evolves according to the $|E|$ independent Markov chains, where each Markov chain determines the state of one link. Given the network realization G_t, the averaging matrix W_t can be chosen, e.g., as the Metropolis or an equal weight matrix [79].

We compute the rate $\mathcal{J}$ for this model, following the reasoning in the proof of Theorem 3.3, and exploit-

ing the decoupled single-link Markov chains. We first find the set of all network realizations at time t. Due to the independence in space of the link failures, and the fact that each link is on/off at time t with positive probability, the set of all network realizations at time t is the set of all subgraphs of G:

$$\mathcal{G}^{\text{Link fail.}} = \left\{(V, E') : E' \subseteq E\right\}.$$

Consider now a fixed disconnected collection $\mathcal{H}$ on $\mathcal{G}^{\text{Link fail.}}$ and let F be $\Gamma(\mathcal{H}) = E \setminus F$; note that F disconnects G. Then $\mathcal{H}$ is necessarily a subset of the (bigger) collection $\mathcal{H}_F = \{(V, E') : E' \subseteq E \setminus F\}$ and thus $\mathbb{P}(G_t \in \mathcal{H}, 0 \leq t \leq k) \leq \mathbb{P}(G_t \in \mathcal{H}_F, 0 \leq t \leq k)$. The latter implies that, in order to find the most likely $\mathcal{H}$ that determines the rate $\mathcal{J}$, we can search over the smaller set $\{\mathcal{H}_F : F \text{ disconnects } G\}$. Thus, we focus on the right hand side of the latter inequality:

$$\begin{aligned}\mathbb{P}(G_t \in \mathcal{H}_F,\ 0 \leq t \leq k) &= \mathbb{P}(Y_{e,t} = 0,\ \text{for } e \in F,\ 0 \leq t \leq k) \\ &= \prod_{e \in F} \mathbb{P}(Y_{e,t} = 0,\ 0 \leq t \leq k) = \prod_{e \in F} (1 - v_e) q_e^k; \end{aligned} \tag{3.19}$$

the second equality in (3.19) follows by the independence of failures of different links. The rate at which the probability in (3.19) decays is equal to $\sum_{e \in F} |\log q_e|$, and thus the rate $\mathcal{J}$ equals

$$\mathcal{J} = \mathcal{J}(\{q_e\}) = \min_{F \subseteq E:\ F \text{ disconnects } G} \sum_{e \in F} |\log q_e|. \tag{3.20}$$

Optimization problem in (3.20) is the minimum cut problem [16], with the cost of edge $e \in E$ equal to $|\log q_e|$. (Recall that q_e is the probability that the link e stays offline, given that in the previous time it was also offline.) Problem (3.20) is a convex problem, and there are efficient numerical algorithms to solve it, e.g., [16].

To get some intuition on the effect of temporal correlations, we let $q = q_e = p_e$, for all $e \in E$, i.e., all the links have the same symmetric 2×2 transition matrix. Note that $q = 1/2$ corresponds to the temporally uncorrelated link failures. When $q < 1/2$, a link *is more likely to change its state* (on/off) with respect to its state in the previous time ("negative correlation") than to maintain it. From (3.20), the rate $\mathcal{J}(q) > \mathcal{J}(1/2)$ for $q < 1/2$. We conclude that a "negative correlation" increases (improves) the rate. Likewise, a "positive correlation" ($q > 1/2$) decreases (degrades) the rate.

3.3 Rate $\mathcal{J}$ for directed networks

In this section we study convergence of products of *asymmetric* stochastic i.i.d. random matrices W_t. Problem setup and preliminaries are given in Subsection 3.3.1. The main result on the large deviation limit $\mathcal{J}$ in (3.2) is stated in Theorem 3.13 in Subsection 3.3.2 and subsequently proven in Subsection 3.3.3. Subsection 3.3.4 then illustrates with examples how to compute the rate $\mathcal{J}$.

3.3.1 Problem setup and preliminaries

Recall that $\mathbb{A}^N$ and $\mathbb{D}^N$ denote, respectively, the set of all stochastic matrices of size N by N and the set of all directed graphs (possibly with self-loops) on the set of vertices $V = \{1, ..., N\}$. Let W_t, $t \geq 1$, be a sequence of random matrices on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ taking values in $\mathbb{A}^N$. We assume the following on the sequence W_t.

Assumption 3.9 1. W_t, $t \geq 1$, are i.i.d.;

2. there exists $\delta > 0$ such that the following two conditions are satisfied:

   - diagonal entries of W_t are a.s. greater or equal to δ; that is, for every t, with probability 1 $[W_t]_{ii} \geq \delta$ for all i;
   - Whenever positive, the entries $[W_t(\omega)]_{ij}$ are a.s. greater or equal to δ; that is, for every t, if $\mathbb{P}\left([W_t(\omega)]_{ij} > 0\right) > 0$ for some i, j, then, conditioned on $\{[W_t(\omega)]_{ij} > 0\}$, $[W_t]_{ij} \geq \delta$ with probability 1.

Similarly as before, for $1 \leq t < s$, $\Phi(s, t)$ denotes the product of the matrices that occur from time $t + 1$ until time s, i.e., $\Phi(s, t) = W_s \cdots W_{t+1}$.

Thus, compared to the random model from Chapter 2, here we allow that matrices W_t have realizations that are *asymmetric*. It can also be seen that we make an additional assumption that positive entries of W_t are bounded away from zero. However, this assumption serves only to simplify the proofs and make them more intuitive. In particular, the same formula for the rate $\mathcal{J}$ holds even when this assumption is relaxed. We don't pursue the proof here, but we note that the same technique from Chapter 2 of using the *family* of improvements times $T_{i,\delta}$ indexed by δ (in the place of a single sequence of stopping times T_i), would yield the result.

Sequence of induced graphs. Introduce a graph operator $G : \mathbb{A}^N \mapsto \mathbb{D}^N$, and let G be defined by

$$G(W) = \left(V, \{(j, i) : W_{ij} > 0\}\right), \tag{3.21}$$

for $W \in \mathbb{A}^N$. Thus, for any $W \in \mathbb{A}^N$, $G(W)$ is the graph on the set of vertices $V = \{1, ..., N\}$ with the set of edges consisting of all (ordered) pairs (j, i) (possibly $i = j$) for which the corresponding entry W_{ij} is positive. Let now G_t, $t \geq 1$, be the sequence of random graphs such that $G_t = G(W_t)$ for each t, that is, G_t is the sequence of the induced (or, support) graphs of W_t's. Note that G_t are i.i.d., the matrices W_t being i.i.d.. Further, for a graph $H \in \mathbb{D}^N$, let p_H denote the probability that an induced graph G_t takes realization H, $p_H = \mathbb{P}(G(W_t) = H)$; for a collection of graphs $\mathcal{H} \subseteq \mathbb{D}^N$, we let $p_{\mathcal{H}} = \mathbb{P}(G(W_t) \in \mathcal{H}) = \sum_{H \in \mathcal{H}} p_H$. Set $\mathcal{G}$ collects all realizations of an induced graph that occur with positive probability:

$$\mathcal{G} := \left\{ H \in \mathbb{D}^N : p_H > 0 \right\}. \tag{3.22}$$

Accumulation graph $\Gamma(k, 0)$ and tree-free collections. For a collection of graphs $\mathcal{H} \subseteq \mathbb{D}^N$, let $\Gamma(\mathcal{H})$ denote the graph that contains all the edges of all the graphs in $\mathcal{H}$, $\Gamma(\mathcal{H}) = (V, \cup_{H \in \mathcal{H}} E(H))$, where $E(H)$ denotes the set of edges of a graph H. We call such a graph union graph (of the graphs in $\mathcal{H}$). With a slight abuse of notation, we use the same symbol Γ for the union of subsequent realizations of G_r over any given time window $1 \leq t < s$:

$$\Gamma(s, t) = \left(V, \bigcup_{r=t+1}^{s} E(G_r) \right), \tag{3.23}$$

in which case we call $\Gamma(s, t)$ the accumulation graph from time t until time s. Next paragraph recalls some concepts from directed graph theory that we will need in our study.

Let $H = (V, E)$ be a given directed graph on the set of vertices $V = \{1, ..., N\}$. A directed path in H is any sequence of nodes $(i_0, i_1, ..., i_L)$, $i_l \in V$, $l = 1, ..., L$, $L \geq 1$, such that $i_l \neq i_k$, for all $0 \leq k, l \leq L$, and, for each l, $(i_l, i_{l+1} \in E$; nodes i_0 and i_L are then, respectively, called initial and end node. Further, it is said that H has a directed spanning tree if there exists a node r, called a root node, such that r can reach every other node j in H. Node i is a neighbor of j in H if $(i, j) \in E$. A subset of nodes $C \subseteq V$ is called a strongly connected component of H if for every $i, j \in C$ both i can reach j in H and j can reach i in H; following the standard practice in the Markov chain literature, we will use also the term *communication class* when referring to a strongly connected component of a graph, see, e.g., [14, 80]. Correspondingly, for a node i that can reach a node j in H we will say that i can communicate to j in H. Finally, a communication class C of H (that is, a strongly connected component of H) is called initial if none of the nodes in C has a neighbor outside of C.

The first step in relating the matrix sequence W_t with the graphs sequence G_t is a simple, but important relation between $\Gamma(k, 0)$ and the induced graph of $\Phi(k, 0)$.

Lemma 3.10 For every $\omega \in \Omega$, and every $k \geq 1$:

1. $\Gamma(k,0)$ is a subgraph of $G(\Phi(k,0))$;

2. two nodes communicate in $\Gamma(k,0)$ if and only if they communicate in $G(\Phi(k,0))$.

Proof To show 1, suppose that $\Gamma(k,0)$ contains an edge (i,j). Then, it must be that $[W_t]_{ij} > 0$, for some $1 \leq t \leq k$. Using the positivity of the diagonal entries, this implies that $[\Phi(k,0)]_{ij} > 0$, showing that $(i,j) \in G(\Phi(k,0))$. TO prove 2 we need to show that for any ordered pair of nodes (i,j), there exists a directed path from i to j in $\Gamma(k,0)$ if and only if there exists a directed path from i to j in $G(\Phi(k,0))$. Now, from part 1 we know that $\Gamma(k,0)$ is a subgraph of $G(\Phi(k,0))$. Thus we only need to prove the sufficiency part. Fix a pair (i,j) and suppose that there exists a path from i to j in $G(\Phi(k,0))$. Then, by construction of $G(\Phi(k,0))$, it must be that $[\Phi(k,0)]_{ij} > 0$, implying that there must exist a sequence of nodes $i_k \equiv i$, i_{k-1},..., $i_0 \equiv j$ such that $[W_t]_{i_t i_{t-1}} > 0$, for $1 \leq t \leq k$. By construction of G_t, we then have $(i_t, i_{t-1}) \in E(G_t)$, implying that $(i_t, i_{t-1}) \in E(\Gamma(k,0))$, for all t. This shows that there exists a directed walk, and thus, a directed path from i to j in $\Gamma(k,0)$, completing the proof of the claim. □

Note that the induced graph of the product matrix, which contains both one-hop information flows and their superpositions in time, contains in general more links than the graph $\Gamma(k,0)$ that registers only one-hop information flows. However, Lemma 3.10 assures that the communication classes of these two graphs are nevertheless the same. We use this observation to derive the key relation between the product matrix $\Phi(k,0)$ and the accumulation graph $\Gamma(k,0)$ which we state in Lemma 3.11.

Lemma 3.11 $|\lambda_2(\Phi(k,0))| < 1$ if and only if $\Gamma(k,0)$ contains a directed tree.

Proof We use (without proof, which can be derived from Lemma (3.20) on page 224 in [80]) the fact that, for every stochastic matrix W with positive diagonals, $|\lambda_2(W)| < 1$ if and only if $G(W)$ has exactly one initial class. Combining this with Lemma 3.10, it follows that $|\lambda_2(\Phi(k,0))| < 1$ if and only $\Gamma(k,0)$ has exactly one initial class. Finally, the last condition is equivalent to the condition that $\Gamma(k,0)$ contains a directed tree. □

We say that a collection of directed graphs $\mathcal{H} \subseteq \mathbb{D}^N$ is tree-free if the union graph $\Gamma(\mathcal{H})$ does not contain a directed tree. Denote with $\Pi(\mathcal{G})$ the set of all tree-free collections $\mathcal{H}$ such that $\mathcal{H} \subseteq \mathcal{G}$:

$$\Pi(\mathcal{G}) = \{\mathcal{H} \subseteq \mathcal{G} : \Gamma(\mathcal{H}) \text{ is tree} - \text{free}\}. \tag{3.24}$$

We illustrate the introduced concepts of accumulation graph and tree-free collection on the example of broadcast gossip algorithm.

Example 3.12 (Broadcast gossip) Let $\widehat{G} = (V, \widehat{E})$ be a directed graph that collects all the available communication links in a network. At each time k, a node is chosen at random in V according to the probability mass function $p_u > 0$, $u \in V$, $\sum_{u \in V} p_u = 1$. Denote by u_k the node chosen at time k. With broadcast gossip [19], the weight matrix W_k has the sparsity pattern of a directed star graph centered at u_k and is given by: $[W_k]_{u_k v} = [W_k]_{vv} = 1/2$, for all v such that $\{u_k, v\} \in \widehat{E}$ (out-neighbors of u_k), $[W_k]_{vv} = 1$ otherwise, and the rest of the entries are zero. The graph realization at time k is then $G_k = (V, \left\{(u_k, v) : v \in V, (u_k, v) \in \widehat{E}\right\})$. Since each node in V has a positive probability of being chosen, we conclude that the collection of realizable graphs with broadcast gossip is the collection of all star subgraphs of $\widehat{G}$ centered at its nodes:

$$\mathcal{G}^{\mathrm{B-gossip}} = \{H_u : u \in V\}, \tag{3.25}$$

where $H_u = \left(V, \left\{(u, v) : v \in V, (u, v) \in \widehat{E}\right\}\right)$.

For concreteness, we consider now a simple case when $\widehat{G}$ is a four node graph, $V = \{1, 2, 3, 4\}$, and with the set of edges $\widehat{E} = \{(1,2), (2,1), (2,3), (3,2), (3,4), (4,3)\}$, as shown in Figure 3.1 a. We can

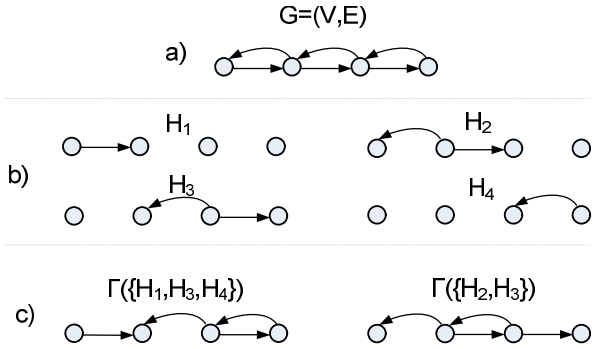

Figure 3.1: Example of a broadcast gossip on a 4-node chain; a) $\widehat{G} = (V, \widehat{E})$ is the total budget of communication links; b) $\mathcal{G} = \{H_1, H_2, H_3, H_4\}$ is the set of realizable graphs; c) $\mathcal{H} = \{H_1, H_3, H_4\}$ is a tree-free collection, whereas $\mathcal{H}' = \{H_2, H_3\}$ is not.

see that, for node 1, its only out-neighbor in $\widehat{G}$ is node 2 – thus, H_1 is the single arc graph $(V, (1,2))$, as shown in Figure 3.1 b. Similarly, node 2 has two out-neighbors in $\widehat{G}$, node 1 and node 3, and thus $H_2 = (V, \{(2,1), (2,3)\})$. Checking the out-neighborhoods of the remaining two nodes, 3 and 4, we conclude that the set of all realizable graphs is $\mathcal{G} = \{H_1, H_2, H_3, H_4\}$, with H_i as in Figure 3.1 b, $i = 1, ..., 4$. To find the tree-free collections on $\mathcal{G}$, we notice first that the removal of the edges of H_2

and H_3 makes $\widehat{G}$ tree-free. Thus, any subset of $\mathcal{G} \setminus \{H_2\}$ is a tree-free collection, and similarly for $\mathcal{G} \setminus \{H_3\}$; for example, $\mathcal{H} = \{H_1, H_3, H_4\}$ shown in Figure 3.1 c (left) is one such a collection (it is easy to see from Figure 3.1 c that $\Gamma(\{H_1, H_3, H_4\})$ does not have a directed spanning tree). On the other hand, simultaneous removal of edges of H_1 and H_4 "does no harm", as $\Gamma(\{H_2, H_3\})$ still contains a directed spanning tree (in fact, it contains two directed spanning trees, as can be seen from Figure 3.1 c (right)). Summing up, the set of all tree-free collections on $\mathcal{G}$ is $\Pi(\mathcal{G}) = 2^{\mathcal{G}\setminus\{H_2\}} \cup 2^{\mathcal{G}\setminus\{H_3\}}$, where $2^{\mathcal{G}\setminus\{H_2\}} = \{\emptyset, \{H_1\}, \{H_3\}, \{H_4\}, \{H_1, H_3\}, \{H_1, H_4\}, \{H_3, H_4\}, \ \{H_1, H_3, H_4\}\}$ is the power set of $\mathcal{G} \setminus \{H_2\}$, and similarly for $\mathcal{G} \setminus \{H_3\}$.

3.3.2 Main result

Theorem 3.9 states the main result on the large deviation rate $\mathcal{J}$ in (3.2).

Theorem 3.13 Let d_k be a sequence of real numbers such that $d_k \in (0, 1]$ and $\log d_k = o(k)$. Then

$$\lim_{k\to+\infty} \frac{1}{k} \log \mathbb{P}\left(|\lambda_2(\Phi(k, 0))| \geq d_k\right) = -\mathcal{J} \tag{3.26}$$

where

$$\mathcal{J} = \begin{cases} |\log p_{\max}|, & \text{If } \Pi(\mathcal{G}) \neq \emptyset \\ +\infty, & \text{otherwise} \end{cases}$$

and $p_{\max} = \max_{\mathcal{H}\in\Pi(\mathcal{G})} p_{\mathcal{H}}$.

In the proof of Theorem 3.13, our general approach is to follow the lines of the proof of Theorem 2.7 in Chapter 2, and then focus only on those steps that require a different argument. Compared to the setup from Chapter 2, we have here an additional difficulty which stems from the fact that the (almost sure) limit $\lim_{k\to 1} W_k \cdots W_1 = 1v^\top$ is no longer deterministic (vector v is random). If we go back to the proof of the upper bound (2.11) in Chapter 2, we can see that it was crucial to use the submultiplicativity of the spectral norm. Because we don't know the limit $1v^\top$, we can no longer use the spectral norm (nor any other matrix norm for that matter); on the other hand, $|\lambda_2(\cdot)|$ is not submultiplicative, i.e. $|\lambda_2(A_1A_2)|$ is not smaller than $|\lambda_2(A_1)||\lambda_2(A_2)|$ in general. Therefore, we need some other matrix function which is both submultiplicative and is able to measure the distance from the set of rank one matrices (to be able to detect the sequence $W_k \cdots W_1$ approaching to a point $1v^\top$ in this set). A class of matrix functions that have these two properties are ergodicity coefficients [14, 81]; for our purpose it will be convenient to use a specific

ergodicity coefficient τ, which we define next[6]

Coefficient of ergodicity τ and scrambling graphs. For $A \in \mathbb{A}^N$, let $\tau : \mathbb{A}^N \mapsto \mathbb{R}$ be defined by

$$\tau(A) = \max_{i,j} \frac{1}{2} \sum_{l=1}^{N} |A_{il} - A_{jl}|. \tag{3.27}$$

Lemma 3.14, borrowed from [81], asserts that the matrix function τ can indeed both detect a rank one matrix (see part 2) and has the submultiplicativity property (see part 4). In addition, we have that τ is always between zero and one (see part 1), and that τ is less than 1 only for (stochastic) matrices that have no two orthogonal rows. The proof of Lemma 3.14 can be found in [81] and is omitted here.

Lemma 3.14 (Properties of τ, [81]) For any $A, A_1, A_2 \in \mathbb{A}^N$:

1. $0 \leq \tau(A) \leq 1$;
2. $\mathrm{rank}(A) = 1$ if and only if $\tau(A) = 0$;
3. $\tau(A) < 1$ if and only if A has no two orthogonal rows (*"scrambleness"*);
4. $\tau(A_1 A_2) \leq \tau(A_1)\tau(A_2)$ (*submultiplicativity*).

Of special importance in our context is a property of τ by which τ upper bounds all eigenvalues of a matrix that are different than 1; hence we state this property in a separate lemma.

Lemma 3.15 (Spectral bound using τ, [81]) For any $A \in \mathbb{A}^n$

$$|\lambda| \leq \tau(A), \text{ for all eigenvalues } \lambda \text{ of } A \text{ s.t. } \lambda \neq 1. \tag{3.28}$$

A stochastic matrix whose coefficient τ is strictly smaller than 1 is called scrambling [14]. We can see from part 2 of Lemma 3.14 that the property of "scrambleness" is a purely topological property: a matrix $A \in \mathbb{A}^N$ is scrambling if for every two rows of A we can find a column in which both rows have a positive element. Motivated by this, following [82], we introduce the concept of a scrambling graph.

Definition 3.16 (Scrambling graph) A graph $H \in \mathbb{D}^N$ is called scrambling if for every $i, j \in V$ there exists $l \in V$ such that the set of edges of H contains both (l, i) and (l, j).

[6] We note that τ is often referred to as the Dobrushin or Markov-Dobrushin coefficient, see, e.g., [14].

In other words, a graph is scrambling if every two nodes in the graph have a common neighbor. Note that if a graph H has a scrambling subgraph then H must be scrambling. It is also easy to see that a stochastic matrix A is scrambling if and only if its induced graph $G(A)$ is scrambling. Using the preceding two observations together with part 1 of Lemma 3.10, it is easy to show that the following lemma holds.

Lemma 3.17 For any ω and s, t, $1 \leq t < s$,

$$\text{if } \Gamma(s,t) \text{ is scrambling then } \tau(\Phi(s,t)) < 1. \tag{3.29}$$

The property of "scrambleness" of a graph and that of the existence of a directed spanning tree are tightly related. First, a necessary condition for a graph to be scrambling is that it has a directed spanning tree. On the other hand, as [82] shows, it turns out that if we take any $N-1$ graphs such that each of the graphs has all the self-loops and contains a directed spanning tree then their union graph will be scrambling. These two claims are formally stated in Lemma 3.18.

Lemma 3.18

1. Any scrambling graph must contain a directed spanning tree.
2. If $A_1, ..., A_{N-1}$ have positive diagonal entries and each of their corresponding induced graphs $G(A_1), ..., G(A_N)$ has a directed spanning tree, then their product $A_1 \cdots A_{N-1}$ is scrambling[7].

Proof By the contraposition law, to prove part 1 it suffices to show the following implication: if a graph H has no directed spanning tree then H is not scrambling. Fix $H \in \mathbb{D}^N$ and suppose that H has no directed spanning tree. Then, H must have (at least) two initial classes, say C_1 and C_2. Fix $i \in C_1$ and $j \in C_2$; because class C_1 is initial, all neighbors of i are in C_1, and, similarly for j, all neighbors of j are in C_2. Finally, since $C_1 \cap C_2 = \emptyset$, it follows that i and j do not have a common neighbor proving that H is not scrambling. Since H was arbitrary, we proved part 1.

For the proof of part 2, see the proof of Theorem 5.1 in [82]. □

We summarize our findings so far. First, by Lemma 3.11 we have that $|\lambda_2(\Phi(k,0))| = 1$ as long as $\Gamma(k,0)$ is tree-free; in other words, the product $\Phi(k,0)$ stays put until a directed spanning tree emerges in $\Gamma(k,0)$. On the other hand, from part 2 of Lemma 3.18, a small step towards the set of rank one matrices is assured ($\tau(\Phi(k,0)) < 1$) each time a directed spanning tree emerges $N-1$ in a row. Therefore, we see that the "improvement" of the product is essentially determined by the number of spanning tree occurrence.

[7]Note that it is not required here that all the graphs $G(A_n)$ have the same directed tree; in particular, part 2 of Lemma 3.18 applies (even) in the case when each $G(A_n)$ has a different spanning tree.

Now, to close the loop, it only remains to bring the two metrics, $\|\lambda_2\|$ and τ, together. The following lemma does the job.

Lemma 3.19 For any $A \in \mathbb{A}^N$

$$|\lambda_2(A)| \leq \tau(A). \tag{3.30}$$

Therefore, an improvement in $\|\lambda_2\|$ is always guaranteed by an improvement in τ. It is easy to show that (3.30) holds: first, for any $A \in \mathbb{A}^N$ such that $\lambda_2(A) \neq 1$ the inequality in (3.30) is true by Lemma 3.15. Pick now A such that $\lambda_2(A) = 1$. Then A must have (at least) two initial classes, and therefore, $G(A)$ has no directed spanning tree. But then, by part 1 of Lemma 3.18), $G(A)$ is not scrambling, implying $\tau(A) = 1$ and inequality (3.30) trivially holds as an equality: $|\lambda_2(A)| = \tau(A) = 1$.

Having the above tools at hand we are now ready to prove Theorem 3.13.

3.3.3 Proof of Theorem 3.13

We first prove Theorem 3.13 for the case $\Pi(\mathcal{G}) \neq \emptyset$ by proving separately the corresponding large deviation upper and the lower bound:

$$\liminf_{k\to\infty} \frac{1}{k} \log \mathbb{P}\left(|\lambda_2(\Phi(k,0))| \geq d_k\right) \leq \log p_{\max} \tag{3.31}$$

$$\limsup_{k\to\infty} \frac{1}{k} \log \mathbb{P}\left(|\lambda_2(\Phi(k,0))| \geq d_k\right) \geq \log p_{\max}. \tag{3.32}$$

Proof of the lower bound (3.32)**.** If all induced graphs G_t from time $t = 1$ until time $t = k$ belong to some tree-free collection on $\mathcal{G}$, then it must be that $\Gamma(k,0) = \Gamma(G_1, ..., G_k)$ is tree-free. In other words, for any $\mathcal{H} \in \Pi(\mathcal{G})$, the following inclusion relation holds between the two events:

$$\{G_1 \in \mathcal{H}, ..., G_k \in \mathcal{H}\} \subseteq \{\Gamma(k,0) \text{ is tree} - \text{free}\}.$$

Computing now the probabilities of the events above and using the fact that G_t, $1 \leq t \leq k$, are i.i.d., it follows

$$\mathbb{P}\left(\Gamma(k,0) \text{ is tree} - \text{free}\right) \geq p_{\mathcal{H}}^k. \tag{3.33}$$

Since equation (3.33) holds for every $\mathcal{H} \in \Pi(\mathcal{G})$, it also holds for the tree-free collection $\mathcal{H}^\star \in \Pi(\mathcal{G})$ that has the highest probability $p_{\mathcal{H}^\star} = p_{\max}$, and therefore:

$$\mathbb{P}\left(\Gamma(k,0) \text{ is tree} - \text{free}\right) \geq p_{\max}^k. \tag{3.34}$$

To relate $p_{\max}^k$ and the probability of our event of interest $\{|\lambda_2(\Phi(k,0))| \geq d_k\}$, we observe that, because $d_k \in (0,1]$:

$$\mathbb{P}\left(|\lambda_2(\Phi(k,0))| \geq d_k\right) \geq \mathbb{P}\left(|\lambda_2(\Phi(k,0))| = 1\right). \tag{3.35}$$

Using now Lemma 3.11 (the "only-if" part) to link (3.35) and (3.34) yields

$$\mathbb{P}\left(|\lambda_2(\Phi(k,0))| \geq d_k\right) \geq p_{\max}^k.$$

Taking the $\log$, dividing by k and taking the $\liminf_{k\to+\infty}$, the large deviation lower bound follows.

Proof of the upper bound (3.31)**.** We start from the following intuitive observation: since δ Assumption 3.9 guarantees a minimal "flow of information" in each realization of W_t, then we must be able to find a minimal improvement in τ over all products $\Phi(s,t) = W_s \cdots W_{t+1}$ of (any) fixed size $s-t$. Lemma 3.20 confirms that this is true.

Lemma 3.20 For any ω, s, t, $1 \leq t < s$, if $\Phi(s,t)$ is scrambling, then

$$\tau(\Phi(s,t)) \leq 1 - \delta^{s-t}. \tag{3.36}$$

Proof Fix ω, s, t such that $\Phi(s,t)$ is scrambling. The proof is complete if we show that: 1) positive entries of $\Phi(s,t)$ bounded below by δ^{s-t}; and 2) for a scrambling matrix A, $\tau(A) \leq 1-a$, where a is the minimum over all positive entries of A. It is easy to show that 1) holds: if for some i,j $[\Phi(s,t)]_{ij} > 0$ then there must existed a sequence of nodes $i_{t+1} \equiv j, i_{t+2}, ..., i_s \equiv i$ such that $[W_r]_{i_{r+1}i_r} > 0$ for all r between $t+1$ and s. But, then, because all positive entries of W_{t+1},..., W_s are by assumption greater than δ, it must be that $[W_r]_{i_{r+1}i_r} > \delta$ for all r between $t+1$ and s; finally, claim 1) follows by noting that $[\Phi(s,t)]_{ij} \geq [W_s]_{ii_s}[W_s]_{i_{s-1}i_{s-2}} \cdots [W_s]_{i_{t+2}j} \geq \delta^{s-t}$. To prove 2), we use the fact that for any stochastic matrix A

$$\tau(A) = 1 - \min_{i,j} \sum_{l=1}^{N} \min\left\{A_{il}, A_{jl}\right\}, \tag{3.37}$$

see [14]. Thus, fix a stochastic matrix A and suppose that it is scrambling; then for any i,j there exists $l^\star = l^\star(i,j)$ such that both $A_{il^\star}$ and $A_{jl^\star}$ are positive. If a is the value of the minimal positive entry of A, then $\min\{A_{il^\star}, A_{jl^\star}\} \geq w_{\min}$ and thus $\sum_{l=1}^N \min\{A_{il}, A_{jl}\} \geq \min\{A_{il^\star}, A_{jl^\star}\} \geq a$. Since the last inequality holds for all pairs i,j, the minimum of the sum $\sum_{l=1}^N \min\{A_{il}, A_{jl}\}$ over all i,j, see (3.37), is also greater than a and the claim in 2) follows. □

Motivated by the result of Lemma 3.20, we introduce the sequence of stopping times S_i that registers

the times when $\Phi(s,t)$ becomes scrambling. For $\omega \in \Omega$ define $S_i : \Omega \mapsto \mathbb{N} \cup \{+\infty\}$, $i = 1, 2, \ldots$ is defined by

$$\begin{aligned} S_i(\omega) &= \min\{t \geq S_{i-1}(\omega) + 1 : \ \Phi(t, S_{i-1})(\omega) \text{ is scrambling}\}, \text{ for } i \geq 1, \\ S_0 &\equiv 0. \end{aligned} \tag{3.38}$$

We observe that, for every i, $\Phi(S_i, S_{i-1})$ is scrambling, which by submultiplicativity of τ implies $\tau\left(\Phi(S_i, 1)\right) \leq \tau\left(\Phi(S_{i-1}, 1)\right) < 1$; therefore, at each new time S_i an improvement is guaranteed with respect to the previous time S_{i-1}. For each $k \geq 1$, we then introduce $M_k : \Omega \mapsto \{1, ..., k\}$ to count the number of improvements made until time k:

$$M_k(\omega) = \max\left\{i \geq 0 : S_i(\omega) \leq k\right\}. \tag{3.39}$$

The following lemma derives a bound on the coefficient of ergodicity of the product at time k in terms of M_k. We can see that this result counterparts Lemma 2.14 from Chapter 2.

Lemma 3.21 For any fixed $\omega \in \Omega$ and $k \geq 1$,

$$\tau(\Phi(k, 0)) \leq \left(1 - \delta^{M_k}\right)^{M_k}. \tag{3.40}$$

Proof Fix an outcome ω and time $k \geq 1$. Let $m = M_k(\omega)$, $s_i = S_i(\omega)$, for $i = 1, ..., m$. Note first that $\tau(\Phi(k, 0)) = \tau(\Phi(k, s_m)\Phi(s_m, 0))$ is, by the submultiplicativity of τ, bounded by $\tau(\Phi(k, s_m))\tau(\Phi(s_m, 0))$. Further, because τ is always between zero and one (see property 1 in Lemma 3.14), the last number is further bounded by $\tau(\Phi(s_m, 0))$. Thus, we have $\tau(\Phi(k, 0)) \leq \tau(\Phi(s_m, 0))$. We now focus on computing $\tau(\Phi(s_m, 0))$. By the construction of the sequence S_i, $\Phi(s_i, s_{i-1})$ is scrambling for each i. Applying Lemma 3.20 to each of the intervals $(s_{i-1}, s_i]$, $i = 1, ..., m$ yields the set of inequalities $\tau(\Phi(s_i, s_{i-1})) \leq 1 - \delta^{s_i - s_{i-1}}$, $i = 1, ..., m$. Submultiplicativity of τ, applied to the consecutive disjoint intervals $(s_{i-1}, s_i]$, used together with the derived set of inequalities yields

$$\begin{aligned} \tau(\Phi(s_m, 0)) &= \tau\left(\Phi(s_m, s_{m-1}) \cdots \Phi(s_2, s_1)\Phi(s_1, 0)\right) \\ &\leq \left(1 - \delta^{s_m - s_{m-1}}\right) \cdots \left(1 - \delta^{s_2 - s_1}\right)\left(1 - \delta^{s_1}\right). \end{aligned} \tag{3.41}$$

To complete the proof, it only remains to consider the logarithm of both sides in (3.41) and then apply the Jensen's inequality to the function $f(\Delta) = \log\left(1 - \delta^{\Delta}\right)$ at the set of points $\Delta_i = s_i - s_{i-1}$ and the set

of convex multipliers $\alpha_i = \frac{1}{m}$, $i = 1, ..., m$; for details, we refer the reader to the proof of Lemma 2.14 in Chapter 2. □

Continuing with mimicking the arguments from Chapter 2, Lemma 2.15 in particular, we can see that if the number of improvements M_k is big enough (i.e., greater than αk, for some $\alpha \in (0, 1)$, as in Lemma 2.15), the coefficient of ergodicity of the products $\tau(\Phi(k, 0))$ cannot stay above d_k (recall here that d_k satisfies $\log d_k = o(k)$– thus, d_k decays slower than exponential). More precisely, it can be shown that there exists $k_0 = k_0(\alpha, \{d_k\})$ such that for all k greater than k_0 the two events below are disjoint:

$$\{M_{k,\delta} \geq \alpha k\} \bigcap \{\tau(\Phi(k, 0)) \geq d_k\} = \emptyset. \tag{3.42}$$

The proof of (3.42) is omitted, as it is the same as the proof of part 1 of Lemma 2.15 from Chapter 2. Relation in (3.42) together with the fact that τ upper bounds $|\lambda_2|$, see eq. (3.30), implies that for all k sufficiently large

$$\mathbb{P}\left(|\lambda_2(\Phi(k, 0))| \geq d_k,\ M_k \geq \alpha k\right) = 0, \tag{3.43}$$

and thus

$$\begin{aligned}\mathbb{P}\left(|\lambda_2(\Phi(k, 0))| \geq d_k\right) &= \mathbb{P}\left(|\lambda_2(\Phi(k, 0))| \geq d_k,\ M_k \geq \alpha k\right) + \mathbb{P}\left(|\lambda_2(\Phi(k, 0))| \geq d_k,\ M_k < \alpha k\right) \\ &\leq \mathbb{P}\left(M_k < \alpha k\right),\end{aligned}$$

for all k sufficiently large. We can see from the preceding inequality that that to prove 3.31 it suffices to prove

$$\limsup_{k \to +\infty} \frac{1}{k} \log \mathbb{P}\left(M_k < \alpha k\right) \leq \log p_{\max}. \tag{3.44}$$

Thus, from now on we focus on computing the probability of the event $\{M_k < \alpha k\}$. To do this, it will prove useful, in addition to recording the emergence of a scrambling graph in $\Phi(s, t)$, to record also the times when a directed spanning tree emerges in $\Gamma(s, t)$. For this purpose we use the sequence $T_i : \Omega \mapsto \mathbb{N} \cup \{+\infty\}$, $i \geq 1$, $i \geq 1$, defined by

$$\begin{aligned}T_i(\omega) &= \min\{t \geq T_{i-1}(\omega) + 1 :\ \Gamma(t, T_{i-1}(\omega)) \text{ has a directed spanning tree}\},\ \text{for } i \geq 1, \\ T_0 &\equiv 0,\end{aligned} \tag{3.45}$$

for $\omega \in \Omega$. Note that in each of the two sequences the increments, $T_{i+1} - T_i$ and $S_{i+1} - S_i$, $i \geq 0$, are i.i.d..

We have the following relations between the two sequences of times, the key behind which is Lemma 3.18.

Lemma 3.22 For any ω and $t \geq 1$:

1. if $T_1(\omega) > t$ then $S_1(\omega) > t$;

2. if $S_1(\omega) > t$ then $T_{N-1}(\omega) > t$.

Proof To prove Lemma 3.22, note first that $S_1 \leq t$ is equivalent to $G(\Phi(t,0))$ being scrambling. Part 1 then says that the necessary condition for $G(\Phi(t,0))$ to be scrambling is that it has a directed spanning tree. But, from part 1 of Lemma 3.18, we know that having a directed spanning tree is a necessary condition for any graph to be scrambling, thus the claim follows. To prove part 2, suppose that for some fixed ω and t $T_{N-1} \leq t$. Then, a directed spanning tree emerged at least $N-1$ times until time t, i.e., there exist times t_n, $0 \equiv t_0 \leq t_1 \leq ... \leq t_{N-1} \leq t$, such that $\Gamma(t_n, t_{n-1})$. Then, by Lemma 3.10 each $G\left(\Phi(t_n, t_{n-1})\right)$ must also have a directed spanning tree. Denote $A_n = \Phi(t_n, t_{n-1})$, $n \geq 1$; note that each A_n has positive diagonal entries and, thus, we are in the setup of Lemma 3.18. Observing that $\Phi(t,0) = A_{N-1} \cdots A_1$ and applying part 2 of Lemma 3.18 proves the claim. □

Passing to probabilities yields a neat "sandwiching" relation:

$$\mathbb{P}\left(T_1 > t\right) \leq \left(S_1 > t\right) \leq \mathbb{P}\left(T_{N-1} > t\right) \tag{3.46}$$

Next lemma provides an exponential bound in t for the probability of the event $\{T_{N-1} > t\}$ in the right hand side of (3.46). As we explain below, Lemma 3.23 is the last piece that was missing to prove the upper bound (3.31).

Lemma 3.23 For all $t \geq 2(N-1)$

$$\mathbb{P}\left(T_{N-1} > t\right) \leq (N-1)\binom{t}{N-1}|\Pi(\mathcal{G})|^{N-1} p_{\max}^{t-(N-1)}. \tag{3.47}$$

To simplify further the right hand side in (3.47), note that for any $\epsilon > 0$ we can find a sufficiently large constant C_ϵ such that $\binom{t}{N-1} e^{-t|\log p_{\max}|} \leq C_\epsilon e^{-t(|\log p_{\max}| - \epsilon)}$ for all t. Therefore, for every $\epsilon > 0$ and $t \geq 1$,

$$\mathbb{P}\left(S_1 > t\right) \leq (N-1)|\Pi(\mathcal{G})|^{N-1} p_{\max}^{-(N-1)} C_\epsilon e^{-t(|\log p_{\max}| - \epsilon)}. \tag{3.48}$$

Now, observe that, by definition of M_k, $\{M_k < \alpha k\} = \{S_{\alpha k} > k\}$; recall that for times $T_{i,\delta}$ from Chapter 2 we had an analogous relations with $M_{k,\delta}$, $\{M_{k,\delta} < \alpha k\} = \{T_{\alpha k,\delta} > k\}$ (see the paragraph before (2.35) in Chapter 2). Further, note the similarity between (3.48) with (2.32) from Chapter 2: they

both provide an exponential bound in t for the probabilities that the corresponding improvement times S_1 and $T_{1,\delta}$ are greater than t. Finally, both sequences S_i and $T_{i,\delta}$ have independent increments. It is easy to check that the proof of part 2 of Lemma 2.15 relied only on the the three listed features of the sequence $T_{i,\delta}$. Therefore, we can derive a counterpart to part 2 of Lemma 2.15 for $M_{k,\delta} = M_k$ and $T_{i,\delta} = S_i$ by simply reusing all the arguments of the corresponding proof: in all the formulas we only need to replace $\overline{\mathcal{J}}_\delta$ with $|\log p_{\max}| - \epsilon$. This will finally result in the bound below:

$$\limsup_{k\to+\infty} \frac{1}{k} \log \mathbb{P}\left(M_k < \alpha k\right) \leq -(|\log p_{\max}| - \epsilon).$$

Passing to the limit $\lim_{\epsilon\to 0}$, (3.44) follows. To complete the proof of the upper bound (3.31), it only remains to prove Lemma 3.23.

Proof [Proof of Lemma 3.23] Remark that $\{T_{N-1} > t\}$ occurs if and only if one of the following disjoint events occurs: $\{T_1 > t\}$, $\{T_1 \leq t \text{ and } T_2 > t\}$,..., $\{T_{N-2} \leq t \text{ and } T_{N-1} > t\}$. Thus,

$$\mathbb{P}\left(T_{N-1} > t\right) = \mathbb{P}\left(\bigcup_{l=1}^{N-2} \{T_n \leq t \text{ and } T_{n+1} > t\}\right) = \sum_{l=1}^{N-2} \mathbb{P}\left(T_n \leq t \text{ and } T_{n+1} > t\right). \tag{3.49}$$

For fixed n, we compute the probability $\mathbb{P}\left(T_n \leq t \text{ and } T_{n+1} > t\right)$ by considering all different realizations of $T_1, \ldots, T_n$. Suppose $T_1 = s_1$,...,$T_n = s_n$. Then, it must be that $\Gamma(s_l - 1, s_{l-1})$, $l = 1, ..., n$, are all tree-free; also, because $T_{n+1} > t$ it must be that $\Gamma(t, s_n)$ is tree-free. Using the fact that the graph realizations from disjoint time intervals are independent, we get

$$\begin{aligned}\mathbb{P}\left(T_1 = t_1, ..., T_n = t_n\right) &\leq \mathbb{P}\left(\Gamma(t, s_l) \text{ is tree} - \text{free, and } \Gamma(s_l - 1, s_{l-1}) \text{ is tree} - \text{free, for } l = 1, ..., n\right) \\ &\leq \mathbb{P}\left(\Gamma(t, s_l) \text{ is tree} - \text{free}\right) \prod_{l=1}^{n} \mathbb{P}\left(\Gamma(s_l - 1, s_{l-1}) \text{ is tree} - \text{free}\right). \end{aligned} \tag{3.50}$$

Now, we can show (similarly as in the proof of Lemma 2.16 in Chapter 2) that, up to a set of probability zero, $\Gamma(s_l - 1, s_{l-1})$ is tree $-$ free implies that all graph realizations in the time window $(s_{l-1}, s_l - 1]$ belong to some tree-free collection $\mathcal{H} \in \Pi(\mathcal{G})$; thus,

$$\mathbb{P}\left(\Gamma(s_l - 1, s_{l-1}) \text{ is tree} - \text{free}\right) \leq \sum_{\mathcal{H}\in\Pi(\mathcal{G})} p_{\mathcal{H}}^{s_l - 1 - s_{l-1}} \leq |\Pi(\mathcal{G})| p_{\max}^{s_l - 1 - s_{l-1}}. \tag{3.51}$$

Using the preceding bound in (3.50) (for each of the factors indexed by $l = 1, ..., n$ and also for the first

factor corresponding to the interval $(s_l, t]$) we get:

$$\mathbb{P}\left(T_1 = t_1, ..., T_n = t_n, T_{n+1} > t\right) \leq |\Pi(\mathcal{G})| p_{\max}^{t-s_n} \prod_{l=1}^{n} |\Pi(\mathcal{G})| p_{\max}^{s_l - 1 - s_{l-1}} = |\Pi(\mathcal{G})|^{n+1} p_{\max}^{t-n}.$$

The preceding bound holds for arbitrary choice of the realization t_1,...,t_n, and since there in total $\binom{t}{n}$ such possible choices, we have

$$\begin{aligned}\mathbb{P}\left(T_n \leq t \text{ and } T_{n+1} > t\right) &= \sum_{1 \leq t_1 \leq ... \leq t_n \leq t} \mathbb{P}\left(T_1 = t1, ..., T_n = t_n, T_{n+1} > t\right) \\ &\leq \binom{t}{n} |\Pi(\mathcal{G})|^{n+1} p_{\max}^{t-n}.\end{aligned}$$

Going back to the probability in (3.49):

$$\mathbb{P}\left(T_{N-1} > t\right) \leq \sum_{n=1}^{N-2} \binom{t}{n} |\Pi(\mathcal{G})|^{n+1} p_{\max}^{t-n};$$

bounding each factor in the product on the right hand side by its maximal value for n between 1 and $N-2$ yields (3.47) □

3.3.4 Computation of $\mathcal{J}$

In this subsection we illustrate computation of rate $\mathcal{J}$ with two random models: leader following with link failures and broadcast gossip.

Leader following with link failures. Let $T = (V, E)$ be a directed tree on N nodes and let r denote the root node in T. We assume that each link e in T may fail with probability $1 - p_e$ and also that links fail independently in time and in space. With leader following algorithm on T, every node in T at each time k transmits its state to all of its children; however, due to link failures, some if its children may not receive the sent information. The corresponding matrix of interactions W_k is then given by $[W_k]_{ut_u} = \alpha$ and $[W_k]_{uu} = 1 - \alpha$ if the node u received the information from its parent denoted by t_u (that is, if the link $(u, t_u) \in E$ was online at time k) and $[W_k]_{ut_u} = 0$ and $[W_k]_{uu} = 1$, otherwise; remark also that $[W_k]_{uv} = 0$ with probability one for all $v \neq t_u$. Using the fact that the links fail independently, we get that each subgraph of T occurs with positive probability and thus the collection of realizable graphs is the collection of all subgraphs of T:

$$\mathcal{G}^{\text{Leader}} = \left\{(V, E') : E' \subseteq E\right\}. \tag{3.52}$$

Now, since $T = (V, E)$ is a directed tree, the easiest way to make T tree-free is to remove an arbitrary arc from T. Thus, a candidate for the most likely tree-free collection on $\mathcal{G}^{\text{Leader}}$ has to be of the following form: $\{(V, E') : E' \subseteq E \setminus e\}$, for some $e \in E$. The probability of such a collection equals $1 - p_e$, $e \in E$. Thus, the one with the highest probability is the one for which p_e is minimal. We conclude that

$$p_{\max}^{\text{Leader}} = 1 - p_{e^\star}, \;\; \mathcal{J}^{\text{Leader}} = |\log(1 - p_{e^\star})|,$$

where $e^\star$ is the "weakest" link in T that has the lowest probability of occurrence.

Broadcast gossip on a tree. In the previous example we explained the broadcast gossip algorithm running on a generic network that is defined by graph $\widehat{G} = (V, \widehat{E})$. We consider now the case when $\widehat{G}$ is a symmetric tree (the undirected graph of $\widehat{G}$ is a tree and for every arc in $\widehat{G}$ its inverted arc also belongs to $\widehat{G}$). Note that the four node graph from the example in Figure 3.1 is of this type. Similarly as in the case of the four node graph, we can see that inhibition of any internal node (non-leaf node) $u \in V$ yields a tree-free network, that is, $\Gamma(\mathcal{G} \setminus \{H_u\})$ is tree-free. Thus, it suffices to remove just one graph of this type from $\mathcal{G}^{\text{B}-\text{gossip}}$, and therefore the most likely tree-free collection must be of the form $\mathcal{G} \setminus \{H_u\}$, where u is some internal node in $\widehat{G}$. The probability of such a collection is $1 - p_u$, and the most likely one is the one for which p_u is minimal:

$$\begin{aligned} p_{\max}^{\text{B}-\text{gossip}} &= 1 - \min_{u \in V : u \text{ is internal}} p_u \\ \mathcal{J}^{\text{B}-\text{gossip}} &= \left|\log p_{\max}^{\text{B}-\text{gossip}}\right|. \end{aligned}$$

For the simplest case when all the nodes have the same probability of activation, equal to $\frac{1}{N}$, the rate $\mathcal{J}^{\text{B}-\text{gossip}} = \left|\log\left(1 - \frac{1}{N}\right)\right| \sim \frac{1}{N}$, for large N.

Chapter 4

Large Deviations for Distributed Inference

4.1 Introduction

In this chapter, we establish large deviations upper and lower bounds for distributed inference algorithms over random networks (see ahead equations (4.3)–(4.4)) for d-dimensional nodes' vector observations, and arbitrary subsets of $\mathbb{R}^d$. Further, for regular networks, we establish the full large deviation principle in the sense of [8].

We explain distributed inference and our results in general terms; for specific applications we refer to Chapter 5. Suppose that each node i in an N-node network observes the samples $Z_{i,t} \in \mathbb{R}^d$ from an unknown distribution μ. The goal is for all nodes to estimate the mean $\overline{Z}$ of the distribution μ. For example, in linear estimation, $Z_{i,t} = \theta + n_{i,t}$, where $\theta \in \mathbb{R}^d$ is an unknown deterministic parameter, $n_{i,t}$ is the zero-mean noise, and $\overline{Z} = \theta$. To motivate our analysis, we briefly describe three types of inference: isolated, centralized, and distributed.

(1) Isolated inference. Node i is isolated from the rest of the network and estimates $\overline{Z}$ through the sample mean of its own observations:

$$x_{i,k}^{(\mathrm{iso})} = \frac{1}{k} \sum_{t=1}^{k} Z_{i,t}. \tag{4.1}$$

(2) Centralized inference assumes a (usually impractical) fusion node that collects the samples from all nodes, at all times, and computes:

$$x_{k}^{(\mathrm{cen})} = \frac{1}{N\,k} \sum_{t=1}^{k} \sum_{i=1}^{N} Z_{i,t}. \tag{4.2}$$

(3) Distributed inference. Each node i updates its estimate $x_{i,k}$ by communicating only with its imme-

diate neighbors in the network, via the following recursive algorithm:

$$\widehat{x}_{i,k} = \sum_{j \in O_{i,k}} W_{ij,k}\, x_{j,k-1} \tag{4.3}$$

$$x_{i,k} = \widehat{x}_{i,k} - \frac{1}{k}\left(\widehat{x}_{i,k} - Z_{i,k}\right). \tag{4.4}$$

Here, $O_{i,k}$ is node i's neighborhood at time k (including), $\widehat{x}_{i,k}$ is an auxiliary variable, and the $W_{ij,k}$'s are positive weights such that at any time k $\sum_{j \in O_{i,k}} W_{ij,k} = 1$ for each i. Distributed inference of type (4.3)–(4.4) has been recently extensively studied in the literature and proved useful in many contexts, numerous references are listed in Chapter 1. The scheme (4.3)–(4.4) has several favorable features: 1) it processes data in real-time, recursively; 2) it utilizes only inter-neighbor communication and avoids the fusion-node bottleneck; and 3) exhibits resilience to inter-node communication failures. The focus of our analysis is on distributed inference.

With all the three types of inference (isolated, centralized, and distributed), full characterization of performance requires knowledge of the full distribution of the random estimate (e.g., $x_{i,k}^{(\mathrm{iso})}$). Even for the simplest, isolated inference, this is usually an intractable task. Hence, one typically resorts to asymptotic measures, like consistency, asymptotic normality, and large deviations. For sample means, hence for the centralized and the isolated inference, all three measures (and, in particular, large deviations) have been studied in the literature and are well understood. Regarding distributed inference, consistency and asymptotic normality have been extensively studied [2]. However, large deviations analysis has not been addressed before. Our focus here is on the large deviations analysis. Specifically, consider a (measurable) set $E \subset \mathbb{R}^d$ that does not contain $\overline{Z}$. For consistent estimators, the probability $\mathbb{P}\left(x_{i,k} \in E\right)$ converges to zero as $k \to \infty$. Here, we find the *exponential decay rate* (the large deviation rate) $I(E)$ of this convergence:

$$\mathbb{P}\left(x_{i,k} \in E\right) \sim e^{-k\, I(E)}. \tag{4.5}$$

That is, we find the function $I(E)$ for any set $E \in \mathbb{R}^d$, and we quantify $I(E)$ in terms of the system parameters – the distribution of the $Z_{i,k}$'s, and the underlying network statistics. We achieve this for *randomly varying* networks, the case which is highly relevant, e.g., with wireless sensor networks, where the packet dropouts may occur at random times.

We give here qualitative explanation of our results, while quantitative statements are in Theorems 4.15 and 4.19. We discover interesting interplay between network and observations that arise from the large deviations approach, and which are not visible through the existing approaches of asymptotic normality

in [2]. First, we show that the performance of distributed inference (4.3)–(4.4) is always (equal or) better than isolated inference and is always worse than (or at best equal to) centralized inference. While the result is highly intuitive, it is technically very involved to prove it. Second, more interestingly, we show a highly nonlinear behavior of distributed inference. To make our point clear, consider the sets E of type

$$E_\delta = \{x \in \mathbb{R}^d : \|x - \overline{Z}\| > \xi\}, \quad \xi > 0.$$

Hence, requiring that the estimate $x_{i,k} \notin E_\xi$ for a very small ξ means requiring a very high estimation precision (high confidence); conversely, a large ξ corresponds to a coarse estimation. Our results show the following nonlinear behavior. Distributed inference is close to centralized performance for very high precisions (very small ξ's) and becomes much worse from the centralized performance for very coarse precisions. Intuitively, reaching high precisions is intrinsically difficult even for the centralized system, and hence the network-wide averaging process in (4.3)–(4.4) has sufficient time to "catch up" with the centralized system. On the other hand, the centralized system reaches a coarse accuracy very quickly, so that the distributed system cannot "catch up." The point $\xi^\star$ where the behavior significantly changes depends on the connectivity of the underlying network, number of nodes, and distribution of $Z_{i,k}$'s. We explicitly quantify this interplay between network and observations in Theorem 4.19.

Notation. We denote by $\mathbb{R}^N$ the N-dimensional real coordinate space; 1_N the vector of all ones in $\mathbb{R}^N$; I_N the identity matrix of size $N \times N$; e_i the i-th canonical vector of $\mathbb{R}^N$; and $\|\cdot\|$ the Euclidean (respectively, spectral) norm of its vector (respectively, matrix) argument; $\mathcal{B}\left(\mathbb{R}^d\right)$ the Borel σ-algebra on $\mathbb{R}^d$; Δ_N the unit simplex in $\mathbb{R}^N$: $\Delta_N = \left\{\alpha \in \mathbb{R}^N : \sum_{i=1}^N \alpha_i = 1,\ \alpha_i \geq 0 \text{ for } i = 1, \ldots, N\right\}$; and $1_{\mathcal{A}}$ the indicator of event $\mathcal{A}$. Further, for a set $E \subseteq \mathbb{R}^d$, E^o and $\overline{E}$ are, respectively, the interior and the closure of E. For a function $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$, we denote its domain by $\mathcal{D}_f := \left\{x \in \mathbb{R}^d : f(x) < +\infty\right\}$. We use notation $B_{x_0,\delta}$ for the closed ball $B_{x_0,\delta} := \{x \in \mathbb{R}^d : \|x - x_0\| \leq \delta\}$.

4.2 Model and distributed inference algorithm

In this section, we detail the distributed inference algorithm (4.3)–(4.4) and the network and observation models.

Distributed inference algorithm. Let $x_{i,k}$ denote the state of node i at time k. The state $x_{i,k}$ is improved over time two-fold: 1) the nodes' states are exchanged and averaged across the neighborhoods to form an intermediate state update $\hat{x}_{i,k}$; and 2) the new observations $Z_{i,k}$ are incorporated into the states.

Mathematically, the algorithm is given by (4.3)–(4.4), where we recall that $O_{i,k}$ is the (random) neighborhood of node i at time k (including i), and $W_{ij,k}$ is the weight that node i at time k assigns to the state of its neighboring node $j \in O_{i,k}$. We assume that the weights are nonnegative at all times. Also, at each node, and at each time, they form a convex combination, i.e., $\sum_{j \in O_{i,k}} W_{ij,k} = 1$. For each k, collect all the weights $W_{ij,k}$ in an $N \times N$ matrix W_k, such that the entry (i, j) of W_k, $[W_k]_{ij}$, takes value of the weight $W_{ij,k}$ for $j \in O_{i,t}$, and equals 0 otherwise.

Unwinding the recursion in (4.3)-(4.4), we rewrite the algorithm in a more compact form:

$$x_{i,k} = \frac{1}{k} \sum_{t=1}^{k} \sum_{j=1}^{N} [\Phi(k, t)]_{ij} \, Z_{j,t}, \tag{4.6}$$

where $\Phi(k, t) = W_k \cdots W_{t+1}$, for $1 \leq t < k$, and $\Phi(k, k)$ is the identity matrix (of size N).

We next state our assumptions on the joint distribution of the W_t's and the $Z_{i,t}$'s.

Assumption 4.1 (Network and observations model) We assume the following:

1. Observations $Z_{i,t}$, $i = 1, \ldots, N$, $t = 1, 2, \ldots$ are independent, identically distributed (i.i.d.), both across nodes and in time;

2. Random matrices W_t, $t = 1, 2, \ldots$ are i.i.d.;

3. Random matrix W_t takes values in the set of stochastic matrices[1];

4. W_t and $Z_{i,s}$ are independent for all i, s, t.

Model 4.1 on matrices W_t is very general. For example, we do not require here that the W_t's are doubly stochastic, nor do we require connectedness (in some sense) of the underlying network that supports the W_t's. Of course, to guarantee high benefits of inter-agent cooperation, we shall assume more structure on the W_t's; this is considered in Section 4.5.

4.3 Preliminaries

This section gives preliminaries by introducing certain large deviation tools needed in the sequel. Subsection 4.3.1 introduces the logarithmic moment generating function and its properties. Subsection 4.3.2 defines the large deviation principle and the Fenchel-Legendre transform. Finally, Subsection 4.3.3 gives the large deviation principle for centralized and isolated inference.

[1] With a stochastic matrix, rows sum to one, and all the entries are nonnegative

4.3.1 Logarithmic moment generating function

To analyze the large deviation performance of algorithm (4.3)–(4.4), we use the well-known tools, namely, the logarithmic moment generating function and Fenchel-Legedre transform. We first introduce the logarithmic moment generating function $\Lambda : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$, of the observations $Z_{i,t}$:

$$\Lambda(\lambda) = \log \mathbb{E}\left[e^{\lambda^\top Z_{i,t}}\right], \ \lambda \in \mathbb{R}^d. \tag{4.7}$$

In general, function Λ can assume values in $\mathbb{R} \cup \{+\infty\}$. We illustrate how to compute Λ when $Z_{i,t}$ is a discrete random vector.

Example 4.2 (Function Λ for a discrete random vector) Suppose that $Z_{i,t}$ is a discrete random vector that takes values in the set $\mathbb{A} = \{a_1, ..., a_L\}$, $a_l \in \mathbb{R}^d$, for $l = 1, ..., L$, according to the probability mass function $p \in \Delta_L$. Then, for any $\lambda \in \mathbb{R}^d$, the value $\Lambda(\lambda)$ is computed by

$$\Lambda(\lambda) = \log\left(\sum_{l=1}^{L} p_l e^{\lambda^\top a_l}\right). \tag{4.8}$$

It can be shown similarly that the logarithmic moment generating function of a d-dimensional Gaussian vector of mean vector m and covariance matrix S is the quadratic function $\mathbb{R}^d \ni \lambda \mapsto 1/2(\lambda - m)^\top S(\lambda - m)$.

We assume that function Λ is finite on whole $\mathbb{R}^d$.

Assumption 4.3 $\mathcal{D}_\Lambda := \left\{\lambda \in \mathbb{R}^d : \Lambda(\lambda) < \infty\right\} = \mathbb{R}^d$, i.e., $\Lambda(\lambda) < +\infty$ for all $\lambda \in \mathbb{R}^d$.

Assumption 4.3 holds for Gaussian vectors and arbitrary discrete random vectors (as we have just shown), but also for many other commonly used distributions; we refer the reader to Chapter 5 for examples beyond the ones here for which $\mathcal{D}_\Lambda = \mathbb{R}^d$.

Logarithmic moment generating function has many nice properties, like convexity and smoothness. They are listed in Lemma 4.4, for future reference.

Lemma 4.4 ([18]) The logarithmic moment generating function Λ of arbitrary random vector Z satisfies:

1. $\Lambda(0) = 0$ and $\nabla\Lambda(0) = \mathbb{E}[Z]$;
2. $\Lambda(\cdot)$ is convex;
3. $\Lambda(\cdot)$ is C^∞, i.e., it has continuous derivatives of all orders.

Proposition 4.5 states one implication of convexity of $\Lambda(\cdot)$ which is important for our analysis and which we frequently use in the sequel.

Proposition 4.5 Let α_i, $1 \leq i \leq N$, be an arbitrary N-tuple of convex multipliers, i.e., the α_i's satisfy $\alpha_i \geq 0$, for each i and $\sum_{i=1}^{N} \alpha_i = 1$. Then, for every $\lambda \in \mathbb{R}^d$:

$$N\Lambda\left(1/N\lambda\right) \leq \sum_{i=1}^{N} \Lambda(\alpha_i \lambda) \leq \Lambda(\lambda). \tag{4.9}$$

Proof We prove first the right hand side inequality in (4.9). To this end, fix $\alpha \in [0,1]$ and recall that $\Lambda(0) = 0$. Then, by convexity of $\Lambda(\cdot)$,

$$\Lambda(\alpha\lambda) = \Lambda(\alpha\lambda + (1-\alpha)0) \leq \alpha\Lambda(\lambda) + (1-\alpha)\Lambda(0) = \alpha\Lambda(\lambda).$$

Plugging $\alpha = \alpha_i$, $i = 1, ..., N$, and summing out the left hand sides and the right hand sides of the resulting i inequalities yields the claim. To prove the left hand side inequality in (4.9), consider the function $g : \mathbb{R}^N \mapsto \mathbb{R}$, $g(\beta) = \sum_{i=1}^{N} \Lambda(\beta_i \lambda)$, for $\beta = (\beta_1, ..., \beta_N) \in \mathbb{R}^N$. Function g is is a sum of convex functions $\mathbb{R}^N \ni \beta \mapsto \Lambda(\beta_i\lambda) = \Lambda(e_i^\top \beta\lambda)$ and this convex (note that each of these functions is convex as a composition of the linear map $e_i^\top \lambda$, for the corresponding i, and the convex function Λ). Therefore, we prove the left hand side inequality in (4.9) if we show that the minimum of $g(\cdot)$ over the unit simplex $\Delta_N := \{a \in \mathbb{R}^d : \sum_{i=1}^{d} a_i = 1,\ a_i \geq 0\, \forall i\}$ is attained at $1/N1_N = (1/N, \ldots, 1/N)$. We show this by verifying that there exists a multiplier $\nu \in \mathbb{R}$ such that the pair $(1/N1_N, \nu)$ satisfies the Karush-Kun-Tucker (KKT) conditions [83]. Let $L(\beta, \nu) = g(\beta) + \nu(1_N^\top \beta - 1)$. Then

$$\partial_{\beta_i} L(\beta, \nu) = \lambda^\top \nabla g(\beta_i) + \nu.$$

Taking $\nu = -\lambda^\top \Lambda(1/N)$ proves the claim. □

4.3.2 Rate function, the large deviation principle, and Cramér's theorem

In this Subsection, we review some concepts from large deviations theory. We start by the rate function I and the large deviation principle. We first informally introduce them and then give formal definitions. Consider a sequence of random vectors $\{Y_k\}$ and their measures $\mu_k : \mathcal{B}(\mathbb{R}^d) \to [0,1]$, $\mu_k(E) := \mathbb{P}(Y_k \in E)$, and suppose that $Y_k \to \overline{Y}$, in probability, where $\overline{Y}$ is a constant vector. Under certain conditions, for a

measurable set E that does not contain $\overline{Y}$, the probability mass vanishes from E exponentially fast, i.e.:

$$\mu_k(E) \sim e^{-k\,\widehat{I}(E)}, \tag{4.10}$$

where $\widehat{I}(E)$ is a nonnegative function that can further be expressed as: $\widehat{I}(E) = \inf_{x\in E} I(x)$, for some function $I : \mathbb{R}^d \to [0,\infty]$. The large deviation principle (LDP) formalizes (4.10) in the following two definitions.

Definition 4.6 (Rate function I) Function $I : \mathbb{R}^d \mapsto [0,+\infty]$ is called a *rate function* if it is lower semicontinuous, or, equivalently, when its level sets are closed.

Definition 4.7 (The large deviation principle) A sequence of measures μ_k on $\left(\mathbb{R}^d, \mathcal{B}\left(\mathbb{R}^d\right)\right)$, $k \geq 1$, is said to satisfy the large deviation principle with rate function I if the following two conditions hold:

- For any closed set $E \subseteq \mathbb{R}^d$,

$$\limsup_{k\to+\infty} \frac{1}{k} \log \mu_k(E) \leq -\inf_{x\in E} I(x); \tag{4.11}$$

- For any open set $F \subseteq \mathbb{R}^d$,

$$\liminf_{k\to+\infty} \frac{1}{k} \log \mu_k(F) \geq -\inf_{x\in F} I(x). \tag{4.12}$$

A set $E \subset \mathbb{R}^d$ for which $\inf_{x\in E^\circ} I(x) = \inf_{x\in \overline{E}} I(x)$ is called an I *continuity set*.

Cramér's theorem gives a canonical way to establish LDP and find the rate function I, when Y_k is a sample mean of i.i.d. random vectors $\mathcal{Z}_t$, i.e., $Y_k := \frac{1}{k}\sum_{t=1}^k \mathcal{Z}_t$. (Note that this is precisely the case with isolated (4.1) and centralized inference (4.2), but it is not the case with distributed inference (4.3)–(4.4).) Namely, the Cramér's theorem states that (the measures of) $\{Y_k\}$ satisfy LDP with the following rate function

$$I(x) = \sup_{\lambda\in\mathbb{R}^d} \lambda^\top x - \Lambda(\lambda),\ \ x \in \mathbb{R}^d, \tag{4.13}$$

where Λ is the logarithmic moment generating function of $\mathcal{Z}_t$; hence, the rate function I is the Fenchel-Legendre transform (conjugate) of Λ.

The Fenchel-Legendre transform I has several interesting properties that are relevant for our study. We state them here for future reference (proofs can be found in [8].)

Lemma 4.8 (Properties of I, [8]) Let I be the Fenchel-Legendre transform of the logarithmic moment generating function of a random vector Z with mean $\overline{Z}$. Then:

1. I is nonnegative;

2. $I\left(\overline{Z}\right) = 0$;

3. I is lower semicontinuous; if $0 \in \mathcal{D}_{\Lambda}^{0}$, then I has compact level sets.

4. I is convex.

Finally, a reader familiar with convex analysis will notice that I is (in the language of convex analysis) the *conjugate* of Λ, see, e.g., Chapter E in [72]. We will use both terms interchangeably.

4.3.3 Large deviation principle: Isolated and centralized inference

We now turn to the three inference algorithms from the introduction: $x_{i,k}^{(\mathrm{iso})}$ (isolated), $x_{k}^{(\mathrm{cen})}$ (centralized), and $x_{i,k}$ (distributed). We apply Cramér's theorem to establish the large deviation performance of the former two, while we analyze distributed inference is Sections 4.4 and 4.5.

(1) Isolated inference. Applying Cramér's theorem [8] to the sample mean in (4.1), we obtain that the sequence of measures $E \mapsto \mathbb{P}\left(x_{i,k}^{(\mathrm{iso})} \in E\right)$ satisfies the large deviation principle with the rate function I given by (4.13): Therefore, for a node working in isolation, its large deviations performance is fully characterized by the Fenchel-Legendre transform I of the logarithmic moment generating function Λ associated with its local samples $Z_{i,t}$.

(2) Centralized inference. Consider now the sample mean at the fusion node in (4.2). To apply Cramér's theorem to the sequence $x_{k}^{(\mathrm{cen})}$, we start from a simple observation that $x_{k}^{(\mathrm{cen})}$ is the sample mean of k samples of the random vector $1/N \sum_{i=1}^{N} Z_{i,t}$. Now, the logarithmic moment generating function of $1/N \sum_{i=1}^{N} Z_{i,t}$ is $\lambda \mapsto N\Lambda\left(1/N\lambda\right)$ (this can be easily shown using the independence of $Z_{i,t}$, $i = 1, ..., N$), and its Fenchel-Legendre transform is given by:

$$\sup_{\lambda \in \mathbb{R}^d} \lambda^{\top} x - N\Lambda(1/N\lambda) = N \sup_{\lambda \in \mathbb{R}^d} \left(1/N\lambda\right)^{\top} x - \Lambda(1/N\lambda) = NI(x), \;\; x \in \mathbb{R}^d.$$

Applying Cramér's theorem to $x_{k}^{(\mathrm{cen})}$, we obtain that the sequence of measures $E \mapsto \mathbb{P}\left(x_{i,k}^{(\mathrm{cen})} \in E\right)$, $E \in \mathcal{B}\left(\mathbb{R}^d\right)$, satisfies the large deviation principle with the rate function $N\,I(\cdot)$, where I is given in (4.13). The advantage of the fusion node over an isolated node in terms of large deviations is now obvious: for any "deviation" set E, $\overline{Z} \notin E$, the probability that the sample mean of the fusion node falls in E, $\mathbb{P}\left(x_{k}^{(\mathrm{cen})} \in E\right)$, decays at an N *times higher rate* with time k than the corresponding probability for $x_{k}^{(\mathrm{iso})}$.

To explain important practical implications of the conclusion above, suppose that we are a given a target accuracy for the estimates of $\overline{Z}$ defined in terms of a Borel set $S \subseteq \mathbb{R}^d$ containing $\overline{Z}$, so that we are satisfied

if, say, with probability $p^{\text{conf}} = 0.95$, our estimate belongs to S. (We allow S to be an arbitrary Borel set that contains $\overline{Z}$.) We now exploit the conclusions from above for the Borel set $E = \mathbb{R}^d \setminus S = S^{\text{c}}$. Recall the definition of the large deviation principle in (4.7) and, with slight abuse of notation, let $I(E) = \inf_{x \in E} I(x)$. Then, assuming that I is continuous on E[2] and ignoring the factors that are slower than exponential[3], we have:

$$\mathbb{P}\left(x_k^{(\text{iso})} \notin S\right) \approx e^{-kI(S^{\text{c}})}, \tag{4.14}$$

and

$$\mathbb{P}\left(x_k^{(\text{cen})} \notin S\right) \approx e^{-kNI(S^{\text{c}})}. \tag{4.15}$$

The earliest time when the estimate $x_k^{(\text{iso})}$ meets the accuracy requirements is the minimal integer k for which

$$\mathbb{P}\left(x_k^{(\text{iso})} \notin S\right) \leq 1 - p^{\text{conf}}.$$

Denote this number by $T^{\text{iso}}_{p^{\text{conf}}}$. Then, exploiting (4.14), we have

$$T^{\text{iso}}_{p^{\text{conf}}} = \left\lceil -\log(1 - p^{\text{conf}})/I(S^{\text{c}}) \right\rceil.$$

Computing the corresponding time for $x_k^{(\text{cen})}$,

$$T^{\text{cen}}_{p^{\text{conf}}} = \left\lceil -\log(1 - p^{\text{conf}})/NI(S^{\text{c}}) \right\rceil.$$

Therefore, the fusion node hits the target set S with high probability p^{conf} N times sooner than an isolated node:

$$T^{\text{cen}}_{p^{\text{conf}}} = \frac{1}{N}\, T^{\text{iso}}_{p^{\text{conf}}}.$$

Moreover, the same ratio between the two times holds for arbitrary design parameters (S, p^{conf}).

(3) Overview of the large deviation results for distributed inference (4.3)–(4.4). Having the LDP with the rate function $I(\cdot)$ for isolated inference and the LDP with the rate function $N\, I(\cdot)$ for a fusion node, it is natural to ask if the LDP, or at least an exponential decay, occurs also with distributed inference (4.3)–(4.4). We answer the above question affirmatively. We detail the results in Sections 4.4 and 4.5, while here we summarize our findings. First, Section 4.4 shows that the performance of distributed inference is

[2]Note that this is satisfied for any $E \subseteq \mathcal{D}_I^0$.

[3]To be more precise, in each of the two equations there exists a constant C_k multiplying the exponential function; however, this constant goes to zero on the exponential scale, $\lim_{k \to +\infty} \log C_k$, and it can be therefore neglected for large k; i.e., $\log \mathbb{P}\left(x_k^{(\text{iso})} \notin S\right) = \log C_k - kI(S^{\text{c}}) \approx kI(S^{\text{c}})$, for k large enough, and similarly for $\mathbb{P}\left(x_k^{(\text{cen})} \notin S\right)$

always bounded between the performance of isolated and centralized algorithms. Although highly intuitive, this result was surprisingly difficult to prove. We remark that this conclusion holds for *arbitrary sequence of stochastic matrices* W_t, $t \geq 1$, for example, with arbitrary correlations among the W_t's, and/or when, e.g., W_t's are permutation matrices. We address the case of more structured W_t's in Section 4.5. Under a "balanced" exchange of information, where each node gives a positive weight to its own opinion, and the matrices are *doubly stochastic*, cooperation guarantees much larger gains in performance. First, for regular networks, we establish the LDP with (4.3)–(4.4), and we provide a closed form formula for the rate function. The formula shows that the rate depends only on the number of nodes N, a single node's rate function I, and the probability that a node is isolated. Further, when the algorithm runs on a graph with i.i.d. random link failures, the LDP holds for every leaf node and, moreover, each leaf node has the same rate function. To explain our third finding, we recall from Chapter 2 the large deviation rate $\mathcal{J}$ for the products $W_k \cdots W_1$. Now, we show that, whenever $\mathcal{J}$ equals $\log |p_{i^\star,\mathrm{isol}}|$, for some node $i^\star$, where $p_{\mathrm{i}^\star,\mathrm{isol}}$ is the probability that node $i^\star$ has no neighbors, then the probability distribution of this node's estimate $x_{i^\star,k}$ satisfies the LDP.

For more general cases, we establish tight bounds $\overline{I}_i$ and $\underline{I}_i$ on the exponential decay of the sequence of measures $E \mapsto \mathbb{P}\left(x_{i,k} \in E\right)$, $E \in \mathcal{B}\left(\mathbb{R}^d\right)$:

$$\limsup_{k\to+\infty} \frac{1}{k} \log\left(x_{i,k} \in E\right) \leq -\underline{I}_i(E) \tag{4.16}$$

$$\liminf_{k\to+\infty} \frac{1}{k} \log\left(x_{i,k} \in E\right) \geq -\overline{I}_i(E), \tag{4.17}$$

and we explicitly characterize these bounds in terms of the number of nodes N, the single node's rate function I, and the statistics of the graphs that support W_t's.

4.4 Large deviations rate for distributed inference: Generic matrices W_t

The main result of this section, Theorem 4.15, asserts that, for any Borel set E, the bounds in (4.16) and (4.17) hold with $\overline{I}_i \equiv NI$ and $\underline{I}_i \equiv I$. Before giving Theorem 4.15, we state and prove the key technical lemma behind this result, Lemma 4.9.

Lemma 4.9 Consider a family of random vectors $Z_{i,t}$ satisfying Assumptions 4.1 and 4.3. Let $\alpha_t = (\alpha_{1,t}, \ldots, \alpha_{N,t})$, $t \geq 1$ be a given sequence of sets of N convex multipliers, i.e., $\alpha_t \in \Delta_N$ for all t. Then, the following holds for the sequence $x_k = \frac{1}{k}\sum_{t=1}^{k}\sum_{i=1}^{N} \alpha_{i,t} Z_{i,t}$, $k \geq 1$:

1. **(No worse than isolation)** For any closed set $E \subseteq \mathbb{R}^d$:

$$\limsup_{k\to+\infty} \frac{1}{k} \log \mathbb{P}\left(x_k \in E\right) \leq -\inf_{x\in E} I(x); \tag{4.18}$$

2. **(No better than centralization)** For any open set $F \subseteq \mathbb{R}^d$:

$$\liminf_{k\to+\infty} \frac{1}{k} \log \mathbb{P}\left(x_k \in F\right) \geq -N \inf_{x\in F} I(x). \tag{4.19}$$

4.4.1 Proof of the "no worse than isolation" bound

To prove (4.18), it suffices to show that for any open ball $G \in \mathbb{R}^d$:

$$\mathbb{P}\left(x_k \in G\right) \leq e^{-k \inf_{x\in G}\left(\lambda^\top x - \Lambda(\lambda)\right)}. \tag{4.20}$$

The upper bound (4.18) can then be established by applying the standard "finite cover" argument for the case of compact E (see, e.g., the proof of the Cramer's theorem in $\mathbb{R}^d$, [8]), combined with the exponential tightness of the sequence of measures $E \mapsto \mathbb{P}\left(x_k \in E\right)$.

The proof of (4.20) is based on two key arguments: exponential Markov inequality [69], and the property of $\Lambda(\cdot)$, proven in Proposition 4.5, that for any λ, and any set of convex multipliers $\alpha \in \Delta_N$, $\Lambda(\lambda)$ upper bounds $\sum_{i=1}^N \Lambda\left(\alpha_i \lambda\right)$.

Fix an arbitrary measurable set $G \subseteq \mathbb{R}^d$. Then, for any $\lambda \in \mathbb{R}^d$, the following statement holds point-wise

$$1_{x_k \in G} \leq e^{k\lambda^\top x_k - k \inf_{x\in G} \lambda^\top x}; \tag{4.21}$$

taking the expectation, yields

$$\mathbb{P}\left(x_k \in G\right) \leq e^{-k \inf_{x\in G} \lambda^\top x} \mathbb{E}\left[e^{k\lambda^\top x_k}\right]. \tag{4.22}$$

We now focus on the right hand side of (4.22). Using the fact that the $Z_{i,t}$'s are independent, together with the definition of the logarithmic moment generating function of $Z_{i,t}$ in (4.7),

$$\mathbb{E}\left[e^{k\lambda^\top x_k}\right] = \mathbb{E}\left[e^{\sum_{t=1}^k \sum_{i=1}^N \alpha_{i,t} \lambda^\top Z_{i,t}}\right] = e^{\sum_{t=1}^k \sum_{i=1}^N \Lambda(\alpha_{i,t}\lambda)}, \tag{4.23}$$

and applying the upper bound from Proposition 4.5, yields (4.20):

$$\mathbb{P}\left(x_k \in G\right) \leq e^{-k\left(\inf_{x\in G} \lambda^\top x - \Lambda(\lambda)\right)}. \tag{4.24}$$

4.4.2 Proof of the "no better than centralization" bound

We prove part 2 of Lemma 4.9 following the general lines of the proof of the Gärtner-Ellis theorem lower bound, see [8]. However, as we will see later in this proof, we encounter several difficulties along the way which force us to depart from the standard Gärtner-Ellis method and use different arguments. The main reason for this is that the sequence of the (scaled) logarithmic moment generating functions of x_k (see ahead (4.26)) does not have a limit in general (that is, for any sequence α_t of convex multipliers). Nevertheless, with the help of Proposition 4.5, we will be able to "sandwich" each member of this sequence between $\Lambda(\cdot)$ and $N\Lambda\left(1/N\cdot\right)$. This is the key ingredient that allows us to derive the lower bound in (4.19).

First, remark that to prove (4.19), it suffices to show that for any $z \in \mathcal{D}_I$,

$$\lim_{\delta\to 0} \liminf_{k\to+\infty} \frac{1}{k} \log \mathbb{P}\left(x_k \in B_{z,\delta}\right) \geq -NI(z). \tag{4.25}$$

To see this, observe that for an open set F and a point $z \in F$ we can find a small neighborhood B_{z,δ_0} that is fully contained in F. Then, for all $\delta \leq \delta_0$

$$\mathbb{P}\left(x_k \in F\right) \geq \mathbb{P}\left(x_k \in B_{z,\delta}\right),$$

implying

$$\liminf_{k\to+\infty} \frac{1}{k} \log \mathbb{P}\left(x_k \in F\right) \geq \lim_{\delta\to+\infty} \liminf_{k\to+\infty} \frac{1}{k} \log \mathbb{P}\left(x_k \in B_{z,\delta}\right).$$

Using now (4.25) to bound the righthand side of the preceding inequality,

$$\liminf_{k\to+\infty} \frac{1}{k} \log \mathbb{P}\left(x_k \in F\right) \geq -NI(z),$$

and taking the supremum over all $z \in F$, proves that (4.25) is a sufficient condition for (4.19) to hold. Thus, from now on we focus on proving (4.25).

We introduce a normalized logarithmic moment generating function $\Lambda_k : \mathbb{R}^d \mapsto \mathbb{R} \cup \{+\infty\}$ of x_k,

defined by:

$$\Lambda_k(\lambda) := \frac{1}{k} \log \mathbb{E}\left[e^{k\lambda^\top x_k}\right], \ \ \lambda \in \mathbb{R}^{d4}. \tag{4.26}$$

Using independence of the $Z_{i,t}$'s, we have that, for any λ,

$$\Lambda_k(\lambda) = \frac{1}{k} \sum_{t=1}^{k} \sum_{i=1}^{N} \Lambda\left(\alpha_{i,t}\lambda\right). \tag{4.27}$$

We now depart from the standard Gärtner-Ellis method and use a regularization of x_k with a Gaussian variable of vanishing probability, see Exercise 2.3.20 in [8] [5]. To this end, introduce a standard multivariate Gaussian variable V independent of the observations $Z_{i,t}$. Fix $M > 0$ and define $y_k = x_k + V/\sqrt{M}k$ for $k \geq 1$. Introduce the normalized logarithmic moment generating function of y_k:

$$\Lambda_{k,M}(\lambda) := \frac{1}{k} \log \mathbb{E}\left[e^{k\lambda^\top y_k}\right].$$

Using that the logarithmic moment generating function of a Gaussian vector of zero mean and covariance matrix Σ is $\frac{1}{2}\lambda^\top \Sigma \lambda$, after simple algebraic manipulations, we obtain:

$$\Lambda_{k,M}(\lambda) = \Lambda_k(\lambda) + \frac{\|\lambda\|^2}{2M}. \tag{4.28}$$

Thus, we can see that adding a small Gaussian noise $V/\sqrt{M}k$ to x_k brings a quadratic term to the (normalized) logarithmic moment generating function Λ_k. We note that both $\Lambda_{k,M}$ and Λ_k are convex and differentiable functions, and with continuous gradients. In addition, $\Lambda_{k,M}$ is strictly convex.

The first step towards proving (4.25) is to show its counterpart (4.29) for the regularized sequence y_k. For each k, let μ_k denote the distribution of y_k: $\mu_k(E) = \mathbb{P}\left(y_k \in E\right)$, $E \in \mathcal{B}\left(\mathbb{R}^d\right)$.

Lemma 4.10 For any $z \in \mathcal{D}_I$,

$$\lim_{\delta \to 0} \liminf_{k \to +\infty} \frac{1}{k} \log \mu_k\left(B_{z,\delta}\right) \geq -NI(z). \tag{4.29}$$

Proof Introduce the conjugate $I_{k,M}$ of $\Lambda_{k,M}$,

$$I_{k,M}(x) = \sup_{\lambda \in \mathbb{R}^d} \lambda^\top x - \Lambda_k(\lambda) - \frac{\|\lambda\|^2}{2M} \tag{4.30}$$

[4] Note that $\Lambda_k(\lambda) = \frac{1}{k} f_k(k\lambda)$, where f_k is the logarithmic moment generating function of x_k.

[5] The reason for this regularization is to be able to handle the case when $z \in \mathcal{D}_I$ is not an exposed point of I_k, as will be clear from later parts of the proof.

for $x \in \mathbb{R}^d$. Exploiting that, for each t, $\sum_{i=1}^N \alpha_{i,t} = 1$, and applying the lower bound from Proposition 4.5 to $\sum_{i=1}^N \Lambda(\alpha_{i,t}\lambda)$ in (4.27) for each t, yields, for each λ,

$$\Lambda_k(\lambda) \geq N\Lambda(1/N\lambda),$$

and thus

$$I_{k,M}(x) \leq \sup_{\lambda\in\mathbb{R}^d} \lambda^\top x - \Lambda_k(\lambda) \leq \sup_{\lambda\in\mathbb{R}^d} \lambda^\top x - N\Lambda(1/N\lambda) = NI(x). \tag{4.31}$$

Note that the preceding inequality holds for all x, k and $M > 0$. In particular, it follows that $\mathcal{D}_I \subseteq \mathcal{D}_{I_{k,M}}$.

The result of the next lemma makes clear the benefit of the regularization: the quadratic term in (4.30) assures that the infimum in (4.30) has a maximizer $\lambda_k = \lambda_k(x)$ for any given x. This in turn implies that every point $x \in \mathbb{R}^d$ is an exposed point of $I_{k,M}$, which is what we need in our proof[6]. Note that we could not guarantee this property for the conjugate of Λ_k.

Lemma 4.11 For any $x \in \mathbb{R}^d$ and any $k \geq 1$, there exists $\lambda_k = \lambda_k(x) \in \mathbb{R}^d$ such that $I_{k,M}(x) = {\lambda_k}^\top x - \Lambda_{k,M}(\lambda_k)$. Moreover, the pair (x, λ_k) satisfies $x = \nabla\Lambda_{k,M}(\lambda_k)$.

Proof Fix $x \in \mathbb{R}^d$ and define $g_x(\lambda) := \lambda^\top x - \Lambda_{k,M}(\lambda)$, $\lambda \in \mathbb{R}^d$; note that

$$I_{k,M}(x) = \sup_{\lambda\in\mathbb{R}^d} g_x(\lambda). \tag{4.32}$$

If we show that g_x has compact level sets, this would imply that g_x has a global maximizer (see, e.g., Proposition 1.1.3. (Weierstrass theorem) in [84]), which in turn would prove that $\lambda_k^\star(x)$ exists. We now show that g_x has compact level sets.

First, observe that Λ_k is convex as a sum of convex functions $\lambda \mapsto 1/k\Lambda(\alpha_{i,t}\lambda)$, $i = 1, ..., N$, $t = 1, ..., k$. Therefore, Λ_k is minorized by its linear approximation at 0:

$$\Lambda_k(\lambda) \;\geq\; \Lambda_k(0) + \nabla\Lambda_k(0)^\top\lambda, \;\; \forall\lambda \in \mathbb{R}^d. \tag{4.33}$$

Computing the gradient of Λ_k at zero, and using the fact that, for all t, the coefficients $\alpha_{i,t}$ satisfy $\sum_{i=1}^N \alpha_{i,t} = 1$, we obtain

$$\nabla\Lambda_k(0) = \frac{1}{k}\sum_{t=1}^k\sum_{i=1}^N \alpha_{i,t}\nabla\Lambda(\alpha_{i,t}0) = \frac{1}{k}\sum_{t=1}^k\sum_{i=1}^N \alpha_{i,t}\overline{Z} = \overline{Z}.$$

[6]See the paragraph before Lemma (4.14) to appreciate the importance of this result

Further, it is easy to see that $\Lambda_k(0) = 0$. Thus, from (4.33) we get

$$\Lambda_k(\lambda) \geq \overline{Z}^\top \lambda,$$

for every λ, implying

$$g_x(\lambda) = \lambda^\top x - \Lambda_{k,M}(\lambda) \leq \lambda^\top (x - \overline{Z}) - \frac{\|\lambda\|^2}{2M}.$$

Since $\lambda \mapsto \lambda^\top (y - \overline{x}) - \frac{\|\lambda\|^2}{2M}$ which majorizes g_x has compact level sets, it implies that g_x has bounded level sets. Finally, using that g_x is continuous on $\mathbb{R}^d$ (which follows by the continuity of Λ on $\mathbb{R}^d$) we prove that g_x has compact level sets. Thus, we proved the existence of $\lambda_k = \Lambda_k(x)$ corresponding to our chosen x. To show the last part of the claim of Lemma 4.11, we observe that, because λ_k is a minimizer of g_x, there holds that $0 = \nabla g_x(\lambda_k) = x - \nabla \Lambda_{k,M}(\lambda_k)$. Finally, noting that x was arbitrary completes the proof. □

Although we use the result of the following lemma later in the proof, it is convenient to state it now because of its relation with the preceding lemma, Lemma 4.11.

Lemma 4.12 Let x be an arbitrary point $\in \mathbb{R}^d$. Consider the sequence λ_k, $k \geq 1$, where, for each k, λ_k is a maximizer of (4.30). Then, for all k:

$$\|\lambda_k\| \leq M \|x - \overline{z}\|,$$

or, equivalently, $\{\lambda_k : k \geq 1\} \subseteq B_{0, M\|x-\overline{z}\|}$.

Proof Fix $x \in \mathbb{R}^d$. Fix $k \geq 1$ and consider the point λ_k from the setup of the lemma. We prove the claim by showing that λ_k cannot go too far in neither direction.

Suppose that $\frac{1}{\|\lambda_k\|}\lambda_k = v$ for some $v \in \mathbb{R}^d$, $\|v\| = 1$. Denote by ρ_v the norm of λ_k, $\rho_v = \|\lambda_k\|$. We show that ρ_v cannot be greater than $Mv^\top(x - \overline{Z})$. Starting from the identity $x = \nabla \Lambda_{k,M}(\lambda_k)$,

$$x = \frac{1}{k} \sum_{t=1}^{k} \sum_{i=1}^{N} \alpha_{i,t} \nabla \Lambda(\alpha_{i,t} \rho_v v) + 1/M \rho_v v, \tag{4.34}$$

and multiplying both sides with $v^\top$, yields

$$v^\top x = \frac{1}{k} \sum_{t=1}^{k} \sum_{i=1}^{N} \alpha_{i,t} v^\top \nabla \Lambda(\alpha_{i,t} \rho_v v) + 1/M \rho_v. \tag{4.35}$$

where, to get the right-most term, we used that $\|v\| = 1$. Now, since Λ is convex, it has non-decreasing

slopes in any direction, and thus along the line ρv, $\rho > 0$ as well; hence,

$$v^\top \nabla \Lambda \left(\alpha_{i,t} \rho_v v\right) \geq v^\top \nabla \Lambda(0) = v^\top \overline{Z}, \ \ \forall i, t.$$

(Note that we use here that $\alpha_{i,t}$ is nonnegative). Combining the preceding inequality with (4.35), we get

$$v^\top x \geq v^\top \overline{Z} + 1/M \rho_v, \tag{4.36}$$

and we conclude that $\rho_v \leq M v^\top (x - \overline{Z})$. Repeating this argument for all directions v, and computing the supremum $\sup_{v \in \mathbb{R}^d, \|v=1\|} M v^\top (x - \overline{Z}) = M \|x - \overline{Z}\|$, proves the claim. □

Now, fix $z \in \mathcal{D}_I$ and fix $\delta > 0$. Then, by Lemma 4.11, there exists a point η_k such that $I_{k,M}(z) = \eta_k^\top z - \Lambda_{k,M}(\eta_k)$; note the subscript k which indicates the dependence of this point on k. For any k we use the η_k to change the measure on $\mathbb{R}^d$ from μ_k to $\widetilde{\mu}_k$ through:

$$\frac{d\widetilde{\mu}_k}{d\mu_k}(x) = e^{\,k\,\eta_k^\top x \,-\, k\,\Lambda_{k,M}(\eta_k)}, \ \ x \in \mathbb{R}^d. \tag{4.37}$$

$x \in \mathbb{R}^d$. Note that, in contrast with the standard method of Gärtner-Ellis Theorem where the change of measure is fixed (once z is given), here we have a different change of measure for each k[7] [8] Using the transformation of measure above, it is easy to show that the logarithmic moment generating function associated with $\widetilde{\mu}_k$ is given by $\widetilde{\Lambda}_{k,M} := \Lambda_{k,M}(\lambda + \eta_k) - \Lambda_{k,M}(\eta_k)$. Fix $\delta > 0$ and consider $\mu_k\left(B_{z,\delta}\right)$ (for the fixed z and δ). Expressing this probability through $\widetilde{\mu}_k$, for each k, we have:

$$\frac{1}{k} \log \mu_k \left(B_{z,\delta}\right) = \Lambda_{k,M}(\eta_k) - \eta_k^\top z + \frac{1}{k} \log \int_{x \in B_{z,\delta}} e^{k \eta_k^\top (z-x)} d\widetilde{\mu}_k(x) \tag{4.38}$$

$$\geq \Lambda_{k,M}(\eta_k) - \eta_k^\top z - \delta \left\|\eta_k\right\| + \frac{1}{k} \log \widetilde{\mu}_k \left(B_{z,\delta}\right). \tag{4.39}$$

The first term in the equation above equals $\Lambda_{k,M}(\eta_k) - \eta_k^\top z = -I_{k,M}(z)$, and since $I_{k,M}(\cdot) \leq N I(\cdot)$ (see eq. (4.31)) it follows

$$\frac{1}{k} \log \mu_k \left(B_{z,\delta}\right) \geq -N I(z) - \delta \left\|\eta_k\right\| + \frac{1}{k} \log \widetilde{\mu}_k \left(B_{z,\delta}\right). \tag{4.40}$$

[7] The obvious reason for this alteration of the standard method is the fact that our sequence of functions $\Lambda_{k,M}$ does not have a limit.

[8] It can be shown that all distributions $\widetilde{\mu}_k$, $k \geq 1$, have the same expected value z; we do not pursue this result here, as it is not crucial for our goals.

Letting now $k \to +\infty$ and then $\delta \to 0$

$$\lim_{\delta \to 0} \liminf_{k \to +\infty} \frac{1}{k} \log \mu_k \left(B_{z,\delta} \right) \geq -NI(z) - \lim_{\delta \to 0} \delta \limsup_{k \to +\infty} \|\eta_k\| + \lim_{\delta \to 0} \liminf_{k \to +\infty} \frac{1}{k} \log \widetilde{\mu}_k \left(B_{z,\delta} \right). \tag{4.41}$$

Applying now Lemma 4.12 to the sequence η_k, we have that $\limsup_{k \to +\infty} \|\eta_k\| \leq M \left\| z - \overline{Z} \right\|$, and, thus, the second term in (4.41) vanishes:

$$\lim_{\delta \to 0} \delta \limsup_{k \to +\infty} \|\eta_k\| = 0.$$

To prove the claim of Lemma 4.10 it remains to show that the third term in (4.41) vanishes as well.

Recall that the logarithmic moment generating function associated with $\widetilde{\mu}_k$ is $\widetilde{\Lambda}_{k,M} := \Lambda_{k,M}(\lambda + \eta_k) - \Lambda_{k,M}(\eta_k)$, and let $\widetilde{I}_{k,M}$ denote the conjugate of $\widetilde{\Lambda}_{k,M}$. It can be shown that

$$\limsup_{k \to +\infty} \widetilde{\mu}_k \left(B_{z,\delta}^c \right) \leq - \liminf_{k \to +infty} \inf_{w \in B_{z,\delta}^c} \widetilde{I}_{k,M}(w). \tag{4.42}$$

If we show that the right hand side in (4.42) is strictly negative, that would imply $\widetilde{\mu}_k \left(B_{z,\delta} \right) \to 1$, as $k \to +\infty$, which in turn yields that the third term in (4.41) vanishes. Thus, from now on we focus on proving that the the right hand side in (4.42) is strictly negative.

Consider

$$\inf_{w \in B_{z,\delta}^c} \widetilde{I}_{k,M}(w). \tag{4.43}$$

Now, from the fact that $\mathcal{D}_{\widetilde{\Lambda}_{k,M}} = \mathbb{R}^d$, we can show that $\widetilde{I}_{k,M}$ has compact level sets (note that $\widetilde{I}_{k,M}$ is lower semicontinuous). Thus, the infimum problem in (4.43) has a solution. Denote a solution by w_k and let ζ_k be a point for which $w_k = \nabla \widetilde{\Lambda}_{k,M} \left(\zeta_k \right) \left(= \nabla \Lambda_{k,M} \left(\zeta_k + \eta_k \right) \right)$ (such a point exists by Lemma 4.11).

We now show that $\|w_k\|$ is uniformly bounded for all k, which, combined with Lemma 4.12, in turn implies that $\eta_k + \zeta_k$ is uniformly bounded.

Lemma 4.13 For any fixed $\delta > 0$ and $M > 0$, there exists $R = R(z, \delta, M) < +\infty$ such that, for all k,

1. $\|w_k\| \leq R$, implying $\{w_k : k \geq 1\} \subseteq B_{0,R}$;

2. $\|\zeta_k + \eta_k\| \leq M \sup_{w \in B_{0,R}} \left\| w - \overline{Z} \right\|$.

Proof Fix $M > 0, \delta > 0$. Define $\overline{f}_M$ and $\underline{f}_M$ as

$$\overline{f}_M(x) = \sup_{\lambda \in \mathbb{R}^d} \lambda^\top x - N\Lambda\left(1/N\lambda\right) - \frac{\|\lambda\|^2}{2M}, \quad \underline{f}_M(x) = \sup_{\lambda \in \mathbb{R}^d} \lambda^\top x - \Lambda(\lambda) - \frac{\|\lambda\|^2}{2M}.$$

$x \in \mathbb{R}^d$. Note that both $\overline{f}_M, \underline{f}_M$ are lower semicontinuous, finite for every x and have compact level sets. Let $c = \inf_{x \in B^c_{z,\delta}} \overline{f}_M \in \mathbb{R}$, and define $S_c = \left\{x \in \mathbb{R}^d : \underline{f}_M(x) \leq c\right\}$.

Fix arbitrary $k \geq 1$. It can be shown with the help of Proposition (4.5) that for any $x \in \mathbb{R}^d$,

$$\underline{f}_M(x) \leq I_{k,M}(x) \leq \overline{f}_M(x). \tag{4.44}$$

Observe now that $I_{k,M}(w_k) = \inf_{x \in B^c_{z,\delta}} I_{k,M}(x) \leq \inf_{x \in B^c_{z,\delta}} f_M(x) \leq c$. On the other hand, taking in (4.44) $x = w_k$, yields $\underline{f}_M(w_k) \leq I_{k,M}(w_k)$, and it thus follows that w_k belongs to S_c.

Finally, as S_c is compact, we can find a ball of some radius $R = R(z, M, \delta) > 0$ that covers S_c, implying $w_k \in B_{0,R}$. Since k was arbitrary, the claim in part 1 follows.

We now prove part 2. Fix $k = k_0 \geq 1$ and consider the pair $w_{k_0}, \zeta_{k_0} + \eta_{k_0}$ for which $w_{k_0} = \nabla\Lambda_{k_0,M}\left(\zeta_{k_0} + \eta_{k_0}\right)$. Applying Lemma 4.12 for for $z = w_{k_0}$ we have that $\|\zeta_{k_0} + \eta_{k_0}\| \leq M\left\|w_{k_0} - \overline{Z}\right\|$. Using now part 1 of this lemma (which asserts that for any k $w_k \in B_{0,R}$, and thus for $k = k_0$) we have $w_{k_0} \in B_{0,R}$. Combining the two preceding conclusions,

$$\|\zeta_{k_0} + \eta_{k_0}\| \leq M\left\|w_{k_0} - \overline{Z}\right\| \leq M \sup_{w \in B_{0,R}} \left\|w - \overline{Z}\right\|. \tag{4.45}$$

Since k_0 was arbitrary, the proof of Lemma 4.13 is complete. □

Observe now that

$$\widetilde{I}_{k,M}\left(w_k\right) = \zeta_k^\top w_k - \Lambda_{k,M}\left(\zeta_k + \eta_k\right) + \Lambda_{k,M}\left(\eta_k\right)$$

By strict convexity of $\Lambda_{k,M}$ (recall the quadratic term $1/2M\|\lambda\|$ in $\Lambda_{k,M}$), we have

$$\Lambda_{k,M}\left(\eta_k\right) > \Lambda_{k,M}\left(\zeta_k + \eta_k\right) + \nabla\Lambda_{k,M}\left(\zeta_k + \eta_k\right)^\top\left(\eta_k - \left(\zeta_k + \eta_k\right)\right)$$

and using that $\nabla\Lambda_{k,M}\left(\zeta_k + \eta_k\right) = w_k$, finally yields

$$\Lambda_{k,M}\left(\eta_k\right) > \Lambda_{k,M}\left(\zeta_k + \eta_k\right) - w_k^\top \zeta_k.$$

Therefore, we obtain that $\widetilde{I}_{k,M}\left(w_k\right)$ is strictly positive. Note also that, since k was arbitrary, the same

conclusion holds for all k. However, we need a stronger claim for our purpose: each member of the sequence of values $\widetilde{I}_{k,M}(w_k)$ be bounded away from zero. The next lemma asserts that this is indeed the case.

Lemma 4.14 For any fixed $z \in \mathbb{R}^d$, $\delta > 0$, and $M > 0$, there exists $\xi = \xi(z, \delta, M) > 0$ such that, for all k,

$$\widetilde{I}_{k,M}(w_k) \geq \xi.$$

Proof Fix z, δ and M and define $r_1 = M\left\|z - \overline{Z}\right\|$, $r_2 = \sup_{w \in B_{0,R}} M\left\|w - \overline{Z}\right\|$, where R is the constant that verifies Lemma 4.13. Fix now $k \geq 1$ and recall that η_k, ζ_k and w_k are chosen such that $z = \nabla\Lambda_{k,M}(\eta_k)$, $\widetilde{I}_{k,M}(w_k) = \inf_{x \in B^c_{z,\delta}}$, and $w_k = \nabla\Lambda_{k,M}(\eta_k + \zeta_k)$. Now, by Lemma 4.12 and 4.13 we have for η_k and ζ_k, $\|\eta_k\| \leq r_1$, $\|\eta_k + \zeta_k\| \leq r_2$. We will now show that there exists some positive constant $r_3 (> 0)$ independent of k such that $\|\eta_k\| \geq r_3$. Consider the gradient map $\lambda \mapsto \nabla\Lambda_{k,M}(\lambda)$, and note that it is continuous, and hence uniformly continuous on every compact set. Note that $\|\eta_k\|, \|\eta_k + \zeta_k\| \leq \max\{r_1, r_2\}$. Suppose now for the sake of contradiction that $\|\eta_k\| \to 0$, as $k \to +\infty$. Then, $\|(\eta + \zeta_k) - \eta_k\| \to 0$, and thus, by the uniform continuity of $\nabla\Lambda_{k,M}(\cdot)$ on $B_{0,\max\{r_1,r_2\}}$ we have

$$\|\nabla\Lambda_{k,M}(\eta_k) - \nabla\Lambda_{k,M}(\eta_k + \zeta_k)\| \to 0, \text{ as } k \to \infty.$$

Recalling that $z = \nabla\Lambda_{k,M}(\eta_k)$, $w_k = \nabla\Lambda_{k,M}(\eta_k)$, yields

$$\|w_k - z\| \to 0.$$

This contradicts with $w_k \in B^c_{z,\delta}$. Thus, we proved the existence of r_3 independent of k such that $\|\eta_k\| \geq r_3$, for all k.

Now, let

$$\Upsilon = \left\{(\eta, \zeta) \in \mathbb{R}^d \times \mathbb{R}^d : \|\eta\| \leq r_1, \|\eta + \zeta\| \leq r_2, \|\zeta\| \geq r_3\right\},$$

and introduce $f : \mathbb{R}^d \times \mathbb{R}^d \mapsto \mathbb{R}$,

$$f(\zeta, \eta) = \Lambda_{k,M}(\eta) - \Lambda_{k,M}(\zeta + \eta) + \nabla\Lambda_{k,M}(\zeta + \eta)^\top \zeta. \tag{4.46}$$

By strict convexity of $\Lambda_{k,M}(\cdot)$, we see that for any η and $\zeta \neq 0$, the value $f(\eta, \zeta)$ is strictly positive. Further,

note that since $\Lambda_{k,M}$ and $\nabla\Lambda_{k,M}$ are continuous, the function f is also continuous. Consider now

$$\xi := \inf_{\Upsilon} f(\eta, \zeta). \tag{4.47}$$

Because Υ is compact, by the Weierstrass theorem, the problem in (4.47) has a solution, that is, there exists $(\eta_0, \zeta_0) \in S$, such that $f(\eta_0, \zeta_0) = \xi$. Finally, because f is strictly positive at each point in Υ (note that $\zeta \neq 0$ in Υ), we conclude that $\xi = f(\eta_0, \zeta_0) > 0$.

Returning to the claim of Lemma 4.14, for every fixed k, $(\eta_k, \eta_k + \zeta_k)$ belongs to Υ, and, thus,

$$\widetilde{I}_{k,M}(w_k) = \Lambda_{k,M}(\eta_k) - \Lambda_{k,M}(\zeta_k + \eta_k) + \nabla\Lambda_{k,M}(\zeta_k + \eta_k)^\top \zeta_k = f(\eta_k, \zeta_k) \geq \xi.$$

This completes the proof of Lemma 4.14. □

We now establish that the third term in (4.41) vanishes. From (4.42), and the existence of $\xi > 0$ from Lemma 4.14, we have

$$\limsup_{k\to+\infty} \frac{1}{k} \log \widetilde{\mu}_k \left(B^{\mathrm{c}}_{z,\delta}\right) \leq -\liminf_{k\to+\infty} \widetilde{I}_{k,M}(w_k) \leq -\xi < 0.$$

This implies that for any fixed $\delta > 0$

$$\mu_k \left(B_{z,\delta}\right) \to 1, \text{ as } k \to +\infty,$$

which finally yields:

$$\lim_{\delta\to 0} \liminf_{k\to+\infty} \log \widetilde{\mu}_k \left(B_{z,\delta}\right) = 0.$$

Thus, the third term in (4.41) vanishes, establishing (4.29). □

Having (4.29), it is easy to establish (4.25). Recall that $x_k = y_k - V/k\sqrt{M}$. Then,

$$\mathbb{P}\left(x_k \in B_{z,2\delta}\right) \geq \mathbb{P}\left(y_k \in B_{z,\delta}, V/k\sqrt{M} \in B_{z,\delta}\right) \geq \mathbb{P}\left(y_k \in B_{z,\delta}\right) - \mathbb{P}\left(V/\sqrt{kM} \notin B_{z,\delta}\right). \tag{4.48}$$

From (4.29), the rate for the probability of the first term above is at most $NI(z)$. On the other hand, the probability that the norm of V is greater than $\sqrt{kM}\delta$ decays exponentially with k with the rate $M\delta^2/2$,

$$\lim_{k\to+\infty} \frac{1}{k} \log \mathbb{P}\left(V/\sqrt{kM} \in B_{z,\delta}\right) = -\frac{M\delta^2}{2}. \tag{4.49}$$

Observing that, for any fixed δ, for all M large enough such that $NI(z) < \frac{M\delta^2}{2}$, the rate of the difference

in (4.48) is dominated by the rate of the first term, $NI(z)$. This finally establishes (4.25), and proves the lower bound 4.19.

4.4.3 Large deviations for distributed inference: Generic matrices W_t

We are now ready to state and prove Theorem 4.15.

Theorem 4.15 Consider distributed inference algorithm (4.3)–(4.4) under Assumptions 4.1 and 4.3. Then, for each node i:

1. **(No worse than isolation)** For any closed set E:

$$\limsup_{k\to+\infty} \frac{1}{k} \log \mathbb{P}\left(x_{i,k} \in E\right) \leq - \inf_{x\in E} I(x) \tag{4.50}$$

2. **(No better than centralization)** For any open set F:

$$\liminf_{k\to+\infty} \frac{1}{k} \log \mathbb{P}\left(x_{i,k} \in F\right) \geq -N \inf_{x\in F} I(x). \tag{4.51}$$

Proof of the "No worse than isolation" bound. Fix $W_1, \ldots, W_k$ and consider the conditional probability $\mathbb{P}\left(x_{i,k} \in E | W_1, \cdots, W_k\right)$. Let G be an arbitrary measurable set in $\mathbb{R}^d$. Then, similarly as in Subsection 4.4.1 we obtain,

$$\mathbb{P}\left(x_{i,k} \in G | W_1, \ldots, W_k\right) \leq e^{-k \inf_{x\in E}\left(\lambda^\top x - \Lambda(\lambda)\right)},$$

which further, by monotonicity of the expectation, yields:

$$\mathbb{P}\left(x_{i,k} \in G\right) = \mathbb{E}\left[\mathbb{P}\left(x_{i,k} \in E | W_1, \ldots, W_k\right)\right] \leq e^{-k\left(\lambda^\top x - \Lambda(\lambda)\right)}.$$

From here, the proof proceeds analogously to the proof of the upper bound (4.18).

Proof of the "No better than centralization" bound. We now prove the lower bound (4.51). Consider again a fixed realization of W_1,...,W_k.

Similarly as in the preceding proof, applying Lemma 4.9, now the lower bound (4.19), to the sequence $x_{i,k}$ (given W_1,..., W_k), yields

$$\liminf_{k\to+\infty} \frac{1}{k} \log \mathbb{P}\left(x_{i,k} \in F | W_1, \ldots, W_k\right) \geq -N \inf_{x\in F} I(x). \tag{4.52}$$

Applying Fatou's lemma [69] to the sequence of random variables $R_k = \frac{1}{k}\log\mathbb{P}\left(x_{i,k} \in F | W_1, \ldots, W_k\right)$, $k = 1, 2, \ldots$,

$$\liminf_{k\to+\infty} \mathbb{E}\left[\frac{1}{k}\log\mathbb{P}\left(x_{i,k} \in F | W_1, \ldots, W_k\right)\right] \geq \mathbb{E}\left[\liminf_{k\to+\infty} \frac{1}{k}\log\mathbb{P}\left(x_{i,k} \in F | W_1, \ldots, W_k\right)\right],$$

and then using the monotonicity of the expectation in (4.52), yields

$$\liminf_{k\to+\infty} \mathbb{E}\left[\frac{1}{k}\log\mathbb{P}\left(x_{i,k} \in F | W_1, \ldots, W_k\right)\right] \geq -N \inf_{x\in F} I(x). \tag{4.53}$$

Now, by concavity of the logarithm:

$$\frac{1}{k}\log\mathbb{P}\left(x_{i,k} \in F\right) = \frac{1}{k}\log\mathbb{E}\left[\mathbb{P}\left(x_{i,k} \in F | W_1, \ldots, W_k\right)\right] \geq \mathbb{E}\left[\frac{1}{k}\log\mathbb{P}\left(x_{i,k} \in F | W_1, \ldots, W_k\right)\right].$$

Finally, passing to $\liminf_{k\to+\infty}$, and using (4.53), completes the proof. □

4.5 Large deviations for distributed inference: Doubly stochastic matrices W_t

This section shows that if an additional structure is assumed in cooperation, namely double-stochasticity of the W_t's, the performance guarantees under cooperation significantly improve.

4.5.1 Model, assumptions, and graph-related objects

To state the corresponding result, we first need to introduce certain concepts.

Definition 4.16 (Convex hull of a function, [72]) Let $f : \mathbb{R}^d \mapsto \mathbb{R} \cup \{+\infty\}$ be minorized by an affine function. Consider the closed convex hull $\overline{\text{co}}\,\text{epi}\, f$ of the epigraph of f, $\text{epi} f = \left\{(x, r) : r \geq f(x), x \in \mathbb{R}^d\right\}$. The closed convex hull of f, denoted by $\overline{\text{co}} f(\cdot)$, is defined by:

$$\overline{\text{co}} f(x) := \inf\{r : (x, r) \in \overline{\text{co}}\,\text{epi}\, f\}. \tag{4.54}$$

The convex hull $\overline{\text{co}} f(x)$ of the function f is constructed by first finding the convex hull of the epigraph of f, and then defining $\overline{\text{co}} f(x)$ to be the function whose epigraph is $\overline{\text{co}}\,\text{epi}\, f$. Intuitively, $\overline{\text{co}} f(x)$ is the best convex approximation of f.

We recall from Chapter 2 some objects related with random matrices W_t. The induced graph of W_t, $G_t = G(W_t)$, is the graph that collects all the edges corresponding to positive entries of W_t. Set $\mathcal{G}$ is then the set of all possible realizations of G_t, i.e., $\mathcal{G}$ is the minimal set for which $\mathbb{P}(G_t \in \mathcal{G}) = 1$. For a collection of graphs $\mathcal{H} \subseteq \mathcal{G}$ on the same set of vertices V, $\Gamma(\mathcal{H})$ denotes the graph that collects all the edges of all the graphs in $\mathcal{H}$, i.e., $\Gamma(\mathcal{H})$ is the union graph of graphs in $\mathcal{H}$. For a given collection of graphs $\mathcal{H}$, we introduce – what we call – the clique of a node i in $\mathcal{H}$.

Definition 4.17 (The clique of node i in the set of graphs $\mathcal{H}$) Let $\mathcal{H}$ be a collection of graphs, and let $C_1, ..., C_M$ be the components of the union graph $\Gamma(\mathcal{H})$. Then, for any node i, $C_{i,\mathcal{H}}$ is the component of $\Gamma(\mathcal{H})$ that contains i, i.e., if $i \in C_m$, then $C_{i,\mathcal{H}} = C_m$, and call it the clique of node i in $\mathcal{H}$; thus,

$$C_{i,\mathcal{H}} = \{j \in V : \text{ there exists a path in } \Gamma(\mathcal{H}) \text{ from } i \text{ to } j\}. \tag{4.55}$$

We explain the intuition behind this definition. Suppose that the sequence of induced graphs G_t takes realizations in a fixed collection of graphs $\mathcal{H} = \{H_1, ..., H_L\} \subset \mathcal{G}$, $H_l \in \mathbb{G}^N$. Consider the union graph $\Gamma(\mathcal{H})$; note that, because $C_{i,\mathcal{H}}$ is by construction a component of $\Gamma(\mathcal{H})$, there are no edges in $\Gamma(\mathcal{H})$ between the nodes in $C_{i,\mathcal{H}}$ and the rest of the network $V \setminus C_{i,\mathcal{H}}$). Now, because each H_l is a subgraph of $\Gamma(\mathcal{H})$ ($\Gamma(\mathcal{H})$ being the union graph of the H_l's), there cannot be any edges between $C_{i,\mathcal{H}}$ and $V \setminus C_{i,\mathcal{H}}$ in the H_l's as well. Thus, none of the realizations along the sequence of graphs G_t can connect $C_{i,\mathcal{H}}$ with the remaining nodes, implying that $C_{i,\mathcal{H}}$ is the only part of the network that node i can "see", i.e., communicate with over time. To explain why we call $C_{i,\mathcal{H}}$ a clique, imagine that, along the sequence G_t, each realization H_l, $l = 1, ..., L$ occurs infinitely often (note that this happens with probability 1). Then, after a finite time, loosely, "the information from every node in $C_{i,\mathcal{H}}$ has reached every other node in $C_{i,\mathcal{H}}$"[9] We end this Subsection by stating an additional Assumption on the matrices W_t.

Assumption 4.18 The random matrices W_t are symmetric and have positive diagonals, with probability one.

4.5.2 Large deviations rate, corollaries, and interpretations

We are now ready to state and prove our main result of this chapter, Theorem 4.19.

Theorem 4.19 Consider distributed inference algorithm (4.3)–(4.4) under Assumptions 4.1, 4.3, and 4.18. Then, for each node i:

[9]To make this statement more precise, note that if each H_l occurs infinitely often, then, if the W_t's have positive diagonal entries, the product matrix $W_k \cdots W_1$ has all the entries strictly positive after a finite time.

1. For any closed set E:

$$\limsup_{k\to+\infty} \frac{1}{k} \log \mathbb{P}\left(x_{i,k} \in E\right) \leq - \inf_{x\in E} I_{\mathcal{J},N}(x), \tag{4.56}$$

$$I_{\mathcal{J},N}(x) = \overline{\text{co}} \inf \left\{I(x) + \mathcal{J}, NI(x)\right\}; \tag{4.57}$$

2. For any open set F:

$$\liminf_{k\to+\infty} \frac{1}{k} \log \mathbb{P}\left(x_{i,k} \in F\right) \geq - \inf_{x\in F} \overline{I}_i(x), \tag{4.58}$$

$$\overline{I}_i(x) = \inf_{\mathcal{H}\subseteq\mathcal{G}} \overline{\text{co}} \left\{|C_{i,\mathcal{H}}|I(x) + |\log p_{\mathcal{H}}|, NI(x)\right\}. \tag{4.59}$$

Before proving Theorem 4.5, we first give the interpretation of the upper bound function $I_{\mathcal{J},N}$, and then we state some important corollaries of Theorem 4.5.

As we can see from part 1 Theorem 4.5, function $I_{\mathcal{J},N}$ gives an upper bound on the exponential decay rate for the large deviation probabilities of each node i in the network. We provide an illustration for $I_{\mathcal{J},N}$ in Figure 4.1. We consider an instance of the inference algorithm (4.3)-(4.4) running on a $N = 3$-node network, with the rate of the matrix products $\mathcal{J} = 5$, and where the observations $Z_{i,t}$ are standard Gaussian (zero mean and variance equal to one); it can be shown that, for this case, $I(x) = \frac{1}{2}x^2$. The more curved blue dotted line plots the function $NI(x) = \frac{1}{2}Nx^2$, the less curved blue dotted line plots the function $I(x) + \mathcal{J} = \frac{1}{2}x^2 + \mathcal{J}$, and the solid red line plots $I_{\mathcal{J},N}$. Observing the figure, we can identify three regions that define $I_{\mathcal{J},N}$. First, there exists a region around the mean value (equal to zero) in which $I_{\mathcal{J},N}$ matches the optimal rate NI. On the other hand, for all values of x that are sufficiently large, $I_{\mathcal{J},N}$ follows the slower rate, $I(\cdot) + \mathcal{J}$. Finally, for the range of values of x that are in between the preceding two regions, function $I_{\mathcal{J},N}$ is linear; this linear part is the tangent line that touches both the epigraph of NI and the epigraph of $I + \mathcal{J}$, and Intuitively, the linear part is responsible for the "convexification" (recall that $\overline{\text{co}}$ of two functions is defined through the convex hull of their corresponding epigraphs) of the point-wise maximum $\max\left\{I(\cdot) + \mathcal{J}, NI(\cdot)\right\}$.

The first corollary of Theorem 4.19 asserts that if every realization of the network topology is connected, then the sequence $x_{i,k}$ satisfies the large deviation principle at each node i in the network and with the best possible rate function, $NI(\cdot)$.

Corollary 4.20 Assume that the induced graph G_t is connected with probability 1. Then, for each $i = 1, \ldots, N$ the sequence of states of node i, $x_{i,k}$, satisfies the large deviation principle with rate function NI.

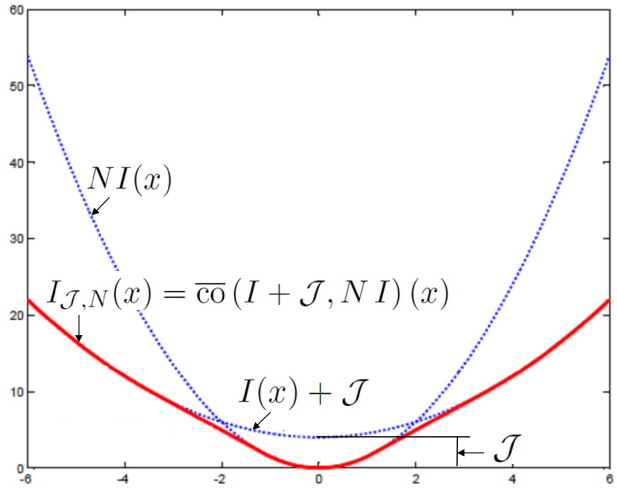

Figure 4.1: Illustration of $I_{\mathcal{J},N}$ for a network of size $N = 3$, with $\mathcal{J} = 5$, and $Z_{i,t} \sim \mathcal{N}(0,1)$. The more curved blue dotted line plots $NI(x) = \frac{1}{2}Nx^2$, the less curved blue dotted line plots $I(x) + \mathcal{J} = \frac{1}{2}x^2 + \mathcal{J}$. The solid red line plots $I_{\mathcal{J},N} = \overline{\text{co}}\,(NI(\cdot), I(\cdot) + \mathcal{J})$.

Proof If the induced graph G_t is connected with probability 1, then by Theorem 2.7 from Chapter 2 we have $\mathcal{J} = +\infty$, and so for any x, $I_{\mathcal{J},N}(x) = NI(x)$. Combining this with the lower bound Theorem 4.15 which asserts that, for any open set F, the exponential rate of the probabilities $\mathbb{P}\,(x_{i,k} \in F)$ is at most $\inf_{x\in F} NI(E)$, the claim follows. □

In particular, it follows that any deterministic cooperation algorithm such that, with probability 1, $W_t \equiv A$, and $|\lambda_2(A)| < 1$ satisfies the large deviation principle with the optimal rate function $NI(\cdot)$.

Consider now a situation when there exists a node i such that $\mathcal{J} = |\log p_{i,\text{isol}}|$, where $p_{i,\text{isol}}$ is the probability that i has no neighbors at some time t. This means that the most likely disconnected collection (recall the definition from Chapter 2) is the one which consists of all the topology realizations that isolate i: $\mathcal{H}^\star = \{H \in \mathcal{G} : C_{i,H} = i\}$, hence,

$$p_{\max} = p_{\mathcal{H}^\star} = \sum_{H\in\mathcal{G}:C_{i,H}=i} p_H. \tag{4.60}$$

Thus, for $\mathcal{H}^\star$, $C_{i,\mathcal{H}^\star} = i$ and so $|C_{i,\mathcal{H}^\star}| = 1$. Consider now the lower bound in (4.58) where, instead of the (tighter) infimum bound we use the looser bound $\overline{\text{co}}\,\{NI(x), I(x) + |\log p_{\mathcal{H}^\star}|\}$. Now, simply noting that $|\log p_{\mathcal{H}^\star}| = \mathcal{J}$, we see that the two functions in the upper and the lower bound coincide, thus implying the large deviation principle for node i. This is formally stated in the next corollary.

Corollary 4.21 (LDP for a critical node) Suppose that for some i, $\mathcal{J} = |\log p_{i,\text{isol}}|$. Then, the sequence of

measures $E \mapsto \mathbb{P}(x_{i,k} \in E)$, $E \in \mathcal{B}\left(\mathbb{R}^d\right)$, satisfies the large deviation principle with the rate function $\overline{\text{co}}\left\{NI(x), I(x) + |\log p_{i,\text{isol}}|\right\}$.

Further important corollaries along the same line of thought are the two corollaries below: LDP for regular networks and the LDP for leaf nodes in a graph.

Corollary 4.22 (LDP for leaf nodes) Suppose that algorithm (4.3)–(4.4) runs on a network $\left(V, \widehat{E}\right)$, where $\widehat{E}$ collects all the links that have positive probability of occurrence. Further, assume that all links in $\widehat{E}$ have the same probability p of being online. Then, for every leaf node i, the corresponding sequence of measures satisfies the LDP with the rate function $\overline{\text{co}}\left\{NI(x), I(x) + |\log(1-p)|\right\}$.

Corollary 4.23 (LDP for regular networks) Suppose that algorithm (4.3)–(4.4) runs on a regular network of degree d in which each link has the same probability p of being online. Then, for every node i, the corresponding sequence of measures satisfies the LDP with the rate function $\overline{\text{co}}\left\{NI(x), I(x) + d\,|\log(1-p)|\right\}$.

The next two subsections prove Theorem 4.19; Subsection 4.5.3 proves the upper bound (4.56) and Subsection 4.5.4 proves the lower bound (4.58).

4.5.3 Proof of the upper bound in (4.56)

For $k \geq 1$, let S_k be defined by

$$S_k = \begin{cases} \max 1 \leq s \leq k : \|\Phi(k,s) - J\| < \frac{1}{k}, & \text{if } \Phi(k,0) < \frac{1}{k} \\ 0, & \text{otherwise} \end{cases}. \tag{4.61}$$

To prove (4.56), it suffices to show that for any open ball G and any $\epsilon > 0$

$$\mathbb{P}\left(x_{i,k} \in G\right) \leq C_\epsilon e^{-k\left(\inf_{x \in G} \lambda^\top x - \max\{\Lambda(\lambda) - (\mathcal{J} - \epsilon), N\Lambda(\lambda/N)\}\right)}, \tag{4.62}$$

where the constant C_ϵ is the constant that verifies

$$\mathbb{P}\left(\|W_k \cdots W_1 - J_N\| \geq \frac{1}{k}\right) \leq C_\epsilon e^{k(\mathcal{J}-\epsilon)}.$$

The claim in (4.56) can then be established by the standard "finite cover" argument, combined with the exponential tightness of $G \mapsto \mathbb{P}(x_{i,k \in G})$, and the fact that the conjugate of $\max\left\{\Lambda(\lambda) - \mathcal{J}, N\Lambda\left(\lambda/N\right)\right\}$ is the convex hull of the conjugates of $\Lambda(\lambda) - \mathcal{J}$ and $N\Lambda(\lambda/N)$, respectively, given by $I(x) + \mathcal{J}$ and $NI(x)$.

By the law of total probability:

$$\mathbb{P}\left(x_{i,k} \in G\right) = \sum_{s=0}^{k} \mathbb{P}\left(x_{i,k} \in G | S_k = s\right) \mathbb{P}\left(S_k = s\right). \tag{4.63}$$

Consider a fixed realization of $W_1, \ldots, W_k$. Similarly as in the proof of Lemma 4.9, see eq. (4.23), by the independence of the terms $[\Phi(k,t)]_{ij} Z_{j,t}$ (given $W_1, ..., W_k$), $j = 1, ..., N$, $t = 1, ..., k$, we have

$$\mathbb{P}\left(x_{i,k} \in G | W_1, \ldots, W_k\right) \leq e^{-k \inf_{x \in G} \lambda^\top x + \sum_{t=1}^{k} \sum_{j=1}^{N} \Lambda([\Phi(k,t)_{ij} \lambda])}. \tag{4.64}$$

Suppose now that, for this fixed realization of the random matrices, $S_k = s$, i.e., $\|\Phi(k,s) - J\| < \frac{1}{k}$ and $\|\Phi(k,s+1) - J\| \geq \frac{1}{k}$. Then, $[\Phi(k,t)]_{ij} \in [\frac{1}{N} - \frac{1}{k}, \frac{1}{N} + \frac{1}{k}]$, for all $1 \leq t \leq s$. By the convexity of $\Lambda(\cdot)$, and the fact that $\Lambda(0) = 0$, this implies

$$\Lambda\left([\Phi(k,t)]_{ij} \lambda\right) \leq \Lambda\left(\left(\frac{1}{N} + \frac{1}{k}\right)\lambda\right) \tag{4.65}$$

, for each t between 1 and s, and for each j. Therefore, summing out over all j's for a fixed t, $1 \leq t \leq s$, yields

$$\sum_{j=1}^{N} \Lambda\left([\Phi(k,t)]_{ij} \lambda\right) \leq N \Lambda\left(\left(\frac{1}{N} + \frac{1}{k}\right)\lambda\right). \tag{4.66}$$

On the other hand, t greater than s, we cannot say much about the entries of $\Phi(k,t)$, so we use the "worst case" bound from Proposition 4.5,

$$\sum_{j=1}^{N} \Lambda\left([\Phi(k,t)]_{ij} \lambda\right) \leq \Lambda\left(\lambda\right). \tag{4.67}$$

Consider now the sum in the exponent in (4.64). Applying (4.66) to the first s terms (i.e., the terms for which $1 \leq t \leq s$) and (4.67) to the remaining $k - s$ terms,

$$\sum_{t=1}^{k} \sum_{j=1}^{N} \Lambda\left([\Phi(k,t)_{ij} \lambda]\right) \leq sN\Lambda\left(\left(\frac{1}{N} + \frac{1}{k}\right)\lambda\right) + (k-s)\Lambda(\lambda). \tag{4.68}$$

Since the preceding bound holds for any realization of W_1,...,W_k for which $S_k = s$, computing the conditional expectation in (4.64) over the event $\{S_k = s\}$ yields

$$\mathbb{P}\left(x_{i,k} \in G | S_k = s\right) \leq e^{-k \inf_{x \in G} \lambda^\top x + sN\Lambda\left(\left(\frac{1}{N} + \frac{1}{k}\right)\lambda\right) + (k-s)\Lambda(\lambda)}. \tag{4.69}$$

We next consider probability of the event $\{S_k = s\}$:

$$\mathbb{P}\left(S_k = s\right) = \mathbb{P}\left(\|\Phi(k,s) - J\| < \frac{1}{k} \text{ and } \|\Phi(k,s+1) - J\| \geq \frac{1}{k}\right) \tag{4.70}$$

$$\leq \mathbb{P}\left(\|\Phi(k,s+1) - J\| \geq \frac{1}{k}\right). \tag{4.71}$$

By Theorem 2.7 from Chapter 2, the probability in the right hand side of (4.70) decays exponentially with $k - s$ and with the rate equal to $\mathcal{J}$ defined in (2.2). In particular, for any $\epsilon > 0$, there exists a constant C_ϵ such that for each k

$$\mathbb{P}\left(\|\Phi(k,s+1) - J\| \geq \frac{1}{k}\right) \leq C_\epsilon e^{-(k-s)\mathcal{J}},$$

implying

$$\mathbb{P}\left(S_k = s\right) \leq C_\epsilon e^{-(k-s)(\mathcal{J}-\epsilon)}. \tag{4.72}$$

Combining (4.72) and (4.69) in (4.63)

$$\mathbb{P}\left(x_{i,k} \in G\right) \leq \sum_{s=0}^{k} C_\epsilon e^{-k\inf_{x\in G}\lambda^\top x + sN\Lambda\left(\left(\frac{1}{N}+\frac{1}{k}\right)\lambda\right) + (k-s)\Lambda(\lambda)} e^{-(k-s)(\mathcal{J}-\epsilon)} \tag{4.73}$$

$$\leq kC_\epsilon e^{-k\inf_{x\in G}\lambda^\top x} e^{\max_{1\leq s\leq k} sN\Lambda\left(\left(\frac{1}{N}+\frac{1}{k}\right)\lambda\right) + (k-s)(\Lambda(\lambda)-(\mathcal{J}-\epsilon))} \tag{4.74}$$

$$\leq ke^{-k\inf_{x\in G}\lambda^\top x} e^{k\max\left\{N\Lambda\left(\left(\frac{1}{N}+\frac{1}{k}\right)\lambda\right),(\Lambda(\lambda)-(\mathcal{J}-\epsilon))\right\}}. \tag{4.75}$$

We next prove the lower bound (4.58).

4.5.4 Proof of the lower bound (4.58)

Fix a collection $\mathcal{H} \subseteq \mathcal{G}$ and consider the following sequence of events

$$\mathcal{E}^\theta_{\mathcal{H},k} = \Bigg\{G_t \in \mathcal{H}, , \lceil\theta k\rceil + 1 \leq t \leq k, \ \left\|[\Phi(k, k - \lceil\sqrt{k}\rceil)]_{C_{i,\mathcal{H}}} - J_M\right\| < \frac{1}{k},$$
$$\left\|\Phi(\lceil\theta k\rceil, \lceil\theta k\rceil - \lceil\sqrt{k}\rceil) - J_N\right\| < \frac{1}{k}\Bigg\}, \tag{4.76}$$

$k \geq 1$, where $\theta \in [0,1]$ and $M = |C_{i,\mathcal{H}}|$. We explain the motivation behind this construction. Fix k, θ and an outcome $\omega \in \mathcal{E}^\theta_{\mathcal{H},k}$, and denote $A_t = W_t(\omega)$, $t \geq 1$[10]. Let $\mathcal{T}_1$, $\mathcal{T}_2$, $\mathcal{T}_3$ and $\mathcal{T}_4$ denote the four disjoint

[10]We see from the definition of $\mathcal{E}^\theta_{\mathcal{H},k}$ that we don't need to consider the realizations A_t of all W_t's but only of those until time k.

intervals according to which $\mathcal{E}^{\theta}_{\mathcal{H},k}$ is defined,

$$\begin{aligned}
\mathcal{T}_1 &= \left\{ t \in [1,k] : 1 \le t \le \lceil \theta k \rceil - \lceil \sqrt{k} \rceil \right\} \\
\mathcal{T}_2 &= \left\{ t \in [1,k] : \lceil \theta k \rceil - \lceil \sqrt{k} \rceil + 1 \le t \le \lceil \theta k \rceil \right\} \\
\mathcal{T}_3 &= \left\{ t \in [1,k] : \lceil \theta k \rceil + 1 \le t \le k - \lceil \sqrt{k} \rceil \right\} \\
\mathcal{T}_4 &= \left\{ t \in [1,k] : k - \lceil \sqrt{k} \rceil + 1 \le t \le k \right\}.
\end{aligned}$$

Note that $\mathcal{T}_1, ..., \mathcal{T}_4$ constitute a partition of the discrete time interval $[1, k]$.

Lemma 4.24 For any $\omega \in \mathcal{E}^{\theta}_{\mathcal{H},k}$,

1. for $t \in \mathcal{T}_3 \cup \mathcal{T}_4$, $[\Phi(k,t)(\omega)]_{i,j} = 0$ for all $j \notin C_{i,\mathcal{H}}$;

2. for $t \in \mathcal{T}_3$, $[\Phi(k,t)(\omega)]_{i,j} \in [1/\left|C_{i,\mathcal{H}}\right| - 1/k, 1/\left|C_{i,\mathcal{H}}\right| + 1/k]$ for all $j \in C_{i,\mathcal{H}}$

3. for $t \in \mathcal{T}_1$, $[\Phi(k,t)(\omega)]_{i,j} \in [1/N - 1/k, 1/N + 1/k]$ for all $j (= 1, ..., N)$.

where $\Phi(k,k)(\omega) \equiv I_N$, and $\Phi(k,t)(\omega) := W_k(\omega) \cdots W_{t+1}(\omega)$, for $1 \le t < k$.

Proof Fix $\omega \in \mathcal{E}^{\theta}_{\mathcal{H},k}$ and denote $A_t = W_t(\omega)$, for $t = 1, ..., k$. To prove part 1, note that, because $\omega \mathcal{E}^{\theta}_{\mathcal{H},k}$, all A_t's in the interval $\mathcal{T}_3 \cup \mathcal{T}_4$ have their induced graphs, $G(A_t)$, in $\mathcal{H}$. Therefore, node i over this time interval can only communicate with the nodes in $C_{i,\mathcal{H}}$. Let $M = |C_{i,\mathcal{H}}|$. Without loss of generality[11], suppose that $C_{i,\mathcal{H}} = \{1, ..., M\}$. Hence, for any $t \in \mathcal{T}_3 \cup \mathcal{T}_4$ A_t has the following block diagonal form

$$A_t = \begin{bmatrix} [A_t]_{C_{i,\mathcal{H}}} & 0_{M \times (N-M)} \\ 0_{M \times (N-M)} & [A_t]_{V \setminus C_{i,\mathcal{H}}} \end{bmatrix},$$

where $[A_t]_{C_{i,\mathcal{H}}}$ is the $(M \times M)$ submatrix of A_t that corresponds to the nodes in $C_{i,\mathcal{H}}$, $0_{M \times (N-M)}$ is the $M \times (N - M)$ matrix of all zeros, and similarly for the two remaining blocks. It follows that each $\Phi(k,t) = A_k \cdots A_t$, $t \in \mathcal{T}_3 \cup \mathcal{T}_4$, has the same sparsity pattern, and thus

$$\Phi(k,t) = \begin{bmatrix} [\Phi(k,t)]_{C_{i,\mathcal{H}}} & 0_{M \times (N-M)} \\ 0_{M \times (N-M)} & [\Phi(k,t)]_{V \setminus C_{i,\mathcal{H}}} \end{bmatrix},$$

proving part 1.

[11] We can always find a permutation matrix that reduces A_t into a block diagonal form.

To prove part 2, consider the second property of the A_t's: $\left\|[\Phi(k, k - \left\lceil\sqrt{k}\right\rceil)]_{C_i,\mathcal{H}} - J_M\right\| < \frac{1}{k}$. Due to the following monotonicity: $\|AD - J_M\| \leq \|A - J_M\|$ that holds for any A and D of size M that are doubly stochastic, we have here that for all $t \in \mathcal{T}_3$

$$\left\|[\Phi(k,t)]_{C_i,\mathcal{H}} - J_M\right\| < \frac{1}{k}.$$

Finally, exploiting the property of the spectral norm by which for any matrix D $|D_{ij}| \leq \|D\|$, for all entries i, j of D, completes the proof of part 2.

To prove part 3, note that that for all $t \in \mathcal{T}_1$ (in fact for all $t < k$)

$$\Phi(k,t) = \Phi\left(k, \lceil\theta k\rceil\right) \Phi\left(\lceil\theta k\rceil, \lceil\theta k\rceil - \lceil\sqrt{k}\rceil\right) \Phi\left(\lceil\theta k\rceil - \lceil\sqrt{k}\rceil, t\right).$$

Using now the fact that $\|ADC - J_N\| \leq \|D - J_N\|$ for doubly stochastic A, D and C of size N, applied to the product of the three matrices above,

$$\|\Phi(k,t) - J_N\| \leq \Phi\left(\lceil\theta k\rceil, \lceil\theta k\rceil - \lceil\sqrt{k}\rceil\right) < \frac{1}{k}, \tag{4.77}$$

which holds for all $t \in \mathcal{T}_1$.

Similarly as in part 2, it follows that $[\Phi(k,t)]_i j \in [1/N - 1/k, 1/N + 1/k]$, for each entry i, j of $[\Phi(k,t)]$, completing the proof of part 3 and the proof of Lemma 4.24. □

Define now the probability distribution $\nu_k : \mathcal{B}\left(\mathbb{R}^d\right) \to [0,1]$ by

$$\nu_k(E) = \frac{\mathbb{P}\left(\{x_{i,k} \in E\} \cap \left\{\mathcal{E}^{\theta}_{\mathcal{H},k}\right\}\right)}{\mathbb{P}\left(\mathcal{E}^{\theta}_{\mathcal{H},k}\right)}, \tag{4.78}$$

that is, ν_k is the conditional probability distribution of $x_{i,k}$ on $\mathcal{E}^{\theta}_{\mathcal{H},k}$; we remark that $\mathbb{P}\left(\mathcal{E}^{\theta}_{\mathcal{H},k}\right) > 0$ for k sufficiently large, as we show later in this proof, see Lemma 4.26. Let Υ_k be the (normalized) logarithmic moment generating function associated with ν_k defined by

$$\Upsilon_k(\lambda) = \frac{1}{k} \log \mathbb{E}\left[e^{k\lambda^\top x_{i,k}} | \mathcal{E}^{\theta}_{\mathcal{H},k}\right], \tag{4.79}$$

for $\lambda \in \mathbb{R}^d$.

Using the properties of the entries of $\Phi(k,t)$ in intervals $\mathcal{T}_1, ..., \mathcal{T}_4$, we establish in Lemma 4.25 that the sequence of functions Υ_k has a point-wise limit, which will allow us to apply the Gärtner-Ellis Theorem to

get a large deviations lower bound for the sequence ν_k. We first state and prove Lemma 4.25.

Lemma 4.25 For any $\lambda \in \mathbb{R}^d$:

$$\lim_{k\to+\infty} \Upsilon_k(\lambda) = (1-\theta)M\Lambda\left(1/M\lambda\right) + \theta N\Lambda\left(1/N\lambda\right). \tag{4.80}$$

Proof Starting from the expected value in (4.79):

$$\begin{aligned}
\mathbb{E}\left[e^{k\lambda^\top x_{i,k}}|\mathcal{E}^\theta_{\mathcal{H},k}\right] &= \frac{1}{\mathbb{P}\left(\mathcal{E}^\theta_{\mathcal{H},k}\right)}\mathbb{E}\left[1_{\mathcal{E}^\theta_{\mathcal{H},k}} e^{k\lambda^\top x_{i,k}}\right] \\
&= \frac{1}{\mathbb{P}\left(\mathcal{E}^\theta_{\mathcal{H},k}\right)}\mathbb{E}\left[\mathbb{E}\left[1_{\mathcal{E}^\theta_{\mathcal{H},k}} e^{k\lambda^\top x_{i,k}}|W_1,...,W_k\right]\right] \qquad (4.81)\\
&= \frac{1}{\mathbb{P}\left(\mathcal{E}^\theta_{\mathcal{H},k}\right)}\mathbb{E}\left[1_{\mathcal{E}^\theta_{\mathcal{H},k}} \mathbb{E}\left[e^{k\lambda^\top x_{i,k}}|W_1,...,W_k\right]\right] \qquad (4.82)
\end{aligned}$$

where in the last equality we used that, given $W_1,...,W_k$, the indicator $1_{\mathcal{E}^\theta_{\mathcal{H},k}}$ is a constant.

Now, fix an outcome $\omega \in \mathcal{E}^\theta_{\mathcal{H},k}$, and let $A_t = W_t(\omega)$, $t = 1,..,k$; focusing on the conditional expectation in the preceding equality for $W_t = A_t$, $t = 1,...,k$, and using the independence of the $Z_{i,t}$'s

$$\mathbb{E}\left[e^{k\lambda^\top x_{i,k}}|W_1 = A_1,...,W_k = A_k\right] = e^{\sum_{t=1}^k \sum_{j=1}^N \Lambda([\Phi(k,t)]_{ij}\lambda)}, \tag{4.83}$$

where $\Phi(k,t) = A_k \cdots A_{t+1}$, for $t = 1,...,k$. We split the sum in the exponent of (4.83) over the disjoint intervals $\mathcal{T}_1,...,\mathcal{T}_4$ that cover $[1,k]$, and analyze each of the obtained summands separately. First, for each $\lambda \in \mathbb{R}^d$, define

$$\overline{h}_k(\lambda) := \max_{\phi\in[1/M-1/k,1/M+1/k]} \Lambda\left(\phi\lambda\right) \quad \underline{h}_k(\lambda) := \min_{\phi\in[1/M-1/k,1/M+1/k]} \Lambda\left(\phi\lambda\right),$$

and

$$\overline{g}_k(\lambda) := \max_{\phi\in[1/N-1/k,1/N+1/k]} \Lambda\left(\phi\lambda\right) \quad \underline{g}_k(\lambda) := \min_{\phi\in[1/N-1/k,1/N+1/k]} \Lambda\left(\phi\lambda\right),$$

for $\lambda \in \mathbb{R}^d$. Note that $\overline{h}_k, \underline{h}_k \to \Lambda\left(1/N\cdot\right)$, as $k \to +\infty$ pointwise (for every λ); likewise, $\overline{g}_k, \underline{g}_k \to \Lambda\left(1/M\cdot\right)$, as $k \to +\infty$, pointwise.

Then, by part 2 of Lemma 4.24,

$$M\underline{h}_k(\lambda) \leq \sum_{j\in C_{i,\mathcal{H}}} \Lambda\left([\Phi(k,t)]_{ij}\lambda\right) \leq M\overline{h}_k(\lambda), \quad \text{for each } t \in \mathcal{T}_3,$$

and, similarly, by part 3 of Lemma 4.24

$$N\underline{g}_k(\lambda) \leq \sum_{j=1}^{N} \Lambda\left([\Phi(k,t)]_{ij}\lambda\right) \leq N\overline{g}_k(\lambda), \quad \text{for each } t \in \mathcal{T}_1.$$

As for the summands in the intervals $\mathcal{T}_2$ and $\mathcal{T}_4$, we apply Proposition 4.5 to get

$$M\Lambda\left(1/M\lambda\right) \leq \sum_{j\in C_{i,\mathcal{H}}} \Lambda\left([\Phi(k,t)]_{ij}\lambda\right) \leq \Lambda(\lambda), \quad \text{for each } t \in \mathcal{T}_4,$$

and

$$N\Lambda\left(1/N\lambda\right) \leq \sum_{j=1}^{N} \Lambda\left([\Phi(k,t)]_{ij}\lambda\right) \leq \Lambda(\lambda), \quad \text{for each } t \in \mathcal{T}_2.$$

Noting that $|\mathcal{T}_1| = \lceil\theta k\rceil - \lceil\sqrt{k}\rceil$, $|\mathcal{T}_3| = k - \lceil\theta k\rceil - \lceil\sqrt{k}\rceil$, and $|\mathcal{T}_2| = |\mathcal{T}_4| = \lceil\sqrt{k}\rceil$, and summing out the upper and lower bounds over all t's in the preceding four inequalities yields:

$$k\underline{\Upsilon}_k\left(\lambda\right) \leq \sum_{t=1}^{k}\sum_{i=1}^{N} \Lambda\left([\Phi(k,t)]_{i,j}\right) \leq k\overline{\Upsilon}_k\left(\lambda\right), \tag{4.84}$$

where

$$\underline{\Upsilon}_k\left(\lambda\right) = \frac{\lceil\theta k\rceil - \lceil\sqrt{k}\rceil}{k} N\underline{g}_k(\lambda) + \frac{\lceil\sqrt{k}\rceil}{k}\left(N\Lambda\left(1/N\lambda\right) + M\Lambda\left(1/M\lambda\right)\right) + \frac{k - \lceil\theta k\rceil - \lceil\sqrt{k}\rceil}{k} M\underline{h}_k(\lambda),$$

and

$$\overline{\Upsilon}_k(\lambda) = \frac{\lceil\theta k\rceil - \lceil\sqrt{k}\rceil}{k} N\overline{g}_k(\lambda) + \frac{\lceil\sqrt{k}\rceil}{k}\left(N\Lambda\left(1/N\lambda\right) + M\Lambda\left(1/M\lambda\right)\right) + \frac{k - \lceil\theta k\rceil - \lceil\sqrt{k}\rceil}{k} M\overline{h}_k(\lambda).$$

Since the preceding inequality holds for any fixed $\omega \in \mathcal{E}^{\theta}_{\mathcal{H},k}$, it follows that

$$1_{\mathcal{E}^{\theta}_{\mathcal{H},k}} e^{k\underline{\Upsilon}_k(\lambda)} \leq 1_{\mathcal{E}^{\theta}_{\mathcal{H},k}} \mathbb{E}\left[e^{k\lambda^\top x_{i,k}} | W_1, ..., W_k\right] \leq 1_{\mathcal{E}^{\theta}_{\mathcal{H},k}} e^{k\overline{\Upsilon}_k(\lambda)}. \tag{4.85}$$

Finally, by the monotonicity of the expectation:

$$\mathbb{P}\left(\mathcal{E}^{\theta}_{\mathcal{H},k}\right) e^{k\underline{\Upsilon}_k(\lambda)}\mathbb{E}\left[1_{\mathcal{E}^{\theta}_{\mathcal{H},k}}\mathbb{E}\left[e^{k\lambda^\top x_{i,k}}|W_1,...,W_k\right]\right] \leq \mathbb{P}\left(\mathcal{E}^{\theta}_{\mathcal{H},k}\right) e^{k\overline{\Upsilon}_k(\lambda)},$$

which combined with (4.81) implies

$$e^{k\underline{\Upsilon}_k(\lambda)} \leq \mathbb{E}\left[e^{k\lambda^\top x_{i,k}}|\mathcal{E}^{\theta}_{\mathcal{H},k}\right] \leq e^{k\overline{\Upsilon}_k(\lambda)} \tag{4.86}$$

Now, taking the logarithm and dividing by k,

$$\overline{\Upsilon}_k(\lambda) \leq \Upsilon_k(\lambda) \leq \overline{\Upsilon}_k(\lambda),$$

and noting that

$$\lim_{k\to+\infty}\overline{\Upsilon}_k(\lambda) = \lim_{k\to+\infty}\overline{\Upsilon}_k(\lambda) = (1-\theta)M\Lambda\left(1/M\lambda\right) + (\theta)N\Lambda\left(1/N\lambda\right),$$

the claim of Lemma 4.25 follows by the sandwiching argument. □

Now, remark that the limiting function $\lambda \mapsto \lim_{k\to}\Upsilon_k(\lambda)$ is finite everywhere. By the Gärtner-Ellis theorem it follows then that the sequence of measures ν_k satisfies the large deviation principle[12]. Therefore, we have that for every open set $F \subseteq \mathbb{R}^d$,

$$\lim_{k\to+\infty}\frac{1}{k}\log\mathbb{P}\left(x_{i,k}\in F|\mathcal{E}^{\theta}_{\mathcal{H},k}\right) = -\inf_{x\in E}\sup_{\lambda\in\mathbb{R}^d}\lambda^\top - (1-\theta)M\Lambda\left(1/N\lambda\right) - \theta N\Lambda\left(1/N\lambda\right). \tag{4.87}$$

We next turn to computing, more precisely, tightly approximating, the probability of the event $\mathcal{E}^{\theta}_{\mathcal{H},k}$, $\mathbb{P}\left(\mathcal{E}^{\theta}_{\mathcal{H},k}\right)$.

Lemma 4.26 For any $\theta \in (0,1)$,

$$\frac{1}{4}p_{\mathcal{H}}^{k-\lceil\theta k\rceil} \leq \mathbb{P}\left(\mathcal{E}^{\theta}_{\mathcal{H},k}\right) \leq p_{\mathcal{H}}^{k-\lceil\theta k\rceil}, \tag{4.88}$$

[12]We use here the variant of the Gärtner-Ellis theorem that claims the (full) LDP for the case when the domain of the limiting function is the whole space $\mathbb{R}^d$; see Exercise 2.3.20 in [8] for the proof of this result.

Proof To start with, by the disjoint blocks theorem [69] applied to the matrices in the two disjoint intervals $\mathcal{T}_1 \cup \mathcal{T}_2$, and $\mathcal{T}_3 \cup \mathcal{T}_4$:

$$\mathbb{P}\left(\mathcal{E}^{\theta}_{\mathcal{H},k}\right) = \mathbb{P}\left(G_t \in \mathcal{H}, \text{ for all } t \in \mathcal{T}_3 \cup \mathcal{T}_4, \ \left\|[\Phi(k, k - \lceil\sqrt{k}\rceil)]_{C_i,\mathcal{H}} - J_M\right\| < \frac{1}{k}\right) \cdot \mathbb{P}\left(\left\|\Phi(\theta k, \theta k - \lceil\sqrt{k}\rceil) - J_N\right\| < \frac{1}{k}\right). \tag{4.89}$$

Conditioning in the first factor on the event $\{G_t \in \mathcal{H}, \text{ for all } t \in \mathcal{T}_3 \cup \mathcal{T}_4\}$ (the probability of which equals $p_{\mathcal{H}}^{k-\lceil\theta k\rceil}$ and thus is non-zero),

$$\mathbb{P}\left(G_t \in \mathcal{H}, \text{ for all } t \in \mathcal{T}_3 \cup \mathcal{T}_4, \ \left\|[\Phi(k, k - \lceil\sqrt{k}\rceil)]_{C_i,\mathcal{H}} - J_M\right\| < \frac{1}{k}\right) \tag{4.90}$$

$$= \mathbb{P}\left(\left\|[\Phi(k, k - \lceil\sqrt{k}\rceil)]_{C_i,\mathcal{H}} - J_M\right\| < \frac{1}{k} | G_t \in \mathcal{H}, \text{ for all } t \in \mathcal{T}_3 \cup \mathcal{T}_4\right) \mathbb{P}\left(G_t \in \mathcal{H}, t \in \mathcal{T}_3 \cup \mathcal{T}_4\right)$$

$$= \mathbb{P}\left(\left\|[\Phi(k, k - \lceil\sqrt{k}\rceil)]_{C_i,\mathcal{H}} - J_M\right\| < \frac{1}{k} | G_t \in \mathcal{H}, t \in \mathcal{T}_3 \cup \mathcal{T}_4\right) p_{\mathcal{H}}^{k-\lceil\theta k\rceil}. \tag{4.91}$$

Computing the probability through the complement

$$\mathbb{P}\left(\left\|[\Phi(k, k - \lceil\sqrt{k}\rceil)]_{C_i,\mathcal{H}} - J_M\right\| < \frac{1}{k} | G_t \in \mathcal{H}, \text{ for all } t \in \mathcal{T}_3 \cup \mathcal{T}_4\right) =$$

$$1 - \mathbb{P}\left(\left\|[\Phi(k, k - \lceil\sqrt{k}\rceil)]_{C_i,\mathcal{H}} - J_M\right\| \geq \frac{1}{k} | G_t \in \mathcal{H}, \text{ for all } t \in \mathcal{T}_3 \cup \mathcal{T}_4\right), \tag{4.92}$$

and we recognize in the righthand side the familiar event from Chapter 2: $\left\{\left\|[\Phi(k, k - \lceil\sqrt{k}\rceil)]_{C_i,\mathcal{H}} - J_M\right\| \geq \frac{1}{k}\right\}$. Using the results from Chapter 2, we can show that the probability in the righthand side goes to zero exponentially fast as $k \to 0$. Therefore, we can find k_0 such that for all $k \geq k_0$, this probability is smaller than $\frac{1}{2}$, which combined with (4.92) shows that the probability of the first factor in (4.89) is bounded between $\frac{1}{2}p_{\mathcal{H}}^{k-k-\lceil\theta k\rceil}$ and $p_{\mathcal{H}}^{k-k-\lceil\theta k\rceil}$ for all $k \geq k_0$. Similarly as with the first factor, we can show that the probability of the second factor in (4.89) is between $\frac{1}{2}$ and 1. The result follows by summarizing the preceding findings. □

To bring the two key arguments together, namely, Lemma 4.26 and the lower bound (4.87), we start from the simple relation

$$\mathbb{P}\left(x_{i,k} \in F\right) \geq \mathbb{P}\left(\{x_{i,k} \in F\} \cap \mathcal{E}^{\theta}_{\mathcal{H},k}\right).$$

Passing to the limit, and exploiting (4.88) and (4.87)

$$\begin{aligned}\liminf_{k\to+\infty}\frac{1}{k}\log\mathbb{P}\left(x_{i,k}\in F\right) &\geq \lim_{k\to+\infty}\frac{1}{k}\log\nu_k(F)+\lim_{k\to+\infty}\frac{1}{k}\log\mathbb{P}\left(\mathcal{E}^{\theta}_{\mathcal{H},k}\right)\\ &= -\inf_{x\in F}\sup_{\lambda\in\mathbb{R}^d}\lambda^{\top}x-(1-\theta)M\Lambda\left(1/M\lambda\right)-\theta N\Lambda\left(1/N\lambda\right)-(1-\theta)|\log p_{\mathcal{H}}|.\end{aligned}$$

Since θ is an arbitrary number in $[0,1]$), we optimize the last bound over $\theta\in[0,1]$:

$$\begin{aligned}&\liminf_{k\to+\infty}\frac{1}{k}\log\mathbb{P}\left(x_{i,k}\in F\right)\geq\\ &-\inf_{\theta\in[0,1]}\inf_{x\in F}\sup_{\lambda\in\mathbb{R}^d}\lambda^{\top}x-(1-\theta)M\Lambda\left(1/M\lambda\right)-\theta N\Lambda\left(1/N\lambda\right)-(1-\theta)|\log p_{\mathcal{H}}|\\ &=\inf_{x\in F}\inf_{\theta\in[0,1]}\sup_{\lambda\in\mathbb{R}^d}\lambda^{\top}x-(1-\theta)M\Lambda\left(1/M\lambda\right)-\theta N\Lambda\left(1/N\lambda\right)-(1-\theta)|\log p_{\mathcal{H}}|.\end{aligned}$$

Fix $x\in F$ and consider the function $f(\theta,\lambda):=\lambda^{\top}x-(1-\theta)M\Lambda\left(1/M\lambda\right)-\theta N\Lambda\left(1/N\lambda\right)-(1-\theta)|\log p_{\mathcal{H}}|$. Function f is convex in θ and concave in λ for every fixed $\theta\in[0,1]$. Also, sets $[0,1]$ and $\mathbb{R}^d$ are convex and set $[0,1]$ is compact. Thus, we can apply the Sion's Minimax theorem [85] to obtain $\inf\sup f=\sup\inf f$:

$$\begin{aligned}&\inf_{\theta\in[0,1]}\sup_{\lambda\in\mathbb{R}^d}\lambda^{\top}x-(1-\theta)M\Lambda\left(1/M\lambda\right)-\theta N\Lambda\left(1/N\lambda\right)-(1-\theta)|\log p_{\mathcal{H}}|=\\ &\sup_{\lambda\in\mathbb{R}^d}\inf_{\theta\in[0,1]}\lambda^{\top}x-(1-\theta)M\Lambda\left(1/M\lambda\right)-\theta N\Lambda\left(1/N\lambda\right)-(1-\theta)|\log p_{\mathcal{H}}|\\ &=\sup_{\lambda\in\mathbb{R}^d}\lambda^{\top}x-\max\left\{M\Lambda\left(1/M\lambda\right)-|\log p_{\mathcal{H}}|,\Lambda\left(1/N\lambda\right)\right\}.\end{aligned}$$

Similarly as in the proof of the upper bound, using properties of conjugation (the conjugate of the maximum of two functions is the convex hull of their respective conjugates,

$$\sup_{\lambda\in\mathbb{R}^d}\lambda^{\top}x-\min\left\{M\Lambda\left(1/M\lambda\right)-|\log p_{\mathcal{H}}|,\Lambda\left(1/N\lambda\right)\right\}=\overline{\mathrm{co}}\left(NI(x),MI(x)+|\log p_{\mathcal{H}}|\right),\tag{4.93}$$

and thus,

$$\liminf_{k\to+\infty}\frac{1}{k}\log\mathbb{P}\left(x_{i,k}\in F\right)\geq-\inf_{x\in F}\overline{\mathrm{co}}\left(NI(x),MI(x)+|\log p_{\mathcal{H}}|\right).\tag{4.94}$$

Optimizing now over all $\mathcal{H}\subseteq\mathcal{G}$, yields the lower bound. This completes the proof of the lower bound and the proof of Theorem 4.19.

4.6 Distributed estimation

Consider distributed estimation where each agent i acquires over time t scalar measurements $y_{i,t}$ with:

$$y_{i,t} = a_i^\top \theta + n_{i,t}. \tag{4.95}$$

Here, $a_i \in \mathbb{R}^d$ is a deterministic vector known to agent i, $\theta \in \mathbb{R}^d$ is the unknown parameter to be estimated, and $n_{i,t}$ is the zero mean additive Gaussian, spatio-temporally uncorrelated noise, with variance equal to 1.

Centralized estimator. For benchmarking purpose, we consider the minimum variance unbiased estimator based on k measurements from all N agents:

$$\widehat{\theta}_{ML,k} = \left(\sum_{i=1}^{N} a_i a_i^\top \right)^{-1} \left(\frac{1}{k} \sum_{t=1}^{k} \sum_{i=1}^{N} a_i y_{i,t} \right). \tag{4.96}$$

Distributed estimator. The estimator assumes that all agents beforehand know the matrix $A := \sum_{i=1}^{N} a_i a_i^\top$. For example, the matrix A can be computed by a consensus algorithm (with $N(N+1)/2$-dimensional variables). Each agent i updates its estimate $x_{i,k}$ of the parameter θ as follows:

$$x_{i,k+1} = \frac{k}{k+1} \sum_{j \in O_{i,k}} [W_k]_{ij} x_{j,k} + \frac{1}{k+1} N A^{-1} a_i y_{i,k+1}. \tag{4.97}$$

Large deviations performance. We are interested in the inaccuracy rates [9]:

$$\widehat{I}(\xi) := \lim_{k \to \infty} -\frac{1}{k} \log \mathbb{P} \left(\|\theta_k - \theta\| \geq \xi \right), \ \xi > 0, \tag{4.98}$$

where θ_k is the estimator after k per-agent samples are processed; e.g., $\theta_k = \widehat{\theta}_{ML,k}$ for the centralized estimator (4.96), and $\theta_k = x_{i,k}$ for the distributed estimator (4.97). It can be shown that the inaccuracy rate for the centralized estimator (4.96) is given by:

$$\widehat{I}^{(\mathrm{cen})}(\xi) = \frac{\xi^2}{2} \lambda_{\min} (A),$$

where $\lambda_{\min}(\cdot)$ denotes the minimal eigenvalue.

In Theorem 4.27 we derive an upper bound for the inaccuracy rate of the distributed estimation algorithm 4.97. We omit the proof of Theorem 4.27, but we remark that the proof exploits the same techniques as in the proof of the upper bound of Theorem 4.19.

Theorem 4.27 For any node i, for any $\xi \geq 0$

$$\limsup_{k \to +\infty} \frac{1}{k} \log \mathbb{P}\left(\|x_{i,k} - \theta\| \geq \xi\right) \leq - \inf_{x \in \mathbb{R}^d \,:\, \|x-\theta\| \geq \xi} \overline{\mathrm{co}} \left(I_1(x) + \mathcal{J}, \ldots, I_N(x) + \mathcal{J}, I^{(\mathrm{cen})}(x)\right) \quad (4.99)$$

where, for $j = 1, \ldots, N$, $I_j(x) = \sup_{\lambda \in \mathbb{R}^d} \lambda^\top x - \frac{N^2}{2} \left(\lambda^\top A^{-1} a_j\right)^2 - N \lambda^\top A^{-1} a_j a_j^\top \theta$, and $I^{(\mathrm{cen})}(x) = \frac{1}{2} \left(x - \theta\right)^\top A \left(x - \theta\right)$.

Chapter 5

Large Deviations for Distributed Detection

In this chapter, we analyze the large deviations performance of distributed detection where agents in a network decide between the two alternative hypothesis H_1 versus H_0. Each agent i improves its (scalar) state $x_{i,k}$ over time k through a distributed detection algorithm of type (4.3)-(4.3), and makes its local decision by comparing $x_{i,k}$ with a zero threshold.[1] Detection performance at agent i and time k is characterized by the probability of false alarm, probability of miss, and average error probability, respectively given by:

$$\alpha_{i,k} = \mathbb{P}\,(\, x_{i,k} \geq 0 \mid H_0\,) = \mathbb{P}\,(\, x_{i,k} \in [\,0, \infty) \mid H_0\,) \tag{5.1}$$

$$\beta_{i,k} = \mathbb{P}\,(\, x_{i,k} < 0 \mid H_1\,) = \mathbb{P}\,(\, x_{i,k} \in (-\infty, 0) \mid H_1\,) \tag{5.2}$$

$$P^{\mathrm{e}}_{i,k} = \pi_0\, \alpha_{i,k} + \pi_1\, \beta_{i,k}, \tag{5.3}$$

where π_1 and π_0 are the prior probabilities. Conditioned on, say, H_1, the state $x_{i,k}$ (under appropriate conditions) converges in probability to a positive number γ_1, so that the probability mass "escapes" from the set $(-\infty, 0)$, and thus $\beta_{i,k}$ converges to zero. Likewise, conditioned on H_0, $x_{i,k}$ converges in probability to a negative number γ_0 and the mass "escapes" from the set $[\,0, \infty)$, thus implying $\alpha_{i,k} \to 0$. In the language of large deviations from Chapter 4, the sets $(-\infty, 0)$ and $[\,0, \infty)$ are deviation sets. In this Chapter, our goal is to find the large deviation rates:

$$\alpha_{i,k} = \mathbb{P}\,(\, x_{i,k} \in [\,0, \infty) \mid H_0\,) \sim e^{-k\, I_{\alpha,i}} \tag{5.4}$$

$$\beta_{i,k} = \mathbb{P}\,(\, x_{i,k} \in (-\infty, 0) \mid H_1\,) \sim e^{-k\, I_{\beta,i}}, \tag{5.5}$$

[1] For simplicity of introductory explanation, we take here a zero threshold, but the chapter analyzes generic thresholds γ.

and to ultimately determine the decay rate of the average error probability:

$$P_{i,k}^{\mathrm{e}} \sim e^{-k\, I_i}. \tag{5.6}$$

We refer to rate I_i as the asymptotic performance of agent i.

We quantify the asymptotic performance in terms of the agents' observations and the underlying random network. Our results reveal a nonlinear, phase transition type behavior with respect to the network connectivity, measured by the quantity $\mathcal{J}$ in (2.1) from Chapter 2[2]. When the network connectivity $\mathcal{J}$ is above a threshold, then the distributed detector is asymptotically optimal at each agent i, i.e., asymptotically equivalent to the optimal centralized detector that collects the observations of all agents. When $\mathcal{J}$ is below the threshold, we quantify what fraction of the centralized performance distributed detector can achieve. Hence, there exists a "sufficient" connectivity $\mathcal{J}^\star$, such that asymptotic detection performance saturates at the optimal (centralized) value for any $\mathcal{J} \geq \mathcal{J}^\star$. Translated in practical terms, once $\mathcal{J}^\star$ is achieved, e.g., by increasing the amount of invested resources, further increase does not pay off. In this chapter, we address the design problem of "targeting" the point $\mathcal{J}^\star$ in a wireless sensor network, where the inter-agent probabilities (and hence, $\mathcal{J}$) of successful communication depend on the invested transmission power. We optimize the agents' transmission powers, such that connectivity $\mathcal{J}^\star$ is achieved with the minimal overall investment.

We discover a very interesting interplay between the distribution of the agent observations (e.g., Gaussian or Laplace) and the connectivity $\mathcal{J}$ of the network. For a network with the same connectivity, a distributed detector with say, Laplace observations distributions, may match the optimal asymptotic performance of the centralized detector, while the distributed detector for Gaussian observations may be suboptimal, even though the centralized detectors for the two distributions, Laplace and Gaussian, have the same optimal asymptotic performance.

For distributed detection, we determine the range on the detection threshold γ for which each agent achieves exponentially fast decay of the error probability (strictly positive error exponent), and we find the optimal γ that maximizes the error exponent. Above the critical (phase transition) value for the network connectivity $\mathcal{J}$, the optimal detector threshold is $\gamma^\star = 0$, mimicking the (asymptotically) optimal threshold for the centralized detector. Below the critical connectivity, the optimal distributed detector threshold may be non zero.

We remark that the results presented in this chapter are based on our work in [63, 64, 62, 86].

Brief review of the literature. A large body of work on distributed detection considers *fusion center (FC)*-

[2]For convenience, in this chapter, we refer to quantity $\mathcal{J}$ simply as the network connectivity.

based architectures, e.g., [43, 44, 45, 46, 47, 42], and, recently, e.g., [48, 87, 88, 89]: [48, 87, 88] consider the problem of selecting a subset of agents that optimizes detection performance at the FC; and [89] optimizes the local linear pre-coding of the agents' messages to the FC, to optimize detection performance subject to a transmit power constraint. References [49, 50, 51, 52] study *consensus-based* detection. *Consensus+innovations* estimation is considered by references [53, 2, 54, 55, 56], while several mutually different variants of *consensus+innovations* detection are studied in [57, 58, 1, 5, 59, 60, 61]. We analyze here running consensus, the variant in [1].

Reference [1] considers asymptotic optimality of running consensus, but in a framework that is very different from ours. Reference [1] studies the asymptotic performance of the distributed detector where the means of the agent observations under the two hypothesis become closer and closer (vanishing signal to noise ratio (SNR)), at the rate of $1/\sqrt{k}$, where k is the number of observations. For this problem, there is an asymptotic, non-zero, probability of miss and an asymptotic, non-zero, probability of false alarm. Under these conditions, running consensus is as efficient as the optimal centralized detector, [66], as long as the network is connected on average. Here, we assume that the means of the distributions stay fixed as k grows. We establish, through large deviations, the rate (error exponent) at which the error probability decays to zero as k goes to infinity. We show that connectedness on average is not sufficient for running consensus to achieve the optimality of centralized detection; rather, phase transition occurs, with distributed becoming as good as centralized, when the network connectivity, measured by $\mathcal{J}$, exceeds a certain threshold.

Chapter outline. Section 5.1 introduces the network and agent observations models and presents distributed detector. Section 5.2 presents and proves our main results on the asymptotic performance of the distributed detector. For a cleaner exposition, this section proves the results for (spatially) identically distributed agent observations. Section 5.3 illustrates our results on several types of agent observation distributions, namely, Gaussian, Laplace, and discrete valued distributions, discussing the impact of these distributions on distributed detection performance. Section 5.4 provides elaborate simulation results for the setup of correlated Gaussian observations and a generic random network. Section 5.5 extends our main results to non-identically distributed agents' observations. Finally, Section 5.6 addresses the problem of optimal power allocation for distributed detection.

Notation. We denote by: A_{ij} the (i,j)-th entry of a matrix A; a_i the i-th entry of a vector a; I, 1, and e_i, respectively, the identity matrix, the column vector with unit entries, and the i-th column of I; J the $N \times N$ ideal consensus matrix $J := (1/N)11^\top$; $\|\cdot\|_l$ the vector (respectively, matrix) l-norm of its vector (respectively, matrix) argument; $\|\cdot\| = \|\cdot\|_2$ the Euclidean (respectively, spectral) norm of its vector (respectively, matrix) argument; $\mu_i(\cdot)$ the i-th largest eigenvalue; $\mathbb{E}[\cdot]$ and $\mathbb{P}(\cdot)$ the expected value and

probability operators, respectively; $\mathcal{I}_{\mathcal{A}}$ the indicator function of the event $\mathcal{A}$; ν^N the product measure of N i.i.d. observations drawn from the distribution with measure ν; $h'(z)$ and $h''(z)$ the first and the second derivatives of the function h at point z.

5.1 Problem formulation

This section introduces the agent observations model, reviews the optimal centralized detector, and presents the distributed detector. The section also reviews relevant properties of the log-moment generating function of a agent's log-likelihood ratio that are needed in the sequel.

5.1.1 Agent observations model

We study the binary hypothesis testing problem H_1 versus H_0. We consider a network of N agents where $Y_{i,t}$ is the observation of agent i at time t, where $i = 1, \ldots, N$, $t = 1, 2, \ldots$

Assumption 5.1 The agents' observations $\{Y_{i,t}\}$ are independent and identically distributed (i.i.d.) both in time and in space, with distribution ν_1 under hypothesis H_1 and ν_0 under H_0:

$$Y_{i,t} \sim \begin{cases} \nu_1, & H_1 \\ \nu_0, & H_0 \end{cases}, i = 1, \ldots, N,\ t = 1, 2, \ldots \tag{5.7}$$

Here ν_1 and ν_0 are mutually absolutely continuous, distinguishable measures. The prior probabilities $\pi_1 = \mathbb{P}(H_1)$ and $\pi_0 = \mathbb{P}(H_0) = 1 - \pi_1$ are in $(0, 1)$.

By spatial independence, the joint distribution of the observations of all agents

$$Y_t := (Y_{1,t}, \ldots, Y_{N,t})^\top \tag{5.8}$$

at any time t is ν_1^N under H_1 and ν_0^N under H_0. Our main results in Section III are derived under Assumption 5.1. Section V extends them to non-identical (but still independent) agents' observations.

5.1.2 Centralized detection, logarithmic moment generating function, and optimal error exponent

The log-likelihood ratio of agent i at time t is $L_{i,t}$ and given by

$$L_{i,t} = \log \frac{f_1\left(Y_{i,t}\right)}{f_0\left(Y_{i,t}\right)},$$

where, $f_l(\cdot)$, $l = 0, 1$, is 1) the probability density function corresponding to ν_l, when $Y_{i,t}$ is an absolutely continuous random variable; or 2) the probability mass function corresponding to ν_l, when $Y_{i,t}$ is discrete valued.

Under Assumption 5.1, the log-likelihood ratio test for k time observations from all agents, for a threshold γ is: [3]

$$D_k := \frac{1}{Nk} \sum_{t=1}^{k} \sum_{i=1}^{N} L_{i,t} \underset{H_0}{\overset{H_1}{\gtrless}} \gamma. \tag{5.9}$$

Logarithmic moment generating function.

Let Λ_l ($l = 0, 1$) denote the logarithmic moment generating function for the log-likelihood ratio under hypothesis H_l:

$$\Lambda_l : \mathbb{R} \longrightarrow (-\infty, +\infty], \quad \Lambda_l(\lambda) = \log \mathbb{E}\left[e^{\lambda L_1(1)} \mid H_l\right]. \tag{5.10}$$

In (5.10), $L_1(1)$ replaces $L_{i,t}$, for arbitrary $i = 1, ..., N$, and $t = 1, 2, ...$, due to the spatial and temporal identically distributed observations, see Assumption 5.1. Recall that we already studied logarithmic moment generating functions in Chapter 4, and gave several properties of these functions. Here, we are interested in the logarithmic moment generating functions for log-likelihood ratios; these functions enjoy additional properties with respect to those in Chapter 4, as we state below. (We repeat here the convexity property for convenience.)

Lemma 5.2 Consider Assumption 5.1. For Λ_0 and Λ_1 in (5.10) the following holds:

(a) Λ_0 is convex;

(b) $\Lambda_0(\lambda) \in (-\infty, 0)$, for $\lambda \in (0, 1)$, $\Lambda_0(0) = \Lambda_0(1) = 0$, and $\Lambda_l'(0) = \mathbb{E}\left[L_1(1)|H_l\right]$, $l = 0, 1$;

(c) $\Lambda_1(\lambda)$ satisfies:

$$\Lambda_1(\lambda) = \Lambda_0(\lambda + 1), \quad \text{for} \ \ \lambda \in \mathbb{R}. \tag{5.11}$$

[3] In (5.9), we re-scale the spatio-temporal sum of the log-likelihood ratios $L_{i,t}$ by dividing the sum by Nk. Note that we can do so without loss of generality, as the alternative test without re-scaling is: $\sum_{t=1}^{k} \sum_{i=1}^{N} L_{i,t} \underset{H_0}{\overset{H_1}{\gtrless}} \gamma'$, with $\gamma' = Nk\gamma$.

Proof For a proof of (a) and (b), see [18]. Part (c) follows from the definitions of Λ_0 and Λ_1, which we show here for the case when the distributions ν_1 and ν_0 are absolutely continuous (the proof for discrete distributions is similar):

$$\Lambda_1(\lambda) = \log \mathbb{E}\left[e^{\lambda L_1(1)}|H_1\right] = \log \int_{y\in\mathbb{R}} \left(\frac{f_1(y)}{f_0(y)}\right)^{\lambda} f_1(y)dy = \log \int_{y\in\mathbb{R}} \left(\frac{f_1(y)}{f_0(y)}\right)^{1+\lambda} f_0(y)dy = \Lambda_0(1+\lambda).$$

□

We further assume that the logarithmic moment generating function of a agent's observation is finite (just like we assumed in Chapter 4).

Assumption 5.3 $\Lambda_0(\lambda) < +\infty, \forall \lambda \in \mathbb{R}$.

In the next two remarks, we give two classes of problems when Assumption 5.3 holds.

Remark I. We consider the signal+noise model:

$$Y_{i,t} = \begin{cases} m + n_{i,t}, & H_1 \\ n_{i,t}, & H_0. \end{cases} \tag{5.12}$$

Here $m \neq 0$ is a constant signal and $n_{i,t}$ is a zero-mean additive noise with density function $f_n(\cdot)$ supported on $\mathbb{R}$; we rewrite $f_n(\cdot)$, without loss of generality, as $f_n(y) = c\, e^{-g(y)}$, where $c > 0$ is a constant. It can be shown that Assumption 5.3 holds under the following mild technical condition: either one of (5.13) or (5.14) *and* one of (5.15) or (5.16) hold:

$$\lim_{y\to+\infty} \frac{g(y)}{|y|^{\tau_+}} = \rho_+, \text{ for some } \rho_+, \tau_+ \in (0, +\infty) \tag{5.13}$$

$$\lim_{y\to+\infty} \frac{g(y)}{(\log(|y|))^{\mu_+}} = \rho_+, \text{ for some } \rho_+ \in (0, +\infty),\ \mu_+ \in (1, +\infty) \tag{5.14}$$

$$\lim_{y\to-\infty} \frac{g(y)}{|y|^{\tau_-}} = \rho_-, \text{ for some } \rho_-, \tau_- \in (0, +\infty) \tag{5.15}$$

$$\lim_{y\to-\infty} \frac{g(y)}{(\log(|y|))^{\mu_-}} = \rho_-, \text{ for some } \rho_- \in (0, +\infty),\ \mu_- \in (1, +\infty). \tag{5.16}$$

In (5.14) and (5.16), we can also allow either (or both) μ_+, μ_- to equal 1, but then the corresponding ρ is in $(1, \infty)$. Note that $f_n(\cdot)$ need not be symmetric, i.e., $f_n(y)$ need not be equal to $f_n(-y)$. Intuitively, the tail of the density $f_n(\cdot)$ behaves regularly, and $g(y)$ grows either like a polynomial of arbitrary finite order in y, or slower, like a power y^τ, $\tau \in (0, 1)$, or like a logarithm $c(\log y)^\mu$. The class of admissible densities $f_n(\cdot)$ includes, e.g., power laws cy^{-p}, $p > 1$, or the exponential families $e^{\theta\,\phi(y) - A(\theta)}$, $A(\theta) :=$

$\log \int_{y=-\infty}^{+\infty} e^{\theta\phi(y)}\chi(dy)$, with: 1) the Lebesgue base measure χ; 2) the polynomial, power, or logarithmic potentials $\phi(\cdot)$; and 3) the canonical set of parameters $\theta \in \Theta = \{\theta :\ A(\theta) < +\infty\}$, [90].

Remark II. Assumption 5.3 is satisfied if $Y_{i,k}$ has arbitrary (different) distributions under H_1 and H_0 with the same, compact support; a special case is when $Y_{i,k}$ is discrete, supported on a finite alphabet.

Centralized detection: Asymptotic performance. We consider briefly the performance of the centralized detector that will benchmark the performance of the distributed detector. Denote by $\gamma_l := \mathbb{E}\left[L_{1,1}|H_l\right]$, $l = 0, 1$. It can be shown [8] that $\gamma_0 < 0$ and $\gamma_1 > 0$. Now, consider the centralized detector in (5.9) with the constant thresholds γ, for all k, and denote by:

$$\alpha_k(\gamma) = \mathbb{P}\left(D_k \geq \gamma | H_0\right),\ \beta_k(\gamma) = \mathbb{P}\left(D_k < \gamma | H_1\right), : P_k^{\mathrm{e}}(\gamma) = \alpha_k(\gamma)\pi_0 + \beta_k(\gamma)\pi_1, \tag{5.17}$$

respectively, the probability of false alarm, probability of miss, and Bayes (average) error probability. In this chapter, we adopt the *minimum Bayes error probability* criterion, both for the centralized and later for our distributed detector, and, from now on, we refer to it simply as the error probability. we relate the quantity $\beta_k(\gamma)$ with large deviation metric in Chapter 4. (Similar relations can be made for $\alpha_k(\gamma)$ and $P_k^{\mathrm{e}}(\gamma)$ as well.) To this end, note that $\mathbb{E}\left[D_k \mid H_1\right] = \gamma_1, \forall k$. Further, by the law of large numbers, D_k converges to γ_1 in probability. Hence, for $\gamma < \gamma_1$, the set $(-\infty, \gamma)$ is a deviation set, i.e., the "probability mass" vanishes from this set exponentially fast. Therefore, the analysis of $\beta_k(\gamma)$ reduces to the large deviation analysis for the set $E := (-\infty, \gamma)$.

We restrict the threshold γ to lie in (γ_0, γ_1). It can be shown that, for any $\gamma \in (\gamma_0, \gamma_1)$, $P_k^{\mathrm{e}}(\gamma)$ converges to zero exponentially fast as $k \to \infty$. On the other hand, for $\gamma \notin (\gamma_0, \gamma_1)$, $P_k^{\mathrm{e}}(\gamma)$ does not converge to zero at all. To see this, assume that H_1 is true, and let $\gamma \geq \gamma_1$. Then, by noting that $\mathbb{E}[D_k|H_1] = \gamma_1$, for all k, we have that $\beta(k, \gamma) = \mathbb{P}(D_k < \gamma|H_1) \geq \mathbb{P}(D_k \leq \gamma_1|H_1) \to \frac{1}{2}$ as $k \to \infty$, by the central limit theorem.

Denote by $I_l(\cdot)$, $l = 0, 1$, the Fenchel-Legendre transform of $\Lambda_l(\cdot)$:

$$I_l(z) = \sup_{\lambda \in \mathbb{R}} \lambda z - \Lambda_l(\lambda),\ z \in \mathbb{R}. \tag{5.18}$$

It can be shown [8] that $I_l(\cdot)$ is nonnegative, strictly convex (unless $L_1(1)$ is an almost sure constant), $I_l(\gamma_l) = 0$, for $l = 0, 1$, and $I_1(z) = I_0(z) - z$, [8]. We now state the result on the centralized detector's asymptotic performance.

Lemma 5.4 Let Assumptions 5.1-5.5 hold, and consider the family of centralized detectors (5.9) with the

constant threshold $\gamma \in (\gamma_0, \gamma_1)$. Then, the best (maximal) error exponent:

$$\lim_{k\to\infty} -\frac{1}{k} \log P_k^{\mathrm{e}}(\gamma)$$

is achieved for the zero threshold $\gamma = 0$ and equals NC_{ind}, where $C_{\mathrm{ind}} = I_0(0)$.

The quantity C_{ind} is referred to as the *Chernoff information* of a single agent observation $Y_{i,t}$. Lemma 5.4 says that the centralized detector' error exponent is N times larger than an individual agent's error exponent. We remark that, even if we allow for time-varying thresholds $\gamma_k = \gamma$, the error exponent NC_{ind} cannot be improved, i.e., the centralized detector with zero threshold is asymptotically optimal over all detectors. We will see that, when a certain condition on the network connectivity holds, the distributed detector is asymptotically optimal, i.e., achieves the best error exponent NC_{ind}, and the zero threshold is again optimal. However, when the network connectivity condition is not met, the distributed detector is no longer asymptotically optimal, and the optimal threshold may be non zero.

Proof [Proof of Lemma 5.4] Denote by $\Lambda_{0,N}$ the logarithmic moment generating function for the log-likelihood ratio $\sum_{i=1}^{N} L_{i,t}$ for the observations of all agents at time t. Then, $\Lambda_{0,N}(\lambda) = N\Lambda_0(\lambda)$, by the i.i.d. in space assumption on the agents' observations. The lemma now follows by the Chernoff lemma (Corollary 3.4.6, [8]):

$$\lim_{k\to\infty} -\frac{1}{k} \log P_k^{\mathrm{e}}(0) = \max_{\lambda\in[0,1]} \{-\Lambda_{0,N}(\lambda)\} = N \max_{\lambda\in[0,1]} \{-\Lambda_0(\lambda)\} = NI_0(0).$$

□

5.1.3 Distributed detection algorithm

We now consider distributed detection when the agents cooperate through a randomly varying network. Specifically, we consider the running consensus distributed detector proposed in [1]. The algorithm is of distributed inference type, similar to the algorithms that we considered in Chapter 4. Each agent i maintains its local decision variable $x_{i,k}$, which is a local estimate of the global optimal decision variable D_k in (5.9). Note that D_k is *not* locally available. At each time k, each agent i updates $x_{i,k}$ in two ways: 1) by incorporating its new observation $Y_{i,k}$ to make an intermediate decision variable $\frac{k-1}{k} x_i(k-1) + \frac{1}{k} L_i(k)$; and 2) by exchanging the intermediate decision variable locally with its neighbors and computing the weighted average of its own and the neighbors' intermediate variables.

More precisely, the update of $x_{i,k}$ is as follows:

$$x_{i,k} = \frac{k-1}{k} \sum_{j \in O_{i,k}} [W_k]_{ij} x_{j,k-1} + \frac{1}{k} L_{i,k}, \;\; k = 0, 1, 2, ... \tag{5.19}$$

$$x_{i,0} = 0. \tag{5.20}$$

Here $O_{i,k}$ is the (random) neighborhood of agent i at time k (including i), and $[W_k]_{ij}$ are the (random) averaging weights. The agent i's local decision test at time k is:

$$x_{i,k} \underset{H_0}{\overset{H_1}{\gtrless}} \gamma, \tag{5.21}$$

i.e., H_1 (respectively, H_0) is decided when $x_{i,k} \geq \gamma$ (respectively, $x_{i,k} < \gamma$).

Write algorithm (5.19) in vector form. Let $x_k = (x_{1,k}, x_{2,k}, ..., x_{N,k})^\top$ and $L_k = (L_{1,k}, ..., L_{N,k})^\top$. Also, collect the averaging weights $[W_k]_{ij}$ in the $N \times N$ matrix W_k, where, clearly, $[W_k]_{ij} = 0$ if the agents i and j do not communicate at time step k. The algorithm (5.19) becomes:

$$x_{k+1} = \frac{k}{k+1} W_{k+1} x_k + \frac{1}{k} L_k, \; k = 0, 1, 2, ... \;\; x_0 = 0. \tag{5.22}$$

Network model. We state the assumption on the random averaging matrices W_k and the observations Y_t.

Assumptions 5.5 The averaging matrices W_k satisfy the following:

(a) The sequence $\{W_k\}_{k=1}^{\infty}$ is i.i.d.

(b) W_k is symmetric, stochastic (row-sums equal 1 and $W_{ij}(k) \geq 0$), and has positive diagonal entries, with probability one, $\forall k$.

(c) W_k and Y_t are mutually independent over all k and t.

Note that we assume here the doubly stochasticity of matrices W_k's.

Define the matrices $\Phi(k, t)$ by:

$$\Phi(k, t) := W_k W_{k-1} ... W_{t+1}, \;\; k \geq t \geq 1. \tag{5.23}$$

It is easy to verify from (5.22) that x_k equals:

$$x_k = \frac{1}{k} \sum_{t=1}^{k} \Phi(k, t) L_t, \;\; k = 1, 2, ... \tag{5.24}$$

Choice of threshold γ. We restrict the choice of threshold γ to $\gamma \in (\gamma_0, \gamma_1)$, $\gamma_0 < 0$, $\gamma_1 > 0$, where we recall $\gamma_l = \mathbb{E}[L_1(1)|H_l]$, $l = 0, 1$. Namely, W_t is a stochastic matrix, hence $W_t 1 = 1$, for all t, and thus $\Phi(k,t)1 = 1$. Also, $\mathbb{E}[L_t|H_l] = \gamma_l 1$, for all t, $l = 0, 1$. Now, by iterating expectation:

$$\mathbb{E}[x_k|H_l] = \mathbb{E}[\mathbb{E}[x_k|H_l, W_1, ..., W_k]] = \mathbb{E}\left[\frac{1}{k}\sum_{t=1}^{k}\Phi(k,t)\mathbb{E}[L_t|H_l]\right] = \gamma_l 1,\ l = 0, 1,$$

and so $\mathbb{E}[x_{i,k}|H_l] = \gamma_l$, for all i, k. Moreover, it can be shown using the results from Chapter 4 that $x_{i,k}$ converges in probability to γ_l under H_l. Now, a similar argument as with the centralized detector in Subsection 5.1.2 shows that for $\gamma \notin (\gamma_0, \gamma_1)$, the error probability with detector (5.19) and (5.21) does not converge to zero when $k \to \infty$. We will show that, for any $\gamma \in (\gamma_0, \gamma_1)$, the error probability converges to 0 exponentially fast, and we find the optimal $\gamma = \gamma^\star$ that maximizes a certain lower bound on the exponent of the error probability.

Network connectivity. From (5.24), we can see that the matrices $\Phi(k,t)$ should be as close to $J = (1/N)11^\top$ as possible for enhanced detection performance. To measure the "quality" of matrices $\Phi(k,t)$, we invoke the quantity $\mathcal{J}$ from Chapter 2. We refer to $\mathcal{J}$ as the connectivity, and we recall the definition:[4]

$$\mathcal{J} := \lim_{(k-t)\to\infty} -\frac{1}{k-t}\log\mathbb{P}\left(\|\Phi(k,t) - J\| > \epsilon\right). \tag{5.25}$$

The following lemma easily follows from (5.25).

Lemma 5.6 Let Assumption 5.5 hold. Then, for any $\delta \in (0, \mathcal{J})$, there exists a constant $C(\delta) \in (0, \infty)$ (independent of $\epsilon \in (0,1)$) such that, for any $\epsilon \in (0,1)$:

$$\mathbb{P}\left(\|\Phi(k,t) - J\| > \epsilon\right) \leq C(\delta)e^{-(k-t)(\mathcal{J}-\delta)},\ \text{for all } k \geq t.$$

5.2 Main results: Error exponents for distributed detection

Subsection 5.2.1 establishes the asymptotic performance of distributed detection under identically distributed agents' observations; subsection 5.2.2 proves the results.

[4]The limit in (5.25) exists and it does not depend on ϵ for $\epsilon \in (0,1)$.

5.2.1 Statement of main results

In this section, we analyze the performance of distributed detection in terms of the detection error exponent, when the number of observations (per agent), or the size k of the observation interval tends to $+\infty$. As we will see next, we show that there exists a threshold on the network connectivity $\mathcal{J}$ such that if $\mathcal{J}$ is above this threshold, each agent in the network achieves asymptotic optimality (i.e., the error exponent at each agent is the total Chernoff information equal to NC_{ind}). When $\mathcal{J}$ is below the threshold, we give a lower bound for the error exponent. Both the threshold and the lower bound are given *solely* in terms of the log-moment generating function Λ_0 and the number of agents N. These findings are summarized in Theorem 5.7 and Corollary 5.8 below.

Let $\alpha_{i,k}(\gamma)$, $\beta_{i,k}(\gamma)$, and $P^{\text{e}}_{i,k}(\gamma)$ denote the probability of false alarm, the probability of miss, and the error probability, respectively, of agent i for the detector (5.19) and (5.21), for the threshold equal to γ:

$$\alpha_{i,k}(\gamma) = \mathbb{P}\left(x_{i,k} \geq \gamma | H_0\right),\ \beta_{i,k}(\gamma) = \mathbb{P}\left(x_{i,k} < \gamma | H_1\right),\ P^{\text{e}}_{i,k}(\gamma) = \pi_0 \alpha_{i,k}(\gamma) + \pi_1 \beta_{i,k}(\gamma), \tag{5.26}$$

where, we recall, π_1 and π_0 are the prior probabilities. Also, recall $I_l(\cdot)$, $l = 0, 1$, in (5.18). Finally, for two functions $\phi_1, \phi_2 : \mathbb{R} \mapsto \mathbb{R}$, recall from Chapter 4 the definition of their convex hull $\phi_3 = \overline{\text{co}}(\phi_1, \phi_2) : \mathbb{R} \mapsto \mathbb{R}$, $\phi_3(x) = \overline{\text{co}}(\phi_1, \phi_2)(x)$.

Theorem 5.7 Let Assumptions 5.1-5.5 hold and consider the family of distributed detectors in (5.19) and (5.21) parameterized by detection thresholds $\gamma \in (\gamma_0, \gamma_1)$. Let λ^{s}_l be the zero of the function:

$$\Delta_l(\lambda) := \Lambda_l(N\lambda) - \mathcal{J} - N\Lambda_l(\lambda),\ \ l = 0, 1, \tag{5.27}$$

and define γ^-_l, γ^+_l, $l = 0, 1$ by

$$\gamma^-_0 = \Lambda'_0(\lambda^{\text{s}}_0),\ \gamma^+_0 = \Lambda'_0(N\lambda^{\text{s}}_0) \geq \gamma^-_0 \tag{5.28}$$

$$\gamma^-_1 = \Lambda'_1(N\lambda^{\text{s}}_1),\ \gamma^+_1 = \Lambda'_1(\lambda^{\text{s}}_1) \geq \gamma^-_1. \tag{5.29}$$

Then, for every $\gamma \in (\gamma_0, \gamma_1)$, at each agent i, $i = 1, \ldots, N$, we have:

$$\liminf_{k\to\infty} -\frac{1}{k}\log \alpha_{i,k}(\gamma) \geq I^0_{\mathcal{J},N}(\gamma), \quad \liminf_{k\to\infty} -\frac{1}{k}\log \beta_{i,k}(\gamma) \geq I^1_{\mathcal{J},N}(\gamma), \tag{5.30}$$

where

$$I^0_{\mathcal{J},N}(\gamma) = \overline{\text{co}}\left(I_0 + \mathcal{J},\, NI_0\right)(\gamma) = \begin{cases} NI_0(\gamma), & \gamma \in (\gamma_0, \gamma_0^-] \\ NI_0(\gamma_0^-) + N\lambda_0^{\text{s}}(\gamma - \gamma_0^-), & \gamma \in (\gamma_0^-, \gamma_0^+) \\ I_0(\gamma) + \mathcal{J}, & \gamma \in [\gamma_0^+, \gamma_1) \end{cases}$$

$$I^1_{\mathcal{J},N}(\gamma) = \overline{\text{co}}\left(I_1 + \mathcal{J},\, NI_1\right)(\gamma) = \begin{cases} I_1(\gamma) + \mathcal{J}, & \gamma \in (\gamma_0, \gamma_1^-] \\ NI_1(\gamma_1^+) + N\lambda_1^{\text{s}}(\gamma - \gamma_1^+), & \gamma \in (\gamma_1^-, \gamma_1^+) \\ NI_1(\gamma), & \gamma \in [\gamma_1^+, \gamma_1). \end{cases}$$

Corollary 5.8 Let Assumptions 5.1-5.5 hold and consider the family of distributed detectors in (5.19) and (5.21) parameterized by detector thresholds $\gamma \in (\gamma_0, \gamma_1)$. Then:

(a)

$$\liminf_{k\to\infty} -\frac{1}{k}\log P^{\text{e}}_{i,k}(\gamma) \geq \min\{I^0_{\mathcal{J},N}(\gamma), I^1_{\mathcal{J},N}(\gamma)\} > 0, \tag{5.31}$$

and the lower bound in (5.31) is maximized for the point $\gamma^\star \in (\gamma_0, \gamma_1)$[5] at which $I^0_{\mathcal{J},N}(\gamma^\star) = I^1_{\mathcal{J},N}(\gamma^\star)$.

(b) Consider $\lambda^\bullet = \arg\min_{\lambda\in\mathbb{R}} \Lambda_0(\lambda)$, and let:

$$\mathcal{J}^\star\left(\Lambda_0, N\right) = \max\{\Lambda_0(N\lambda^\bullet) - N\Lambda_0(\lambda^\bullet), \Lambda_0(1 - N(1-\lambda^\bullet)) - N\Lambda_0(\lambda^\bullet)\}. \tag{5.32}$$

Then, when $\mathcal{J} \geq \mathcal{J}^\star(\Lambda_0, N)$, each agent i with the detector threshold set to $\gamma = 0$, is asymptotically optimal:

$$\lim_{k\to\infty} -\frac{1}{k}\log P^{\text{e}}_{i,k}(0) = NC_{\text{ind}}.$$

(c) Suppose $\Lambda_0(\lambda) = \Lambda_0(1-\lambda)$, for $\lambda \in [0,1]$. Then, $\gamma^\star = 0$, irrespective of the value of r (even when $\mathcal{J} < \mathcal{J}^\star(\Lambda_0, N)$.)

Figure 5.1 (left) illustrates the error exponent lower bounds $I^0_{\mathcal{J},N}(\gamma)$ and $I^1_{\mathcal{J},N}(\gamma)$ in Theorem 5.7, while Figure 5.1 (right) illustrates the quantities in (5.28). (See the definition of the function $\Phi_0(\lambda)$ in (5.44) in the proof of Theorem 5.7.) Note that $B_l(\cdot)$ is the convex envelope of the functions $NI_l(\cdot)$ and $I_l(\cdot) + \mathcal{J}$, $l = 0, 1$. We consider $N = 3$ agents and a discrete distribution of $Y_{i,t}$ over a 5-point alphabet, with the distribution $[.2, .2, .2, .2, .2]$ under H_1, and $[0.01, 0.01, 0.01, 0.01, 0.96]$ under H_0. We set here $\mathcal{J} \approx 0.92$

Remark. Consider part (c) of Corollary 5.8. When $\Lambda_0(\lambda) = \Lambda_0(1-\lambda)$, for $\lambda \in [0,1]$, it can be shown that

[5] As we show in the proof, such a point exists and is unique.

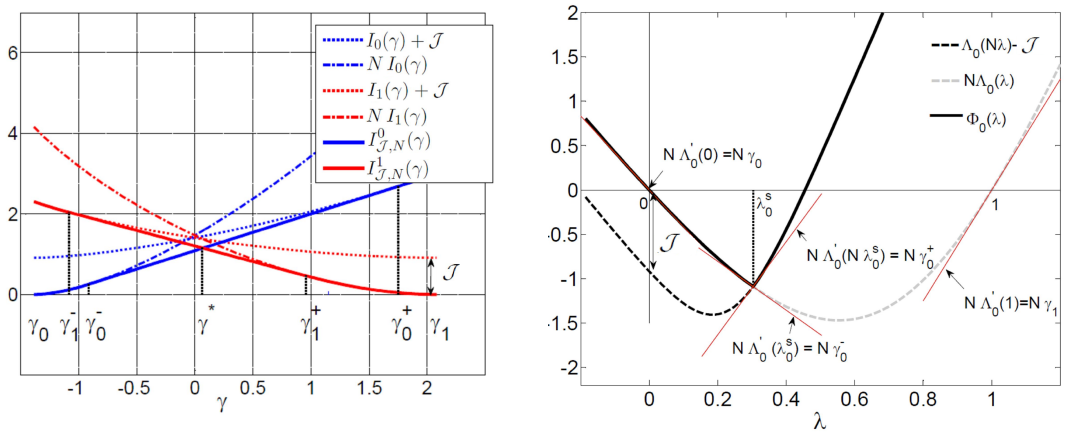

Figure 5.1: **Left:** Illustration of the error exponent lower bounds $I^0_{\mathcal{J},N}(\gamma)$ and $I^1_{\mathcal{J},N}(\gamma)$ in Theorem 5.7; **Right:** Illustration of the function $\Phi_0(\lambda)$ in (5.44), and the quantities in (5.28). We consider $N = 3$ agents and a discrete distribution of $Y_{i,t}$ over a 5-point alphabet, with the distribution $[.2, .2, .2, .2, .2]$ under H_1, and $[0.01, 0.01, 0.01, 0.01, 0.96]$ under H_0. We set here $\mathcal{J} \approx 0.92$.

$\gamma_0 = -\gamma_1 < 0$, and $I^0_{\mathcal{J},N}(\gamma) = I^1_{\mathcal{J},N}(-\gamma)$, for all $\gamma \in (\gamma_0, \gamma_1)$. This implies that the point $\gamma^\star$ at which $I^0_{\mathcal{J},N}$ and $I^1_{\mathcal{J},N}$ are equal is necessarily zero, and hence the optimal detector threshold $\gamma^\star = 0$, irrespective of the network connectivity $\mathcal{J}$ (even when $\mathcal{J} < \mathcal{J}^\star(\Lambda_0, N)$.) This symmetry holds, e.g., for the Gaussian and Laplace distribution considered in Section 5.3.

Corollary 5.8 states that, when the network connectivity $\mathcal{J}$ is above a threshold, the distributed detector in (5.19) and (5.21) is asymptotically equivalent to the optimal centralized detector. The corresponding optimal detector threshold is $\gamma = 0$. When $\mathcal{J}$ is below the threshold, Corollary 5.8 determines what value of the error exponent the distributed detector can achieve, for any given $\gamma \in (\gamma_0, \gamma_1)$. Moreover, Corollary 5.8 finds the optimal detector threshold $\gamma^\star$ for a given r; $\gamma^\star$ can be found as the unique zero of the strictly decreasing function $\Delta_B(\gamma) := I^1_{\mathcal{J},N}(\gamma) - I^0_{\mathcal{J},N}(\gamma)$ on $\gamma \in (\gamma_0, \gamma_1)$, see the proof of Corollary 5.8, e.g., by bisection on (γ_0, γ_1).

Corollary 5.8 establishes that there exists a "sufficient" connectivity, say $\mathcal{J}^\star$, so that further improvement on the connectivity (and further spending of resources, e.g., transmission power) does not lead to a pay off in terms of asymptotic detection performance. Hence, Corollary 5.8 is valuable in the practical design of a network, as it says how much connectivity (resources) is sufficient to achieve asymptotically optimal detection.

Equation (5.31) says that the distribution of the agent observations (through the logarithmic moment generating function) plays a role in determining the performance of distributed detection. We illustrate and explain by examples this effect in Section 5.3.

5.2.2 Proofs of the main results

In this subsection, we prove Theorem 5.7 and Corollary 5.8. We follow essentially the same arguments as in the proof of Theorem 4.19 in Chapter 4. Differently from the latter proof, 1) we explicitly characterize the convex hull of the functions $NI_l(\cdot)$ and $I_l(\cdot) + \mathcal{J}$, $l = 0, 1$; and 2) we analyze the (average) error probability and the choice of thresholds γ. (The average error probability and thresholds are specific to distributed detection and are not considered in Chapter 4.) For convenience, we include here the full proof of Theorem 5.7 and Corollary 5.8.

Proof [Proof of Theorem 5.7] Consider the probability of false alarm $\alpha_{i,k}(\gamma)$ in (5.26). We upper bound $\alpha_{i,k}(\gamma)$ using the exponential Markov inequality [69] parameterized by $\zeta \geq 0$:

$$\alpha_{i,k}(\gamma) = \mathbb{P}\left(x_{i,k} \geq \gamma \,|\, H_0\right) = \mathbb{P}\left(e^{\zeta x_{i,k}} \geq e^{\zeta\gamma} \,|\, H_0\right) \leq \mathbb{E}\left[e^{\zeta x_{i,k}} | H_0\right] e^{-\zeta\gamma}. \tag{5.33}$$

Next, by setting $\zeta = N\,k\,\lambda$, with $\lambda \geq 0$, we obtain:

$$\begin{aligned} \alpha_{i,k}(\gamma) &\leq \mathbb{E}\left[e^{Nk\lambda x_{i,k}} | H_0\right] e^{-Nk\lambda\gamma} && (5.34)\\ &= \mathbb{E}\left[e^{N\lambda\sum_{t=1}^{k}\sum_{j=1}^{N}[\Phi(k,t)]_{i,j}L_{j,t}} | H_0\right] e^{-Nk\lambda\gamma}. && (5.35) \end{aligned}$$

The terms in the sum in the exponent in (5.35) are conditionally independent, given the realizations of the averaging matrices W_t, $t = 1, \ldots, k$, Thus, by iterating the expectations, and using the definition of Λ_0 in (5.10), we compute the expectation in (5.35) by conditioning first on W_t, $t = 1, \ldots, k$:

$$\begin{aligned} \mathbb{E}\left[e^{N\lambda\sum_{t=1}^{k}\sum_{j=1}^{N}[\Phi(k,t)]_{i,j}L_{j,t}} | H_0\right] &= \mathbb{E}\left[\mathbb{E}\left[e^{N\lambda\sum_{t=1}^{k}\sum_{j=1}^{N}[\Phi(k,t)]_{i,j}L_{j,t}} | H_0, W_1, \ldots, W_k\right]\right] \\ &= \mathbb{E}\left[e^{\sum_{t=1}^{k}\sum_{j=1}^{N}\Lambda_0(N\lambda[\Phi(k,t)]_{i,j})}\right]. && (5.36) \end{aligned}$$

Partition of the sample space. We handle the random matrix realizations W_t, $t = 1, \ldots, k$, through a suitable partition of the underlying probability space. Adapting the argument from Subsection 4.5.3 in Chapter 4, partition the probability space *based on the time of the last successful averaging*. In more detail, for a fixed k, introduce the partition $\mathcal{P}_k$ of the sample space that consists of the disjoint events $\mathcal{A}_{s,k}$, $s =$

$0, 1, ..., k$, given by:

$$\mathcal{A}_{s,k} = \{\|\Phi(k,s) - J\| \leq \epsilon \text{ and } \|\Phi(k,s+1) - J\| > \epsilon\},$$

for $s = 1, ..., k-1$, $\mathcal{A}_{0,k} = \{\|\Phi(k,1) - J\| > \epsilon\}$, and $A_{k,k} = \{\|\Phi(k,k) - J\| \leq \epsilon\}$. For simplicity of notation, we drop the index k in the sequel and denote event $\mathcal{A}_{s,k}$ by $\mathcal{A}_s$, $s = 0, \ldots, k$. for $\epsilon > 0$. Intuitively, the smaller t is, the closer the product $\Phi(k,t)$ to J is; if the event $\mathcal{A}_s$ occurred, then the largest t for which the product $\Phi(k,t)$ is still ϵ-close to J equals s. We now show that $\mathcal{P}_k$ is indeed a partition. We need the following simple lemma. The lemma shows that convergence of $\Phi(k,s) - J$ is monotonic, for any realization of the matrices $W_1, W_2, ..., W_k$.

Lemma 5.9 Let Assumption 5.5 hold. Then, for any realization of the matrices $W_1, ..., W_k$:

$$\|\Phi(k,s) - J\| \leq \|\Phi(k,t) - J\|, \text{ for } 1 \leq s \leq t \leq k.$$

Proof Since every realization of W_t is stochastic and symmetric for every t, we have that $W_t 1 = 1$ and $1^\top W_t = 1^\top$, and, so: $\Phi(k,s) - J = W_k \cdots W(s) - J = (W_k - J) \cdots (W(s) - J)$. Now, using the sub-multiplicative property of the spectral norm, we get

$$\begin{aligned}\|\Phi(k,s) - J\| &= \|(W_k - J) \cdots (W_t - J)(W(t-1) - J) \cdots (W(s) - J)\| \\ &\leq \|(W_k - J) \cdots (W_t - J)\| \|(W(t-1) - J)\| \cdots \|(W(s) - J)\|.\end{aligned}$$

To prove Lemma 5.9, it remains to show that $\|W_t - J\| \leq 1$, for any realization of W_t. To this end, fix a realization W of W_t. Consider the eigenvalue decomposition $W = QMQ^\top$, where $M = \text{diag}(\mu_1, \ldots, \mu_N)$ is the matrix of eigenvalues of W, and the columns of Q are the orthonormal eigenvectors. As $\frac{1}{\sqrt{N}}1$ is the eigenvector associated with eigenvalue $\mu_1 = 1$, we have that $W - J = QM'Q^\top$, where $M = \text{diag}(0, \mu_2, \ldots, \mu_N)$. Because W is stochastic, we know that $1 = \mu_1 \geq \mu_2 \geq ... \geq \mu_N \geq -1$, and so $\|W - J\| = \max\{|\mu_2|, |\mu_N|\} \leq 1$. □

We now show that $\mathcal{P}_k$ is indeed a partition. Note first that (at least) one of the events $\mathcal{A}_0, ..., \mathcal{A}_k$ necessarily occurs. It remains to show that the events $\mathcal{A}_s$ are disjoint. We carry out this by fixing arbitrary $s = 1, ..., k$, and showing that, if the event $\mathcal{A}_s$ occurs, then $\mathcal{A}_t$, $t \neq s$, does not occur. Suppose that $\mathcal{A}_s$ occurs, i.e., the realizations $W_1, ..., W_k$ are such that $\|\Phi(k,s) - J\| \leq \epsilon$ and $\|\Phi(k,s+1) - J\| > \epsilon$. Fix any $t > s$. Then, event $\mathcal{A}_t$ does not occur, because, by Lemma 5.9, $\|\Phi(k,t) - J\| \geq \|\Phi(k,s+1) - J\| > \epsilon$. Now,

fix any $t < s$. Then, event $\mathcal{A}_t$ does not occur, because, by Lemma 5.9, $\|\Phi(k,t+1)-J\| \leq \|\Phi(k,s)-J\| \leq \epsilon$. Thus, for any $s = 1, ..., k$, if the event $\mathcal{A}_s$ occurs, then $\mathcal{A}_t$, for $t \neq s$, does not occur, and hence the events $\mathcal{A}_s$ are disjoint.

Using the total probability law over $\mathcal{P}_k$, the expectation (5.36) is computed by:

$$\mathbb{E}\left[e^{\sum_{t=1}^{k}\sum_{j=1}^{N}\Lambda_0(N\lambda[\Phi(k,t)]_{i,j})}\right] = \sum_{s=0}^{k}\mathbb{E}\left[e^{\sum_{t=1}^{k}\sum_{j=1}^{N}\Lambda_0(N\lambda[\Phi(k,t)]_{i,j})}\mathcal{I}_{\mathcal{A}_s}\right], \tag{5.37}$$

where, we recall, $\mathcal{I}_{\mathcal{A}_s}$ is the indicator function of the event $\mathcal{A}_s$. The following lemma explains how to use the partition $\mathcal{P}_k$ to upper bound the expectation in (5.37).

Lemma 5.10 Let Assumptions 5.1-5.5 hold. Then:

(a) For any realization of the random matrices W_t, $t = 1, 2, ..., k$:

$$\sum_{j=1}^{N}\Lambda_0\left(N\lambda[\Phi(k,t)]_{i,j}\right) \leq \Lambda_0\left(N\lambda\right), \ \forall t = 1, \ldots, k.$$

(b) Further, consider a fixed s in $\{0, 1, ..., k\}$. If the event $\mathcal{A}_s$ occurred, then, for $i = 1, \ldots, N$: $\Lambda_0\left(N\lambda[\Phi(k,t)]_{i,j}\right) \leq \max\left(\Lambda_0\left(\lambda - \epsilon N\sqrt{N}\lambda\right), \Lambda_0\left(\lambda + \epsilon N\sqrt{N}\lambda\right)\right)$, $\forall t = 1, \ldots, s$, $\forall j = 1, \ldots, N$.

Proof To prove part (a) of the lemma, by convexity of Λ_0, the maximum of $\sum_{j=1}^{N}\Lambda_0(N\lambda a_j)$ over the simplex $\left\{a \in \mathbb{R}^N : \sum_{j=1}^{N} a_j = 1, \ a_j \geq 0, \ j = 1, \ldots, N\right\}$ is achieved at a corner point of the simplex. The maximum equals: $\Lambda_0(N\lambda) + (N-1)\Lambda_0(0) = \Lambda_0(N\lambda)$, where we use the property from Lemma 5.2, part (b), that $\Lambda_0(0) = 0$. Finally, since for any realization of the matrices $W_1, \ldots, W_k$, the set of entries $\{[\Phi(k,t)]_{i,j} : j = 1, \ldots, N\}$ is a point in the simplex, the claim of part (a) of the lemma follows.

To prove part (b) of the lemma, suppose that event $\mathcal{A}_s$ occurred. Then, by the definition of $\mathcal{A}_s$,

$$\|\Phi(k,s) - J\| = \|W_k \cdot \ldots \cdot W(s) - J\| \leq \epsilon.$$

Now, by Lemma 5.9:

$$\|\Phi(k,t) - J\| = \|W_k \cdot \ldots \cdot W_t - J\| \leq \epsilon,$$

for every $t \leq s$. Then, by the equivalence of the 1-norm and the spectral norm, it follows that:

$$\left|[\Phi(k,t)]_{i,j} - \frac{1}{N}\right| \leq \sqrt{N}\epsilon, \text{ for } t = 1, \ldots, s, \text{ for all } i, j = 1, \ldots, N.$$

Finally, since Λ_0 is convex (Lemma 5.2, part (a)), its maximum in $\left[\lambda - \epsilon N\sqrt{N}\lambda, \lambda + \epsilon N\sqrt{N}\lambda\right]$ is attained at a boundary point and the claim follows. □

We now fix $\delta \in (0, \mathcal{J})$. Using the results from Lemma 5.6 and Lemma 5.10, we next bound the expectation in (5.37) as follows:

$$\begin{aligned}\sum_{s=0}^{k} \mathbb{E}\left[e^{\sum_{t=1}^{k}\sum_{j=1}^{N}\Lambda_0(N\lambda[\Phi(k,t)]_{i,j})}\mathcal{I}_{\mathcal{A}_s}\right] &\leq \sum_{s=0}^{k}\left(e^{sN\max\left(\Lambda_0\left(\lambda-\epsilon N\sqrt{N}\lambda\right),\Lambda_0\left(\lambda+\epsilon N\sqrt{N}\lambda\right)\right)+(k-s)\Lambda_0(N\lambda)}\right) \\ &\times \left(C(\delta)e^{-(k-(s+1))(\mathcal{J}-\delta)}\right). \end{aligned} \tag{5.38}$$

To simplify the notation, we introduce the function:

$$g_0 : \mathbb{R}^2 \longrightarrow \mathbb{R},\ g_0(\epsilon,\lambda) := \max\left(\Lambda_0\left(\lambda - \epsilon N\sqrt{N}\lambda\right), \Lambda_0\left(\lambda + \epsilon N\sqrt{N}\lambda\right)\right). \tag{5.39}$$

We need the following property of $g_0(\cdot,\cdot)$.

Lemma 5.11 Consider $g_0(\cdot,\cdot)$ in (5.39). Then, for every $\lambda \in \mathbb{R}$, the following holds:

$$\inf_{\epsilon>0} g_0(\epsilon,\lambda) = \Lambda_0(\lambda).$$

Proof Since $\Lambda_0(\cdot)$ is convex, for $\epsilon' < \epsilon$ and for a fixed λ, we have that

$$g_0(\epsilon,\lambda) = \max_{\delta\in[-\epsilon,\epsilon]}\Lambda_0\left(\lambda+\delta N\sqrt{N}\lambda\right) \geq \max_{\delta\in[-\epsilon',\epsilon']}\Lambda_0\left(\lambda+\delta N\sqrt{N}\lambda\right) = g_0(\epsilon',\lambda).$$

Thus, for a fixed λ, $f(\cdot,\lambda)$ is non-increasing, and the claim of the lemma follows. □

We proceed by bounding further the right hand side in (5.38), by rewriting $e^{-(k-(s+1))(\mathcal{J}-\delta)}$ as $\frac{1}{re^{\delta}}\, e^{-(k-s)(\mathcal{J}-\delta)}$:

$$\begin{aligned} & \sum_{s=0}^{k}\frac{C(\delta)}{re^{\delta}}\, e^{sNg_0(\epsilon,\lambda)+(k-s)\Lambda_0(N\lambda)-(k-s)(\mathcal{J}-\delta)} \\ \leq\ & (k+1)\max_{s\in\{0,\ldots,k\}}\frac{C(\delta)}{re^{\delta}}e^{[\,sNg_0(\epsilon,\lambda)+(k-s)\,(\Lambda_0(N\lambda)-(\mathcal{J}-\delta))\,]} \\ =\ & (k+1)\,\frac{C(\delta)}{re^{\delta}}\, e^{\max_{s\in\{0,\ldots,k\}}[\,sNg_0(\epsilon,\lambda)+(k-s)(\Lambda_0(N\lambda)-(\mathcal{J}-\delta))\,]} \\ \leq\ & (k+1)\,\frac{C(\delta)}{re^{\delta}}\, e^{k\max_{\theta\in[0,1]}[\,\theta Ng_0(\epsilon,\lambda)+(1-\theta)(\Lambda_0(N\lambda)-(\mathcal{J}-\delta))]} \\ =\ & (k+1)\,\frac{C(\delta)}{re^{\delta}}\, e^{k\,\max\{\,Ng_0(\epsilon,\lambda),\,\Lambda_0(N\lambda)-(\mathcal{J}-\delta)\,\}}. \end{aligned} \tag{5.40}$$

The second inequality follows by introducing $\theta := \frac{s}{k}$ and by enlarging the set for θ from $\left\{0, \frac{1}{k}, \ldots, 1\right\}$ to the continuous interval $[0, 1]$. Taking the $\log$ and dividing by k, from (5.34) and (5.40) we get:

$$\frac{1}{k} \log \alpha_{i,k}(\gamma) \quad \leq \quad \frac{\log(k+1)}{k} + \frac{\log \frac{C(\delta)}{re^{\delta}}}{k} + \max\left\{N g_0(\epsilon, \lambda), \Lambda_0(N\lambda) - (\mathcal{J} - \delta)\right\} - N\gamma\lambda. \quad (5.41)$$

Taking the $\limsup$ when $k \to \infty$, the first two terms in the right hand side of (5.41) vanish; further, changing the sign, we get a bound on the exponent of $\alpha_i(k)$ that holds for every $\epsilon \in (0, 1)$:

$$\liminf -\frac{1}{k} \log \alpha_{i,k}(\gamma) \quad \geq \quad -\max\left\{N g_0(\epsilon, \lambda),\ \Lambda_0(N\lambda) - (\mathcal{J} - \delta)\right\} + N\gamma\lambda.$$

By Lemma 5.11, as $\epsilon \to 0$, $N g_0(\epsilon, \lambda)$ decreases to $N\,\Lambda_0(\lambda)$; further, letting $\delta \to 0$, we get

$$\liminf -\frac{1}{k} \log \alpha_{i,k}(\gamma) \quad \geq \quad -\max\left\{N\Lambda_0(\lambda),\ \Lambda_0(N\lambda) - \mathcal{J}\right\} + N\gamma\lambda. \quad (5.42)$$

The previous bound on the exponent of the probability of false alarm holds for any $\lambda \geq 0$. To get the best bound, we maximize the expression on the right hand side of (5.42) over $\lambda \in [0, \infty)$. (We refer to Figure 5.1, left and right, to help us illustrate the bounds $I^0_{\mathcal{J},N}(\gamma)$ and $I^1_{\mathcal{J},N}(\gamma)$ for the discrete valued observations $Y_{i,t}$ over a 5-point alphabet.) More precisely, we calculate the convex hull $I^0_{\mathcal{J},N}(\gamma)$ from Theorem 5.7:

$$I^0_{\mathcal{J},N}(\gamma) = \max_{\lambda \geq 0} N\gamma\lambda - \Phi_0(\lambda), \quad (5.43)$$

where

$$\Phi_0(\lambda) := \max\left\{N\Lambda_0(\lambda),\ \Lambda_0(N\lambda) - \mathcal{J}\right\}. \quad (5.44)$$

To calculate $I^0_{\mathcal{J},N}(\gamma)$ in (5.43), we need to find an optimizer $\lambda^\star = \lambda^\star(\gamma)$ (if it exists) of the objective in (5.43); from the first order optimality conditions, $\lambda^\star$ is a point that satisfies:

$$N\gamma \in \partial\Phi_0(\lambda^\star), \quad \lambda^\star \geq 0, \quad (5.45)$$

where $\partial\Phi_0(\lambda)$ denotes the subdifferential set of Φ_0 at λ. We next characterize $\partial\Phi_0(\lambda)$, for $\lambda \geq 0$. Recall the zero λ^{s}_0 of $\Delta_0(\cdot)$ from Theorem 5.7. The function $\Phi_0(\lambda)$ in (5.44) equals: 1) $N\Lambda_0(\lambda)$ on $\lambda \in [0, \lambda^{\mathrm{s}}_0)$; 2) $N\Lambda_0(\lambda^{\mathrm{s}}_0) = \Lambda_0(N\lambda^{\mathrm{s}}_0) - \mathcal{J}$ at $\lambda = \lambda^{\mathrm{s}}_0$; and 3) $\Lambda_0(N\lambda) - \mathcal{J}$ on $\lambda > \lambda^{\mathrm{s}}_0$. Thus, by the rule for the

subdifferential of a pointwise maximum of two convex functions, the subdifferential $\partial\Phi_0(\lambda)$ is:

$$\partial\Phi_0(\lambda) = \begin{cases} \{N\Lambda_0'(\lambda)\}, & \text{for} \quad \lambda \in [0, \lambda_0^{\mathrm{s}}) \\ [N\Lambda_0'(\lambda), N\Lambda_0'(N\lambda)], & \text{for} \quad \lambda = \lambda_0^{\mathrm{s}} \\ \{N\Lambda_0'(N\lambda)\}, & \text{for} \quad \lambda > \lambda_0^{\mathrm{s}}. \end{cases} \tag{5.46}$$

Geometrically, $\partial\Phi_0(\lambda)$ is the set of slopes of the tangent lines to the graph of $\Phi_0(\cdot)$ at the point λ, see Figure 5.1, right. We next find $I^0_{\mathcal{J},N}(\gamma)$ for any $\gamma \in (\gamma_0, \gamma_1)$, by finding $\lambda^\star = \lambda^\star(\gamma)$ for any $\gamma \in (\gamma_0, \gamma_1)$. Geometrically, from Figure 5.1, right, given a slope $\gamma \in (\gamma_0, \gamma_1)$, finding a $\lambda^\star$ corresponds to finding a point at which a tangent line to the graph of $\Phi_0(\cdot)$ has a slope γ. Recall γ_0^- and γ_0^+ from Theorem 5.7. We consider separately three regions: 1) $\gamma \in [\gamma_0, \gamma_0^-]$; 2) $\gamma \in (\gamma_0^-, \gamma_0^+)$; and 3) $\gamma \in [\gamma_0^+, \gamma_1]$. For the first region, (5.45) reduces to finding $\lambda^\star \geq 0$ such that $\gamma = \Lambda_0'(\lambda^\star)$. Recall that $\Lambda_0'(0) = \gamma_0$, i.e., for $\gamma = \gamma_0$, equation (5.45) holds (only) for $\lambda^\star = 0$. Also, for $\gamma = \gamma_0^-$, we have, *by definition of* γ_0^-, that $\Lambda_0'(\lambda_0^{\mathrm{s}}) = \gamma_0^-$, i.e., equation (5.45) holds (only) for $\lambda^\star = \lambda_0^{\mathrm{s}}$. Because $\Lambda_0'(\lambda)$ is continuous and strictly increasing on $\lambda \in [0, \lambda_0^{\mathrm{s}}]$, it follows that, for any $\gamma \in [\gamma_0, \gamma_0^-]$ there exists a solution $\lambda^\star$ to (5.45), it is unique, and lies in $[0, \lambda_0^{\mathrm{s}}]$. Now, we calculate $I^0_{\mathcal{J},N}(\gamma)$:

$$I^0_{\mathcal{J},N}(\gamma) = N\lambda^\star\gamma - \Phi_0(\lambda^\star) = N\lambda^\star\gamma - N\Lambda_0(\lambda^\star) \tag{5.47}$$

$$= N(\lambda^\star\gamma - \Lambda_0(\lambda^\star)) = N\sup_{\lambda\geq 0}(\lambda\gamma - \Lambda_0(\lambda)) = NI_0(\gamma), \tag{5.48}$$

where we used the fact that $\Phi_0(\lambda^\star) = N\Lambda_0(\lambda^\star)$ (because $\lambda^\star \leq \lambda_0^{\mathrm{s}}$), and the definition of the function $I_0(\cdot)$ in (5.18). We now consider the second region. Fix $\gamma \in (\gamma_0^-, \gamma_0^+)$. It is trivial to verify, from (5.46), that $\lambda^\star = \lambda_0^{\mathrm{s}}$ is the solution to (5.45). Thus, we calculate $I^0_{\mathcal{J},N}(\gamma)$ as follows:

$$I^0_{\mathcal{J},N}(\gamma) = N\lambda_0^{\mathrm{s}}\gamma - \Phi_0(\lambda_0^{\mathrm{s}}) = N\lambda_0^{\mathrm{s}}\gamma - N\Lambda_0(\lambda_0^{\mathrm{s}}) \tag{5.49}$$

$$= N\lambda_0^{\mathrm{s}}(\gamma - \gamma_0^-) + N\lambda_0^{\mathrm{s}}\gamma_0^- - N\Lambda_0(\lambda_0^{\mathrm{s}}) = N\lambda_0^{\mathrm{s}}(\gamma - \gamma_0^-) + NI_0(\gamma_0^-), \tag{5.50}$$

where we used the fact that $\lambda_0^{\mathrm{s}}\gamma_0^- - \Lambda_0(\lambda_0^{\mathrm{s}}) = \sup_{\lambda\geq 0}\lambda\gamma_0^- - \Lambda_0(\lambda) = I_0(\gamma_0^-)$. The proof for the third region is analogous to the proof for the first region. Hence, the part of the proof for $\alpha_{i,k}(\gamma)$ is complete.

For a proof of the claim on the probability of miss $\beta_{i,k}(\gamma) = \mathbb{P}\left(x_{i,k} < \gamma | H_1\right)$, we proceed analogously to (5.33), where instead of $\zeta \geq 0$, we now use $\zeta \leq 0$ (and, hence, the proof proceeds with $\lambda \leq 0$). □

Proof [Proof of Corollary 5.8] We first prove part (a). Consider the error probability $P^{\mathrm{e}}_{i,k}(\gamma)$ in (5.26).

By Lemma 1.2.15 in [8], we have that:

$$\begin{aligned}\liminf_{k\to\infty} -\frac{1}{k}\log P^{\mathrm{e}}_{i,k}(\gamma) &= \min\left\{\liminf_{k\to\infty} -\frac{1}{k}\log(\alpha_{i,k}(\gamma)\pi_0), \liminf_{k\to\infty} -\frac{1}{k}\log(\beta_{i,k}(\gamma)\pi_1)\right\}\\ &= \min\left\{\liminf_{k\to\infty} -\frac{1}{k}\log\alpha_{i,k}(\gamma), \liminf_{k\to\infty} -\frac{1}{k}\log\beta_{i,k}(\gamma)\right\}\\ &\geq \min\{I^0_{\mathcal{J},N}(\gamma), I^1_{\mathcal{J},N}(\gamma)\},\end{aligned}$$

where the last inequality is by Theorem 5.7; thus, the left inequality in (5.31). We now show the right inequality in (5.31), i.e., $\min\{I^0_{\mathcal{J},N}(\gamma), I^1_{\mathcal{J},N}(\gamma)\} > 0$ for all $\gamma \in (\gamma_0, \gamma_1)$. First, from the expression for $I^0_{\mathcal{J},N}(\gamma)$ in Theorem 5.7, for $\mathcal{J} > 0$, we have: $I^0_{\mathcal{J},N}(\gamma_0) = NI_0(\gamma_0) = 0$, and ${I^0_{\mathcal{J},N}}'(\gamma) = NI_0'(\gamma) > 0$ for any $\gamma \in (\gamma_0, \gamma_0^-)$. As the function $I^0_{\mathcal{J},N}(\cdot)$ is convex, we conclude that $I^0_{\mathcal{J},N}(\gamma) > 0$, for all $\gamma > \gamma_0$. (The same conclusion holds under $\mathcal{J} = 0$, by replacing $NI_0(\gamma)$ with $I_0(\gamma) + \mathcal{J} = I_0(\gamma)$.) Analogously, it can be shown that $I^1_{\mathcal{J},N}(\gamma) > 0$ for all $\gamma < \gamma_1$; thus, $\min\{I^0_{\mathcal{J},N}(\gamma), I^1_{\mathcal{J},N}(\gamma)\} > 0$, for all $\gamma \in (\gamma_0, \gamma_1)$.

We now calculate $\max_{\gamma\in(\gamma_0,\gamma_1)} \min\{I^0_{\mathcal{J},N}(\gamma), I^1_{\mathcal{J},N}(\gamma)\}$. Consider the function $\Delta_B(\gamma) := I^1_{\mathcal{J},N}(\gamma) - I^0_{\mathcal{J},N}(\gamma)$. Using the definition of $I^0_{\mathcal{J},N}(\gamma)$ in Theorem 5.7, and taking the subdifferential of $I^0_{\mathcal{J},N}(\gamma)$ at any point $\gamma \in (\gamma_0, \gamma_1)$, it is easy to show that ${I^0_{\mathcal{J},N}}'(\gamma) > 0$, for any subgradient ${I^0_{\mathcal{J},N}}'(\gamma) \in \partial I^0_{\mathcal{J},N}(\gamma)$, which implies that $I^0_{\mathcal{J},N}(\cdot)$ is strictly increasing on $\gamma \in (\gamma_0, \gamma_1)$. Similarly, it can be shown that $I^1_{\mathcal{J},N}(\cdot)$ is strictly decreasing on $\gamma \in (\gamma_0, \gamma_1)$. Further, using the properties that $I_0(\gamma_0) = 0$ and $I_1(\gamma_1) = 0$, we have $\Delta_B(\gamma_0) = I^1_{\mathcal{J},N}(\gamma_0) > 0$, and $\Delta_B(\gamma_1) = -I^0_{\mathcal{J},N}(\gamma_1) < 0$. By the previous two observations, we have that $\Delta_B(\gamma)$ is strictly decreasing on $\gamma \in (\gamma_0, \gamma_1)$, with $\Delta_B(\gamma_0) > 0$ and $\Delta_B(\gamma_1) < 0$. Thus, $\Delta_B(\cdot)$ has a unique zero $\gamma^\star$ on (γ_0, γ_1). Now, the claim: $\max_{\gamma\in(\gamma_0,\gamma_1)} \min\{I^0_{\mathcal{J},N}(\gamma), I^1_{\mathcal{J},N}(\gamma)\} = I^0_{\mathcal{J},N}(\gamma^\star) = I^1_{\mathcal{J},N}(\gamma^\star)$ holds trivially because $I^0_{\mathcal{J},N}(\cdot)$ is strictly increasing on $\gamma \in (\gamma_0, \gamma_1)$, $I^1_{\mathcal{J},N}(\cdot)$ is strictly decreasing on $\gamma \in (\gamma_0, \gamma_1)$, and $I^0_{\mathcal{J},N}(\cdot)$ and $I^1_{\mathcal{J},N}(\cdot)$ intersect at $\gamma^\star \in (\gamma_0, \gamma_1)$. This completes the proof of part (a).

We now prove part (b). Suppose that $\mathcal{J} \geq \mathcal{J}^\star(\Lambda_0, N)$. We show that, for $\gamma = 0$:

$$I^0_{\mathcal{J},N}(0) = NI_0(0), \quad I^1_{\mathcal{J},N}(0) = NI_1(0) = NI_0(0). \tag{5.51}$$

(The last equality in (5.51) holds because $I_1(0) = (I_0(\gamma) - \gamma)|_{\gamma=0} = I_0(0)$.)

We prove only the equality for $I^0_{\mathcal{J},N}$ in (5.51) as the equality for $I^1_{\mathcal{J},N}$ follows similarly. Because $\mathcal{J} \geq \mathcal{J}^\star(\Lambda_0, N)$, we have, by the definition of $\Phi_0(\cdot)$ in (5.44), that $\Phi_0(\lambda^\bullet) = N\Lambda_0(\lambda^\bullet)$. Recall that $I^0_{\mathcal{J},N}(0) = -\Phi_0(\lambda^\star)$, where $\lambda^\star$ is a point for which (5.45) holds for $\gamma = 0$. However, because $\partial\Phi_0(\lambda^\bullet) = \{N\Lambda_0'(\lambda^\bullet)\}$, and $\Lambda_0'(\lambda^\bullet) = 0$, it follows that $\lambda^\star = \lambda^\bullet$ and $I^0_{\mathcal{J},N}(0) = -\Phi_0(\lambda^\bullet) = -N\Lambda_0(\lambda^\bullet) = NI_0(0)$, which proves (5.51).

Now, (5.51) means that $I^0_{\mathcal{J},N}(0) = I^1_{\mathcal{J},N}(0)$. Further, $0 \in (\gamma_0, \gamma_1)$, and, from part (a), $\gamma^\star$ is unique, and so $\gamma^\star$ has to be 0. This shows that $\sup_{\gamma\in(\gamma_0,\gamma_1)} \min\{I^0_{\mathcal{J},N}(\gamma), I^1_{\mathcal{J},N}(\gamma)\} = NI_0(0) = NC_{\text{ind}}$, and so, by part (a):

$$\liminf_{k\to\infty} -\frac{1}{k}\log P^{\text{e}}_{i,k}(0) \geq NC_{\text{ind}}. \tag{5.52}$$

On the other hand,

$$\limsup_{k\to\infty} -\frac{1}{k}\log P^{\text{e}}_{i,k}(0) \leq NC_{\text{ind}}, \tag{5.53}$$

because, by the Chernoff lemma [8], for *any test* (with the corresponding error probability $P^{\text{e}\prime}_k(\gamma)$,) we have that $\limsup_{k\to\infty} -\frac{1}{k}\log P^{\text{e}\prime}_k(\gamma) \leq NC_{\text{ind}}$. Combining (5.52) and (5.53) yields‘

$$NC_{\text{ind}} \leq \liminf_{k\to\infty} -\frac{1}{k}\log P^{\text{e}}_{i,k}(0) \leq \limsup_{k\to\infty} -\frac{1}{k}\log P^{\text{e}}_{i,k}(0) \leq NC_{\text{ind}}.$$

Thus, the result in part (b) of the lemma. □

5.3 Examples

This section illustrates our main results for several examples of the distributions of the agent observations. Subsection 5.3.1 compares the Gaussian and Laplace distributions, both with a finite number of agents N and when $N \to \infty$. Subsection 5.3.2 considers discrete distributions with finite support, and, in more detail, binary distributions. Finally, Subsection 5.3.3 numerically demonstrates that our theoretical lower bound on the error exponent (5.31) is tight. Subsection 5.3.3 also shows through a symmetric, tractable example how distributed detection performance depends on the network topology (agents' degree and link occurrence/failure probability.)

5.3.1 Gaussian distribution versus Laplace distribution

Gaussian distribution. We now study the detection of a signal in additive Gaussian noise; $Y_{i,t}$ has the following density:

$$f_{\text{G}}(y) = \begin{cases} \frac{1}{\sqrt{2\pi}\sigma_{\text{G}}} e^{-\frac{(y-m_{\text{G}})^2}{2\sigma^2_{\text{G}}}}, & H_1 \\ \frac{1}{\sqrt{2\pi}\sigma_{\text{G}}} e^{-\frac{y^2}{2\sigma^2_{\text{G}}}}, & H_0. \end{cases}$$

The logarithmic moment generating function is given by: $\Lambda_{0,\mathrm{G}}(\lambda) = -\frac{\lambda(1-\lambda)}{2}\frac{m_{\mathrm{G}}^2}{\sigma_{\mathrm{G}}^2}$. The minimum of $\Lambda_{0,\mathrm{G}}$ is achieved at $\lambda^\bullet = \frac{1}{2}$, and the per agent Chernoff information is

$$C_{\mathrm{ind,G}} = \frac{m_{\mathrm{G}}^2}{8\sigma_{\mathrm{G}}^2}.$$

Applying Corollary 5.8, we get the sufficient condition for optimality:

$$\mathcal{J} \geq \Lambda_{0,\mathrm{G}}\left(\frac{N}{2}\right) - N\Lambda_{0,\mathrm{G}}\left(\frac{1}{2}\right) = N(N-1)C_{\mathrm{ind,G}}. \tag{5.54}$$

Since $\Lambda_0(\lambda) = \Lambda_1(\lambda)$, the two conditions from the Corollary here reduce to a single condition in (5.31).

Now, let the number of agents $N \to \infty$, while keeping the total Chernoff information constant, i.e., not dependent on N; that is, $C_{\mathrm{G}} := NC_{\mathrm{ind,G}} = \mathrm{const}$, $C_{\mathrm{ind,G}}(N) = C_{\mathrm{G}}/N$. Intuitively, as N increases, we deploy more and more agents over a region (denser deployment), but, on the other hand, the agents' quality becomes worse and worse. The increase of N is balanced in such a way that the total information offered by all agents stays constant with N. Our goal is to determine how the optimality threshold on the network connectivity $\mathcal{J}^\star(N, \Lambda_{0,\mathrm{G}})$ depends on N. We can see from (5.54) that the optimality threshold for the distributed detector in the Gaussian case equals:

$$\mathcal{J}^\star(\Lambda_{0,\mathrm{G}}, N) = (N-1)C_{\mathrm{G}}. \tag{5.55}$$

Laplace distribution. We next study the optimality conditions for the agent observations with Laplace distribution. The density of $Y_{i,t}$ is:

$$f_{\mathrm{L}}(y) = \begin{cases} \frac{1}{2b_{\mathrm{L}}}e^{-\frac{|y-m_{\mathrm{L}}|}{b_{\mathrm{L}}}}, & H_1 \\ \frac{1}{2b_{\mathrm{L}}}e^{-\frac{|y|}{b_{\mathrm{L}}}}, & H_0. \end{cases}$$

The logarithmic moment generating function has the following form:

$$\Lambda_{0,\mathrm{L}}(\lambda) = \log\left(\frac{1-\lambda}{1-2\lambda}e^{-\lambda\frac{m_{\mathrm{L}}}{b_{\mathrm{L}}}} - \frac{\lambda}{1-2\lambda}e^{-(1-\lambda)\frac{m_{\mathrm{L}}}{b_{\mathrm{L}}}}\right).$$

Again, the minimum is at $\lambda^\bullet = \frac{1}{2}$, and the per agent Chernoff information is $C_{\mathrm{ind,L}} = \frac{m_{\mathrm{L}}}{2b_{\mathrm{L}}} - \log\left(1 + \frac{m_{\mathrm{L}}}{2b_{\mathrm{L}}}\right)$.

The optimality condition in (5.31) becomes:

$$\begin{aligned}\mathcal{J} &\geq \Lambda_{0,\mathrm{L}}\left(\frac{N}{2}\right) - N\Lambda_{0,\mathrm{L}}\left(\frac{1}{2}\right) \\ &= \log\left(\frac{2-N}{2-2N}e^{-\frac{N}{2}\frac{m_{\mathrm{L}}}{b_{\mathrm{L}}}} - \frac{N}{2-2N}e^{-(1-\frac{N}{2})\frac{m_{\mathrm{L}}}{b_{\mathrm{L}}}}\right) - N\log\left(1+\frac{m_{\mathrm{L}}}{2b_{\mathrm{L}}}\right) + N\frac{m_{\mathrm{L}}}{2b_{\mathrm{L}}}.\end{aligned} \tag{5.56}$$

Gaussian versus Laplace distribution. It is now interesting to compare the Gaussian and the Laplace case under equal per agent Chernoff information $C_{\mathrm{ind,L}} = C_{\mathrm{ind,G}}$. Figure 5.2 (left) plots the logarithmic moment generating function for the Gaussian and Laplace distributions, for $N = 10$, $C_{\mathrm{ind}} = C_{\mathrm{ind,L}} = C_{\mathrm{ind,G}} = 0.0945$, $b_{\mathrm{L}} = 1$, $m_{\mathrm{L}} = 1$, and $m_{\mathrm{G}}^2/\sigma_{\mathrm{G}}^2 = 0.7563 = 8C_{\mathrm{ind}}$. By (5.32), the optimality threshold equals

$$|N\Lambda_0(1/2)| + |\Lambda_0(N/2)|,$$

as $\lambda^\bullet = 1/2$, for both the Gaussian and Laplace distributions. The threshold can be estimated from Figure 5.2 (left): solid lines plot the functions $\Lambda_0(N\lambda)$ for the two different distributions, while dashed lines plot the functions $N\,\Lambda_0(\lambda)$. For both solid and dashed lines, the Gaussian distribution corresponds to the more curved functions. We see that the threshold is larger for the Gaussian case. This means that, for a certain range $\mathcal{J} \in (\mathcal{J}_{\min}, \mathcal{J}_{\max})$, the distributed detector with Laplace agents is asymptotically optimal, while with Gaussian agents the distributed detector may not be optimal, even though it uses the same network infrastructure (equal r) and has equal per agent Chernoff information. (See also Figure 5.2 (right) for another illustration of this effect.)

We now compare the Gaussian and Laplace distributions when $N \to \infty$, and we keep the Gaussian total Chernoff information C_{G} constant with N. Let the Laplace distribution parameters vary with N as:

$$m_{\mathrm{L}} = m_{\mathrm{L}}(N) = \frac{2\sqrt{2C_{\mathrm{G}}}}{\sqrt{N}}, \quad b_{\mathrm{L}} = b_{\mathrm{L}}(N) = 1.$$

We can show that, as $N \to \infty$, the total Chernoff information $C_{\mathrm{L}}(N) \to C_{\mathrm{G}}$ as $N \to \infty$, and so the Gaussian and the Laplace centralized detectors become equivalent. On the other hand, the threshold for the Gaussian distributed detector is given by (5.55) while, for the Laplace detector, using (5.56) and a Taylor expansion, we get that the optimality threshold is approximately:

$$\mathcal{J}^\star(\Lambda_{0,\mathrm{L}}, N) \approx \sqrt{2C_{\mathrm{G}}N}.$$

Hence, the required $\mathcal{J}$ to achieve the optimal error exponent grows much slower with the Laplace distribu-

tion than with the Gaussian distribution.

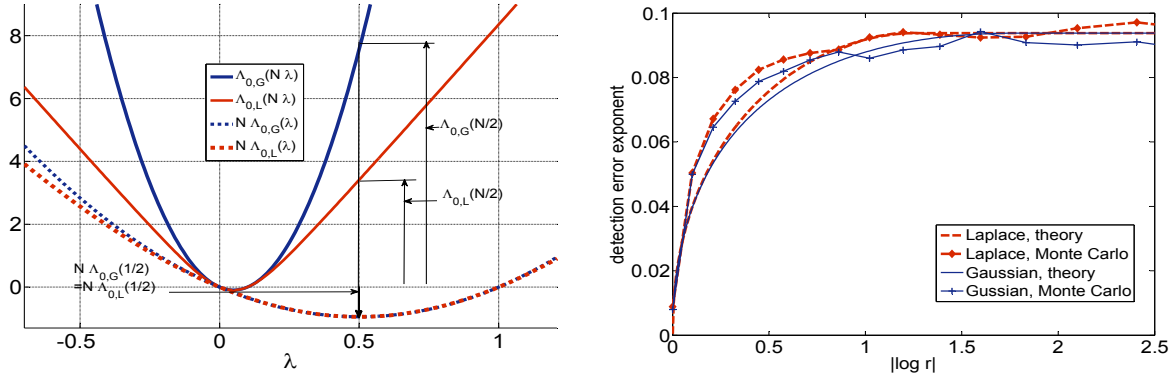

Figure 5.2: **Left:** Logarithmic moment generating functions for Gaussian and Laplace distributions with equal per agent Chernoff information, for $N = 10$, $C_{\text{ind}} = C_{\text{ind,L}} = C_{\text{ind,G}} = 0.0945$, $b_{\text{L}} = 1$, $m_{\text{L}} = 1$, and $m_{\text{G}}^2/\sigma_{\text{G}}^2 = 0.7563 = 8C_{\text{ind}}$. Solid lines plot the functions $\Lambda_0(N\lambda)$ for the two distributions, while dashed lines plot the functions $N\Lambda_0(\lambda)$. For both solid and dashed lines, the Gaussian distribution corresponds to the more curved functions. The optimality threshold in (5.32) is given by $|N\Lambda_0(1/2)| + |\Lambda_0(N/2)|$, as $\lambda^\bullet = 1/2$. **Right:** Lower bound on the error exponent in (5.31) and the Monte Carlo estimate of the error exponent versus $\mathcal{J}$ for the Gaussian and Laplace agent observations: $N = 20$, $C_{\text{ind}} = C_{\text{ind,L}} = C_{\text{ind,G}} = 0.005$, $b_{\text{L}} = 1$, $m_{\text{L}} = 0.2$, and $m_{\text{G}}^2/\sigma_{\text{G}}^2 = 0.04 = 8C_{\text{ind}}$.

5.3.2 Discrete distributions

We now consider the case when the support of the agent observations under both hypothesis is a finite alphabet $\{a_1, a_2, ..., a_M\}$. This case is of practical interest when, for example, the sensing device has an analog-to-digital converter with a finite range; hence, the observations take only a finite number of values. Specifically, the distribution of $Y_{i,k}$, $\forall i$, $\forall k$, is given by:

$$\mathbb{P}(Y_{i,k} = a_m) = \begin{cases} q_m, & H_1 \\ p_m, & H_0 \end{cases}, \quad m = 1, ..., M. \tag{5.57}$$

Then, the logarithmic moment generating function under H_0 equals:

$$\Lambda_0(\lambda) = \log\left(\sum_{m=1}^{M} q_m^\lambda p_m^{1-\lambda}\right).$$

Note that $\Lambda_0(\lambda)$ is finite on $\mathbb{R}$. Due to concavity of $-\Lambda_0(\cdot)$, the argument of the Chernoff information $\lambda^\bullet$ ($C_{\text{ind}} = \max_{\lambda\in[0,1]}\{-\Lambda_0(\lambda)\} = -\Lambda_0(\lambda^\bullet)$) can, in general, be efficiently computed numerically, for example, by the Netwon method (see, e.g., [83], for details on the Newton method.) It can be shown,

defining $c_m = \log\left(\frac{q_m}{p_m}\right)$, that the Newton direction, e.g., [83], equals:

$$d(\lambda) = -\left(\Lambda_0''(\lambda)\right)^{-1}\Lambda_0'(\lambda) = -\frac{1}{\frac{\sum_{m=1}^M c_m^2 p_m e^{\lambda c_m}}{\sum_{m=1}^M c_m p_m e^{\lambda c_m}} - \frac{\sum_{m=1}^M c_m p_m e^{\lambda c_m}}{\sum_{m=1}^M c_m e^{\lambda c_m}}}.$$

Binary observations. To gain more intuition and obtain analytical results, we consider (5.57) with $M = 2$, i.e., binary agents,

$$\mathbb{P}(Y_{i,t} = \gamma_m) = \begin{cases} q_m, & H_1 \\ p_m, & H_0 \end{cases}, \quad m = 1, 2,$$

with $p_2 = 1 - p_1 = 1 - p$, $q_2 = 1 - q_1 = 1 - q$. Suppose further that $p < q$. We can show that the negative of the per agent Chernoff information $\Lambda_{0,\text{bin}}$ and the quantity $\lambda^\bullet$ are:

$$-C_{\text{ind}} = \Lambda_{0,\text{bin}}(\lambda^\bullet) = \lambda^\bullet \log\left(\frac{q}{p}\right) + \log p + \log\left(1 - \frac{\log\frac{q}{p}}{\log\frac{1-q}{1-p}}\right), \quad \lambda^\bullet = \frac{\log\frac{p}{1-p} + \log\left(\frac{\log\frac{p}{q}}{\log\frac{1-q}{1-p}}\right)}{\log\left(\frac{1-q}{1-p}\right) - \log\left(\frac{q}{p}\right)}.$$

Further, note that:

$$\Lambda_{0,\text{bin}}(N\lambda^\bullet) \;=\; \log\left(p\left(\frac{q}{p}\right)^{N\lambda^\bullet} + (1-p)\left(\frac{1-q}{1-p}\right)^{N\lambda^\bullet}\right) \leq \log\left(\frac{q}{p}\right)^{N\lambda^\bullet} = N\lambda^\bullet \log\left(\frac{q}{p}\right) \quad (5.58)$$

Also, we can show similarly that:

$$\Lambda_{0,\text{bin}}(1 - N(1-\lambda^\bullet)) \leq N(1-\lambda^\bullet)\log\left(\frac{1-p}{1-q}\right). \quad (5.59)$$

Combining (5.58) and (5.59), and applying Corollary 5.8 (equation (5.31)), we get that a sufficient condition for asymptotic optimality is:

$$\mathcal{J} \geq \max\left\{N\log\frac{1}{p} - N\log\left(1 + \frac{\left|\log\frac{q}{p}\right|}{\left|\log\frac{1-q}{1-p}\right|}\right), N\log\frac{1}{1-q} - N\log\left(1 + \frac{\left|\log\frac{1-q}{1-p}\right|}{\left|\log\frac{q}{p}\right|}\right)\right\}.$$

From the equation above, we can further obtain a very simplified sufficient condition for optimality:

$$\mathcal{J} \geq N\max\left\{|\log p|, |\log(1-q)|\right\}. \quad (5.60)$$

The expression in (5.60) is intuitive. Consider, for example, the case $p = 1/2$, so that the right hand side in (5.60) simplifies to: $N|\log(1-q)|$. Let q vary from $1/2$ to 1. Then, as q increases, the per agent

Chernoff information increases, and the optimal centralized detector has better and better performance (error exponent.) That is, the centralized detector has a very low error probability after a very short observation interval k. Hence, for larger q, the distributed detector needs more connectivity to be able to "catch up" with the performance of the centralized detector. We compare numerically Gaussian and binary distributed detectors with equal per agent Chernoff information, for $N = 32$ agents, $C_{\rm ind} = 5.11 \cdot 10^{-4}$, $m_G^2/\sigma_G^2 = 8C_{\rm ind}$, $p = 0.1$, and $q = 0.12$. Binary detector requires more connectivity to achieve asymptotic optimality ($\mathcal{J} \approx 1.39$), while Gaussian detector requires $\mathcal{J} \approx 0.69$

5.3.3 Tightness of the error exponent lower bound in (5.31) and the impact of the network topology

Assessment of the tightness of the error exponent lower bound in (5.31)**.** We note that the result in (5.31) is a theoretical lower bound on the error exponent. In particular, the condition $\mathcal{J} \geq \mathcal{J}^\star(\Lambda_0, N)$ is proved to be a sufficient, but not necessary, condition for asymptotically optimal detection; in other words, (5.31) does not exclude the possibility of achieving asymptotic optimality for a certain value of $\mathcal{J}$ smaller than $\mathcal{J}^\star(\Lambda_0, N)$. In order to assess the tightness of (5.31) (for both the Gaussian and Laplace distributions,) we perform Monte Carlo simulations to estimate the actual error exponent and compare it with (5.31). We consider $N = 20$ agents and fix the agent observation distributions with the following parameters: $C_{\rm ind} = C_{\rm ind,L} = C_{\rm ind,G} = 0.005$, $b_{\rm L} = 1$, $m_{\rm L} = 0.2$, and $m_{\rm G}^2/\sigma_{\rm G}^2 = 0.04 = 8C_{\rm ind}$. We vary $\mathcal{J}$ as follows. We construct a (fixed) geometric graph with N agents by placing the agents uniformly at random on a unit square and connecting the agents whose distance is less than a radius. Each link is a Bernoulli random variable, equal to 1 with probability p (link online), and equal to 0 with probability $1 - p$ (link offline). The link occurrences are independent in time and space. We change $\mathcal{J}$ by varying p from 0 to 0.95 in the increments of 0.05. We adopt the standard time-varying Metropolis weights: whenever a link $\{i, j\}$ is online, we set $[W_k]_{ij} = 1/(1 + \max(d_{i,k}, d_{j,k}))$, where $d_{i,k}$ is the number od neighbors of agent i at time k; when a link $\{i, j\}$ is offline, $[W_k]_{ij} = 0$; and $[W_k]_{ii} = 1 - \sum_{j \in O_{i,k}} [W_k]_{ij}$, where we recall that $O_i(k)$ is the neighborhood of agent i. We obtain an estimate of the error probability $\widehat{P}^{\rm e}_{i,k}$ at agent i and time k using 30,000 Monte Carlo runs of (5.19) per each hypothesis. We then estimate the agent-wide average error exponent as:

$$\frac{1}{N} \sum_{i=1}^{N} \frac{\log \widehat{P}^{\rm e}_{i,K_1} - \log \widehat{P}^{\rm e}_{i,K_2}}{K_2 - K_1},$$

with $K_1 = 40$, $K_2 = 60$. That is, we estimate the error exponent as the average slope (across agents) of the error probability curve in a semi-log scale. Figure 5.2 (right) plots both the theoretical lower bound on

the error exponent in (5.31) and the Monte Carlo estimate of the error exponent versus $\mathcal{J}$ for Gaussian and Laplace distributions. We can see that the bound (5.31) is tight for both distributions. Hence, the actual distributed detection performance is very close to the performance predicted by (5.31). (Of course, above the optimality threshold, (5.31) and the actual error exponent coincide and are equal to the total Chernoff information.) Also, we can see that the theoretical threshold on optimality $\mathcal{J}^\star(\Lambda_0, N)$ and the threshold value computed from simulation are very close. Finally, the distributed detector with Laplace observations achieves asymptotic optimality for a smaller value of $\mathcal{J}$ ($\mathcal{J} \approx 1.2$) than the distributed detector with Gaussian observations ($\mathcal{J} \approx 1.6$), even though the corresponding centralized detectors are asymptotically equivalent.

Impact of the network topology. We have seen in the previous two subsections how detection performance depends on $\mathcal{J}$. In order to understand how $\mathcal{J}$ depends on the network topology, we consider a symmetric network structure, namely a regular network. For this case, we can express $\mathcal{J}$ as an explicit (closed form) function of the agents' degrees and the link occurrence probabilities. (Recall that the larger $\mathcal{J}$ is, the better the network connectivity.)

Consider a connected regular network with N agents and degree $d \geq 2$. Suppose that each link is a Bernoulli random variable, equal to 1 with probability p (link online) and 0 with probability $1 - p$ (link offline,) with spatio-temporally independent link occurrences. Then, as we show in Chapter 2, $\mathcal{J}$ equals:

$$\mathcal{J} = d|\log(1-p)|. \tag{5.61}$$

This expression is very intuitive. When p increases, i.e., when the links are online more often, the network (on average) becomes more connected, and hence we expect that the network connectivity $\mathcal{J}$ increases (improves). This is confirmed with expression (5.61). Further, when d increases, the network becomes more connected, and hence the network speed again improves. Note also that $\mathcal{J}$ is a linear function of d.

We now recall Corollary 5.8 to relate distributed detection performance with p and d. For example, for a fixed p, the distributed detection optimality condition becomes $d > \frac{\mathcal{J}^\star(\Lambda_0,N)}{|\log(1-p)|}$, i.e., distributed detection is asymptotically optimal when the agents' degree is above a threshold. Further, because $d \leq N$, it follows that, for a large value of $\mathcal{J}^\star(\Lambda_0, N)$ and a small p, even the networks with a very large degree (say, $d = N - 1$) do not achieve asymptotic optimality. Intuitively, a large $\mathcal{J}^\star(\Lambda_0, N)$ means that the corresponding centralized detector decreases the error probability so fast in k that, because of the intermittent link failures, the distributed detector cannot parallel in performance the centralized detector. Finally, when $p = 1$, the optimality condition becomes $d > 0$, i.e., distributed detection is asymptotically optimal for any $d \geq 2$.

This is because, when $p = 1$, the network is always connected, and the distributed detector asymptotically "catches up" with an arbitrarily fast centralized detector. In fact, it can be shown that an arbitrarily connected network with no link failures achieves asymptotic optimality for any value of $\mathcal{J}^\star(\Lambda_0, N)$. (Such a network has the network connectivity $\mathcal{J}\ +\infty$.)

5.4 Simulations

In this section, we present a simulation example for the distributed detection algorithm where the observations $Y_{i,t}$ are correlated Gaussian. To account for the correlations, we correspondingly adjust the values of the innovations $L_{i,t}$; we detail this adjustment below. Our simulation results demonstrate that the distributed detector with correlated observations exhibits a similar behavior with respect to the network connectivity $\mathcal{J}$ as its i.i.d. counterpart: when $\mathcal{J}$ is sufficiently large, the error probability of each node in the network decays at the optimal rate equal to the total Chernoff information in the network. Further, we demonstrate that a sensor with poor connectedness to the rest of the network cannot be an optimal detector,and, moreover, its performance approaches the performance of an isolated sensor, i.e., a sensor that works as an individual detector, as connectedness becomes worse and worse.

Simulation setup. We consider a network (V, E) with $N = 40$ nodes and $M = 247$ edges. Nodes are uniformly distributed on a unit square and nodes within distance less than a radius ℓ are connected by an edge. As averaging weights, we use the standard Metropolis weights.The link failures are spatially and temporally independent. Each link $\{i, j\} \in E$ has the same probability of formation, i.e., the probability of being online at a time, $q_{ij} = q$. This network and weight model satisfy Assumption 5.5.

We assume equal prior probabilities, $\pi_0 = \pi_1 = 0.5$, and thus we set the threshold γ to be the optimal threshold $\gamma = 0$. We assume that the vector of all observations Y_t is Gaussian with mean value $m_1 = 1_N$ under hypothesis H_1 and mean value $m_0 = 0_N$ under hypothesis H_0, with the same covariance matrix S under both hypothesis. We generate randomly the covariance matrix S, as follows. We generate: a $N \times N$ matrix M_S, with the entries drawn independently from $U[0, 1]$–the uniform distribution on $[0, 1]$; we set $R_S = M_S M_S^\top$; we decompose R_S via the eigenvalue decomposition: $R_S = Q_S \Lambda_S Q_S^\top$; we generate a $N \times 1$ vector u_S with the entries drawn independently from $U[0, 1]$; finally, we set $S = \alpha_S\, Q_S \mathrm{Diag}(u_S) Q_S^\top$, where $\alpha_S > 0$ is a parameter. It can be shown that the log-likelihood ratio for this problem (after k observations are processed) has the form $D_k := \sum_{t=1}^{k} (m_1 - m_0)^\top S^{-1} \left(Y_t - \frac{m_1+m_0}{2}\right)$, and the decision test consists of comparing D_k with 0. Observe that D_k is Gaussian random variable with mean value $m_{D,1} := \frac{(m_1-m_0)^\top S^{-1}(m_1-m_0)}{2}$ under hypothesis H_1, mean value $m_{D,0} := -m_{D,1}$ under the hypoth-

esis H_0, and variance $(m_1 - m_0)^\top S^{-1}(m_1 - m_0)$ (under both hypotheses). This implies that, for the optimal centralized detector, the error probability $P_k^{\mathrm{e}}(0)$ can be computed by the Q function and equals: $P_k^{\mathrm{e}}(0) = \pi_0 \mathbb{P}\left(D_k \geq 0 | H_0\right) + \pi_1 \mathbb{P}\left(D_k < 0 | H_1\right) = Q\left(k \frac{(m_1-m_0)^\top S^{-1}(m_1-m_0)}{8}\right)$. Also, the total Chernoff information in the network is equal to $C_{\mathrm{tot}} = \frac{(m_1-m_0)^\top S^{-1}(m_1-m_0)}{8}$.

Taking into account the correlations between the observations $Y_{i,t}$, we run the distributed detector (5.19) with $L_{i,t} = v_i \left(Y_{i,t} - \frac{m_{i,1}-m_{i,0}}{2}\right)$, where $v = S^{-1}(m_1 - m_0)$. Note that this choice of $L_{i,t}$'s aims at computing the optimal decision statistics D_k in a distributed way: at each time t, $\sum_{i=1}^N L_{i,t} = D_t$. We evaluate $P_{i,k}^{\mathrm{e}}(0)$ by Monte Carlo simulations with 20,000 sample paths (20,000 for each hypothesis H_l, $l = 0, 1$) of the running consensus algorithm.

Exponential rate of decay of the error probability vs. the network connectivity $\mathcal{J}$. First, we examine the asymptotic behavior of distributed detection when the network connectivity $\mathcal{J}$ varies. To this end, we fix the graph $G = (V, E)$, and then we vary the formation probability of links q from 0 to 0.75. Figure 5.3 (right) plots the estimated exponential rate of decay, averaged across sensors, versus q. For q greater than 0.1 the rate of decay of the error probability is approximately the same as for the optimal centralized detector $\mathbf{C}_{\mathrm{tot}}$–the simulation estimate of $\mathbf{C}_{\mathrm{tot}}$ is 0.0106. [6] For $q < 0.1$ the detection performance becomes worse and worse as q decreases. Figure 5.3 (left) plots the estimated error probability, averaged across sensors, for different values of q. We can see that the curves are "stretched" for small values of q; after q exceeds a threshold (on the order of 0.1,) the curves cluster, and they have approximately the same slope (the error probability has approximately the same decay rate,) equal to the optimal slope.

Study of a sensor with poor connectivity to the rest of the network. Next, we demonstrate that a sensor with poor connectivity to the rest of the network cannot be an asymptotically optimal detector;its performance approaches the performance of an individual detector-sensor, when its connectivity becomes worse and worse. For the i-th individual detector-sensor (no cooperation between sensors), it is easy to show that the Bayes probability of error, $P_{i,k}^{\text{e no cooper.}}$ equals: $P_{i,k}^{\text{e no cooper.}} = Q\left(\sqrt{k} \frac{m_{i,\text{no cooper.}}}{\sigma_{\text{i,no cooper.}}}\right)$, where $m_{i,\text{no cooper.}} = \frac{1}{2}\frac{[m_1-m_0]_i^2}{S_{ii}}$, and $\sigma^2_{\text{i,no cooper.}} = \frac{[m_1-m_0]_i^2}{S_{ii}}$. It is easy to show that the Chernoff information (equal to $\lim_{k\to\infty} \frac{1}{k} \log P_{i,k}^{\text{e no cooper.}}$) for sensor i, in the absence of cooperation, is given by $\frac{1}{8}\frac{[m_1]_i^2}{S_{ii}}$.

We now detail the simulation setup. We consider a network with $N = 35$ nodes and $M = 263$ edges. We initially generate the graph as a geometric disc graph, but then we isolate sensor 35 from the rest of the network, by keeping it connected only to sensor 3. We then vary the formation probability of the link $\{3, 35\}$, $q_{3,35}$, from 0.05 to 0.5 (see Figure 5.4.) All other links in the supergraph have the formation

[6] In this numerical example, the theoretical value of $\mathbf{C}_{\mathrm{tot}}$ is 0.009. The estimated value shows an error because the decay of the error probability, for the centralized detection, and for distributed detection with a large q, tends to slow down slightly when k is very large; this effect is not completely captured by the simulation with $k < 700$.

probability of 0.8. Figure 5.4 plots the error probability for: 1) the optimal centralized detection; 2) the distributed detection at each sensor, with cooperation (running consensus;) and 3) the distributed detection at each sensor, without cooperation (sensors do not communicate.) Figure 5.4 shows that, when $q_{3,35} = 0.05$, sensor 35 behaves almost as bad as the individual sensors that do not communicate (cooperate) with each other. As q increases, the performance of sensor 35 gradually improves.

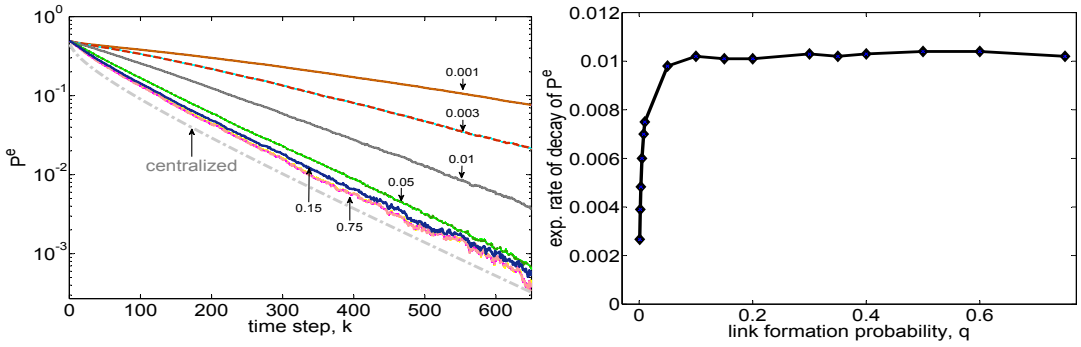

Figure 5.3: Monte Carlo estimate of the performance of distributed detection for different values of the link formation probability q. Left: Error probability averaged across N sensors. Each line is labeled with the value of q; performance of centralized detection is plotted in gray. Right: Estimated exponential rate of decay of the error probability vs. q.

5.5 Non-identically distributed observations

We extend Theorem 5.7 and Corollary 5.8 to the case of (independent) *non-identically* distributed observations. First, we briefly explain the measurement model and define the relevant quantities. As before, let $Y_{i,t}$ denote the observation of agent i at time t, $i = 1, \ldots, N$, $t = 1, 2, \ldots$.

Assumption 5.12 The observations of agent i are i.i.d. in time, with the following distribution:

$$Y_{i,t} \sim \begin{cases} \nu_{i,1}, & H_1 \\ \nu_{i,0}, & H_0 \end{cases}, \quad i = 1, ..., N, \ t = 1, 2, ...$$

(Here we assume that $\nu_{i,1}$ and $\nu_{i,0}$ are mutually absolutely continuous, distinguishable measures, for $i = 1, \ldots, N$). Further, the observations of different agents are independent both in time and across agents, i.e., for $i \neq j$, $Y_{i,t}$ and $Y_{j,k}$ are independent for all t and k.

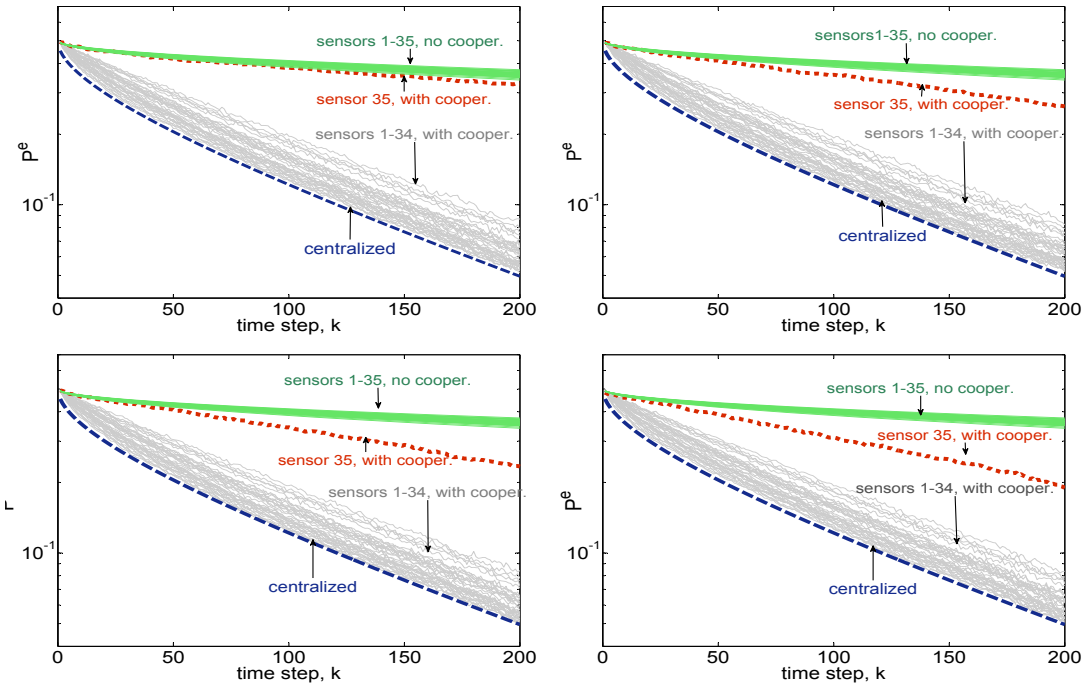

Figure 5.4: Error probability averaged across sensors for the optimal centralized detection, distributed detection at each sensor (with cooperation), and detection at each sensor, without cooperation. The formation probability $q_{3,35}$ of the link $\{3, 35\}$ varies between 0.05 and 0.5: $q_{3,35}$=0.05 (top right); 0.2 (top left); 0.3 (bottom left); 0.5 (bottom right).

Under Assumption 5.12, the form of the log-likelihood ratio test remains the same as under Assumption 5.1:

$$D_k := \frac{1}{Nk} \sum_{t=1}^{k} \sum_{i=1}^{N} L_{i,t} \underset{H_0}{\overset{H_1}{\gtrless}} \gamma,$$

where the log-likelihood ratio at agent i, $i = 1, ..., N$, is now:

$$L_{i,t} = \log \frac{f_{i,1}(Y_{i,t})}{f_{i,0}(Y_{i,t})},$$

and $f_{i,l}$, $l = 0, 1$, is the density (or the probability mass) function associated with $\nu_{i,l}$. We now discuss the choice of detector thresholds γ. Let $\overline{\gamma}_l = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N} L_{i,t}|H_l\right] = \left(\sum_{i=1}^{N} \gamma_{i,l}\right)/N$. We can show that, if $\mathcal{J} > 0$, then any $\gamma \in (\overline{\gamma}_0, \overline{\gamma}_1)$ yields an exponentially fast decay of the error probability, at any agent. The condition $\mathcal{J} > 0$ means that the network is connected on average, see Chapter 2; if met, then, it is not difficult to show that for all i, $\mathbb{E}[x_{i,k}|H_l] \to \overline{\gamma}_l$ as $k \to \infty$, $l = 0, 1$. Clearly, under identical agents, $\gamma_{i,l} = \gamma_{j,l}$ for any i, j, and hence the range of detector thresholds becomes the one assumed in Section 5.1.3.

Denote by $\Lambda_{i,0}$ the logarithmic moment generating function of $L_{i,t}$ under hypothesis H_0:

$$\Lambda_{i,0} : \mathbb{R} \longrightarrow (-\infty, +\infty], \quad \Lambda_{i,0}(\lambda) = \log \mathbb{E}\left[e^{\lambda L_{i,1}}|H_0\right].$$

We assume finiteness of $\Lambda_{i,0}(\cdot)$ of all agents. Assumption 5.3 is restated explicitly as Assumption 5.13.

Assumption 5.13 For $i = 1, \ldots N$, $\Lambda_{i,0}(\lambda) < +\infty$, $\forall \lambda \in \mathbb{R}$.

The optimal centralized detector, with highest error exponent, is the log-likelihood ratio test with zero threshold $\gamma = 0$ [8], its error exponent is equal to the Chernoff information of the vector of all agents observations, and can be expressed in terms of the logarithmic moment generating functions as:

$$C_{\text{tot}} = \max_{\lambda \in [0,1]} -\sum_{i=1}^{N} \Lambda_{i,0}(\lambda) = -\sum_{i=1}^{N} \Lambda_{i,0}(\lambda^{\bullet}).$$

Here, $\lambda^{\bullet}$ is the minimizer of $\sum_{i=1}^{N} \Lambda_{i,0}$ over $[0, 1]$. We are now ready to state our results on the error exponent of the distributed detector for the case of non-identically distributed observations. (We continue to use $\alpha_{i,k}(\gamma)$, $\beta_{i,k}(\gamma)$, and $P^{\text{e}}_{i,k}(\gamma)$ to denote the false alarm, miss, and Bayes error probabilities of distributed detector at agent i.)

Theorem 5.14 Let Assumptions 5.5, 5.12, and 5.13 hold, and let, in addition, $\mathcal{J} > 0$. Consider the family

of distributed detectors in (5.19) and (5.21) with thresholds $\gamma \in (\overline{\gamma}_0, \overline{\gamma}_1)$. Then, at each agent i:

$$\liminf_{k\to\infty} -\frac{1}{k}\log \alpha_{i,k}(\gamma) \geq I^0_{\mathcal{J},N}(\gamma) > 0, \;\; \liminf_{k\to\infty} -\frac{1}{k}\log \beta_{i,k}(\gamma) \geq I^1_{\mathcal{J},N}(\gamma) > 0, \tag{5.62}$$

where

$$I^0_{\mathcal{J},N}(\gamma) = \max_{\lambda\in[0,1]} N\lambda\gamma - \max\left\{\sum_{i=1}^N \Lambda_{i,0}(\lambda), \max_{i=1,\ldots,N} \Lambda_{i,0}(N\lambda) - \mathcal{J}\right\} \tag{5.63}$$

$$I^1_{\mathcal{J},N}(\gamma) = \max_{\lambda\in[-1,0]} N\lambda\gamma - \max\left\{\sum_{i=1}^N \Lambda_{i,1}(\lambda), \max_{i=1,\ldots,N} \Lambda_{i,1}(N\lambda) - \mathcal{J}\right\}. \tag{5.64}$$

Corollary 5.15 Let Assumptions 5.5, 5.12, and 5.13 hold, and let, in addition, $\mathcal{J} > 0$. Consider the family of distributed detectors in (5.19) and (5.21) with thresholds $\gamma \in (\overline{\gamma}_0, \overline{\gamma}_1)$. Then:

(a) At each agent i:

$$\liminf_{k\to\infty} -\frac{1}{k}\log P^{\mathrm{e}}_{i,k}(\gamma) \geq \min\{I^0_{\mathcal{J},N}(\gamma), I^1_{\mathcal{J},N}(\gamma)\} > 0, \tag{5.65}$$

and the lower bound in (5.65) is maximized for the point $\gamma^\star \in (\overline{\gamma}_0, \overline{\gamma}_1)$ at which $I^0_{\mathcal{J},N}(\gamma^\star) = I^1_{\mathcal{J},N}(\gamma^\star)$.

(b) Consider $\lambda^\bullet = \arg\min_{\lambda\in[0,1]} \sum_{i=1}^N \Lambda_{i,0}(\lambda)$, and let:

$$\mathcal{J}^\star\left(\Lambda_{1,0},\ldots,\Lambda_{N,0}\right) = \tag{5.66}$$
$$\max\left\{\max_{i=1,\ldots,N} \Lambda_{i,0}(N\lambda^\bullet) - \sum_{i=1}^N \Lambda_{i,0}(\lambda^\bullet), \max_{i=1,\ldots,N} \Lambda_{i,0}(1-N(1-\lambda^\bullet)) - \sum_{i=1}^N \Lambda_{i,0}(\lambda^\bullet)\right\}.$$

Then, when $\mathcal{J} \geq \mathcal{J}^\star\left(\Lambda_{1,0},\ldots,\Lambda_{N,0}\right)$, each agent i with the detector threshold set to $\gamma = 0$, is asymptotically optimal:

$$\lim_{k\to\infty} -\frac{1}{k}\log P^{\mathrm{e}}_{i,k}(0) = C_{\mathrm{tot}}.$$

Comparing Theorem 5.7 with Theorem 5.14, we can see that, under non-identically distributed observations, it is no longer possible to analytically characterize the lower bounds on the error exponents, $I^0_{\mathcal{J},N}(\gamma)$ and $I^1_{\mathcal{J},N}(\gamma)$. However, the objective functions (in the variable λ) in (5.63) and (5.64) are concave (by convexity of the logarithmic moment generating functions) and the underlying optimization variable λ is a scalar, and, thus, $I^0_{\mathcal{J},N}(\gamma)$ and $I^1_{\mathcal{J},N}(\gamma)$ can be efficiently found by a one dimensional numerical optimization procedure, e.g., a subgradient algorithm [72].

The proof of Theorem 5.14 mimics the proof of Theorem 5.7; we focus only on the steps that account for different agents' logarithmic moment generating functions. The proof of Corollary 5.15 is omitted.

Proof [Proof of Theorem 5.14] First, expression (5.36) that upper bounds the probability of false alarm $\alpha_{i,k}(\gamma)$ for the case of non-identically distributed observations becomes:

$$\begin{aligned}\mathbb{E}\left[e^{N\lambda\sum_{t=1}^{k}\sum_{j=1}^{N}[\Phi(k,t)]_{i,j}L_{j,t}}|H_0\right] &= \mathbb{E}\left[\mathbb{E}\left[e^{N\lambda\sum_{t=1}^{k}\sum_{j=1}^{N}[\Phi(k,t)]_{i,j}L_{j,t}}|H_0,W_1,\ldots,W_k\right]\right]\\ &= \mathbb{E}\left[e^{\sum_{t=1}^{k}\sum_{j=1}^{N}\Lambda_{j,0}(N\lambda[\Phi(k,t)]_{i,j})}\right].\end{aligned}$$

Next, we bound the sum in the exponent of the previous equation, conditioned on the event $\mathcal{A}_s$, for a fixed s in $\{0,1,\ldots,k\}$, deriving a counterpart to Lemma 5.10.

Lemma 5.16 Let Assumptions 5.5, 5.12, and 5.13 hold. Then,

(a) For any realization of W_t, $t = 1, 2, ..., k$:

$$\sum_{j=1}^{N}\Lambda_{j,0}\left(N\lambda[\Phi(k,t)]_{i,j}\right) \leq \max_{j=1,\ldots,N}\Lambda_{j,0}\left(N\lambda\right),\ \forall t=1,\ldots,k.$$

(b) Consider a fixed s in $\{0,1,...,k\}$. If the event $\mathcal{A}_s$ occurred, then, for $i = 1, ..., N$:

$$\sum_{j=1}^{N}\Lambda_{j,0}\left(N\lambda[\Phi(k,t)]_{i,j}\right) \leq \sum_{j=1}^{N}\max\left(\Lambda_{j,0}\left(\lambda-\epsilon N\sqrt{N}\lambda\right),\Lambda_{j,0}\left(\lambda+\epsilon N\sqrt{N}\lambda\right)\right),\ \forall t=1,\ldots,s.$$

The remainder of the proof proceeds analogously to the proof of Theorem 5.7. □

5.6 Power allocation

Part b of Corollary 5.8 says that there is a sufficient large deviation rate $\mathcal{J}^\star$ such that the distributed detector is asymptotically optimal; a further increase of $\mathcal{J}$ above $\mathcal{J}^\star$ does not improve the exponential decay rate of the error probability. Also, as we have shown in Subsection 2.5.2 of Chapter 2, the large deviation rate $\mathcal{J}$ is a function of the link occurrence probabilities, which are further dependent on the sensors' transmission power. In summary, Part b of Corollary 5.8 suggests that there is a sufficient (minimal required) transmission power that achieves detection with the optimal exponential decay rate. This observation motivates us to formulate the optimal power allocation problem of minimizing the total transmission power per time k

subject to the optimality condition $\mathcal{J} \geq \mathcal{J}^\star$. Before presenting the optimization problem, we detail the inter-sensor communication model.

Inter-sensor communication model. We adopt a symmetric Rayleigh fading channel model, a model similar to the one proposed in [74] (reference [74] assumes asymmetric channels). At time k, sensor j receives from sensor i:

$$y_{ij,k} = g_{ij,k}\sqrt{\frac{S_{ij}}{d_{ij}^{\alpha}}}x_{i,k} + n_{ij,k},$$

where S_{ij} is the transmission power that sensor i uses for transmission to sensor j, $g_{ij,k}$ is the channel fading coefficient, $n_{ij,k}$ is the zero mean additive Gaussian noise with variance σ_n^2, d_{ij} is the inter-sensor distance, and α is the path loss coefficient. We assume that the channels (i,j) and (j,i) at time k experience the same fade, i.e., $g_{ij,k} = g_{ji,k}$; $g_{ij,k}$ is i.i.d. in time; and $g_{ij,t}$ and $g_{lm,s}$ are mutually independent for all t, s. We adopt the following link failure model. Sensor j successfully decodes the message from sensor i (i.e., the link (i,j) is online) if the signal to noise ratio exceeds a threshold, i.e., if: $\mathrm{SNR} = \frac{S_{ij}g_{ij,k}^2}{\sigma_n^2 d_{ij}^{\alpha}} > \tau$, or, equivalently, if $g_{ij,k}^2 > \frac{\tau\sigma_n^2 d_{ij}^{\alpha}}{S_{ij}} := \frac{K_{ij}}{S_{ij}}$. The quantity $g_{ij,k}^2$ is, for the Rayleigh fading channel, exponentially distributed with parameter 1. Hence, we arrive at the expression for the probability of the link (i,j) being online:

$$P_{ij} = \mathbb{P}\left(g_{ij,k}^2 > \frac{K_{ij}}{S_{ij}}\right) = e^{-\frac{K_{ij}}{S_{ij}}}. \tag{5.67}$$

We constrain the choice of the transmission powers by $S_{ij} = S_{ji}$[7], so that the link (i,j) is online if and only if the link (j,i) is online, i.e., the graph realizations are undirected graphs. Hence, the underlying communication model is the link failure model, with the link occurrence probabilities P_{ij} in (5.67) that are dependent on the transmission powers S_{ij}.

With this model, the large deviation rate $\mathcal{J}$ is given by (2.55), where the weight c_{ij} associated with link (i,j) is:

$$c_{ij}(S_{ij}) = -\log\left(1 - e^{-K_{ij}/S_{ij}}\right).$$

We denote by $\{S_{ij}\}$ the set of all powers S_{ij}, $\{i,j\} \in E$.

Lemma 5.17 The function $\mathcal{J}\left(\{S_{ij}\}\right) = \mathrm{mincut}(V, E, C)$, with $c_{ij} = -\log(1 - e^{-K_{ij}/S_{ij}})$, for $\{i,j\} \in E$, and $c_{ij} = 0$ else, is concave.

[7] We assumed equal noise variances $\sigma_n^2 = \mathrm{Var}(n_{ij,k}) = \mathrm{Var}(n_{ji,k})$ so that $K_{ij} = K_{ji}$, which implies the constraint $S_{ij} = S_{ji}$. Our analysis easily extends to unequal noise variances, in which case we would require $\frac{K_{ij}}{S_{ij}} = \frac{K_{ji}}{S_{ji}}$; this is not considered here.

Proof Note that the function $\mathcal{J}(\{S_{ij}\}) = \mathrm{mincut}(V, E, C)$ can be expressed as

$$\min_{E' \subset E:\, G'=(V,E') \text{ is disconnected}} \sum_{\{i,j\} \in E \setminus E'} c_{ij}(S_{ij}).$$

On the other hand, $c_{ij}(S_{ij})$ is concave in S_{ij} for $S_{ij} \geq 0$, which can be shown by computing the second derivative and noting that it is non-positive. Hence, $\mathcal{J}(\{S_{ij}\})$ is a pointwise minimum of concave functions, and thus it is concave. □

Power allocation problem formulation. We now formulate the power allocation problem as the problem of minimizing the total transmission power used at time k, $2\sum_{\{i,j\} \in E} S_{ij}$, so that the distributed detector achieves asymptotic optimality. This translates into the following optimization problem:

$$\begin{array}{ll} \text{minimize} & \sum_{\{i,j\} \in E} S_{ij} \\ \text{subject to} & \mathcal{J}(\{S_{ij}\}) \geq \mathcal{J}^\star. \end{array} \quad . \tag{5.68}$$

The cost function in (5.68) is linear, and hence convex. Also, the constraint set $\{\{S_{ij}\} :\ \mathcal{J}(\{S_{ij}\}) \geq \mathcal{J}^\star\} = \{\{S_{ij}\} :\ -\mathcal{J}(\{S_{ij}\}) \leq -\mathcal{J}^\star\}$ is convex, as a sub level set of the convex function $-\mathcal{J}(\{S_{ij}\})$. (See Lemma 5.17.) Hence, we have just proved the following lemma.

Lemma 5.18 The optimization problem (5.68) is convex.

Convexity of (5.68) allows us to find a globally optimal solution.

5.6.1 Simulation example

We first describe the simulation setup. We consider a geometric network with $N = 14$ sensors. We place the sensors uniformly over a unit square, and connect those sensors whose distance d_{ij} is less than a radius. The total number of (undirected) links is 38. (These 38 links are the failing links, for which we want to allocate the transmission powers S_{ij}.) We set the coefficients $K_{ij} = 6.25 d_{ij}^{\alpha}$, with $\alpha = 2$. For the averaging weights, we use Metropolis weights, i.e., if link $\{i, j\}$ is online, we assign $W_{ij,k} = 1/(1 + \max\{d_{i,k}, d_{j,k}\})$, where $d_{i,k}$ is the degree of node i at time k and $W_{ij,k} = 0$ otherwise; also, $W_{ii,k} = 1 - \sum_{j \in O_{i,k}} W_{ij,k}$. For the sensors' measurements, we use the Gaussian distribution $f_1 \sim \mathcal{N}(m, \sigma^2)$, $f_0 \sim \mathcal{N}(0, \sigma^2)$, with $\sigma^2 = 1$. For a lower signal-to-noise ratio (SNR) case, we set $m = 0.0447$, and for a higher SNR case, we set $m = 2 \cdot 0.0447$. The corresponding values are $\mathcal{J}^\star = (N-1)N\frac{m^2}{8\sigma^2} = 0.0455$, for a lower SNR, and and $\mathcal{J}^\star = 0.182$, for a higher SNR; see [64].

To obtain the optimized power allocation, we solve the optimization problem (5.68) by applying the

subgradient algorithm with constant stepsize $\beta = 0.0001$ on the unconstrained exact penalty reformulation of (5.68), see, e.g., [72], which is to minimize $\sum_{\{i,j\}\in E} S_{ij} + \mu \max\{0, -\text{mincut}(V, E, C) + \mathcal{J}^\star\}$, where $C = [c_{ij}]$, $c_{ij} = -\log(1 - e^{-K_{ij}/S_{ij}})$, for $\{i, j\} \in E$, and zero else; and μ is the penalty parameter that we set to $\mu = 500$. We used the MATLAB implementation [91] of the min-cut algorithm from [73]. Note that the resulting power allocation is optimal over the class of *deterministic* power allocations, i.e., the power allocations that: 1) use the same total power across all links per each time step; and 2) use *deterministic* power assignment policy at each time step.

Results. Figure 5.5 (left) plots the detection error probability for a lower SNR case, of the worst sensor $\max_{i=1,\dots,N} P^{\text{e}}_{i,k}$ versus time k. We compare: 1) the optimized power allocation $\{S^\star_{ij}\}$ (solid blue line); 2) the uniform power allocation $S_{ij} = S$ across all links, such that the total power per k over all links $2\sum_{\{i,j\}\in E} S_{ij} = 2\sum_{\{i,j\}\in E} S^\star_{ij} =: \mathcal{S}$; and 3) a *random*, gossip like, power allocation, where, at a time step k, only one out of all links is activated (uniformly across all links) such that the power $\mathcal{S}$ is invested in it (half of $\mathcal{S}$ in each direction of the communication.) Note that this allocation is *random*, hence outside of the class that we optimize over. The optimized power allocation significantly outperforms the uniform power allocation. For example, to achieve the error probability 0.1, the optimized power allocation scheme requires about 550 time steps, hence the total consumed power is $550\mathcal{S}$; in contrast, the uniform power allocation needs more than $2000\mathcal{S}$ for the same target error 0.1. In addition, Figure 5.5 plots the detection performance for the uniform power allocation with the total power per k equal to $\text{sr} \times 3\mathcal{S}$. This scheme takes more than 700 time steps to achieve an error of 0.1, hence requiring the total power of $700 \times 3 \times \mathcal{S} = 2100\mathcal{S}$ to achieve an error of 0.1. Further, we can see that, for a lower SNR case, the random, gossip policy achieves – exactly as the optimized policy – the best detection error exponent $\mathcal{D}$. (Note that the two corresponding lines are parallel.) This is not a contradiction as the random policy is outside of the class of deterministic allocations that we optimize over. Furthermore, the randomized gossip policy is slightly better than the optimized policy (It has a better constant C in the detection error $P^{\text{e}}_k \approx Ce^{-k\mathcal{D}}$). However, for a larger SNR (Figure 5.5, right), the gossip policy no longer achieves the optimal slope $\mathcal{D}$, and the optimized policy becomes better. In particular, for the 10^{-1} detection error, the optimized policy saves about 50 time steps (from 200 to 150), with respect to gossip, hence saving 25% of total required power.

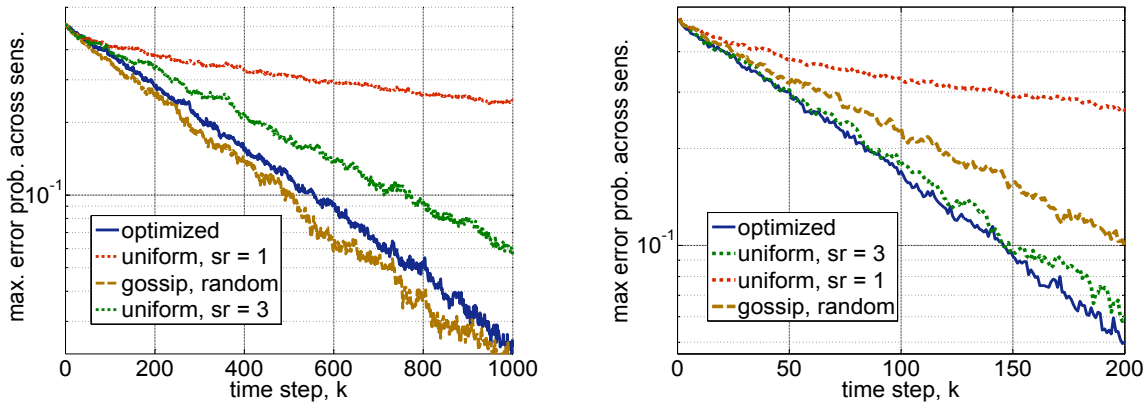

Figure 5.5: Detection error probability of the worst sensor versus time step k for the optimized power allocation, the uniform power allocation with $\mathrm{sr} = 1, 3$, ($\mathrm{sr} = \frac{\text{total power per k for uniform allocation}}{\text{total power per k for optimal allocation}}$,) and for the random, gossip allocation; **Left:** lower SNR; **Right:** higher SNR.

Chapter 6

Conclusion

This thesis analyzed the large deviations performance of consensus-based distributed algorithms for inference (detection, estimation) in networks. Consensus-based distributed algorithms have recently attracted much attention in the literature, in the context of distributed estimation and detection in sensor networks [2], modeling swarm behavior of robots/animals [3, 4], detection of a primary user in cognitive radio networks [5], and power grid state estimation [6].

In contrast with existing literature that usually adopts asymptotic normality or asymptotic consistency metrics, the metric that we adopt are the rates of large deviations. This enables us to quantify an interesting interplay between the underlying random network parameters and the distributions of the agents' observations.

We recapitulate chapter-by-chapter contributions of this thesis.

Chapter 2: Products of random stochastic matrices: The symmetric i.i.d. case

We consider a sequence of i.i.d., symmetric, stochastic matrices $\{W_k\}$ with positive diagonal entries. We characterize the large deviation limit: $\mathcal{J} := \lim_{k\to\infty} -\frac{1}{k}\log\mathbb{P}\left(\|W_k\cdots W_2W_1 - J\| \geq \epsilon\right),\ \epsilon \in (0,1]$, The quantity $\mathcal{J}$ has not been computed in the literature before. We show that $\mathcal{J}$ is solely a function of the graphs induced by the matrices W_k and the corresponding probabilities of occurrences of these graphs, and we show that $\mathcal{J}$ does not depend on $\epsilon \in (0,1]$.

Computation of the quantity $\mathcal{J}$ is in general a hard problem. However, for commonly used gossip and link failure models, we show that $\mathcal{J} = |\log(1-c)|$, where c is the min-cut value (or connectivity [16]) of a graph whose links are weighted by the gossip link probabilities. Similarly, we show that $\mathcal{J}$ is also computed via min-cut for *link failures on general graphs*. Finally, we find tight approximations for $\mathcal{J}$ for a

symmetrized broadcast gossip and similar averaging models.

Besides distributed inference algorithms, the result on the characterization of $\mathcal{J}$ is of independent interest in the theory of products of stochastic matrices [12, 13], non-homogenous Markov chains [14], and consensus algorithms, e.g., [15].

Chapter 3: Products of Random Stochastic Matrices: Temporal Dependencies and Directed Networks

We go beyond symmetric i.i.d. matrices from Chapter 1 by studying: 1) temporally dependent, symmetric matrices W_k; and 2) temporally i.i.d., asymmetric (not necessarily doubly stochastic) matrices W_k.

Our temporally dependent model of the W_k's associates a state of a Markov chain to each of the possible realizations G_t of graphs that supports W_t. The distribution of the graphs G_t, $t \geq 1$, is determined by a $\mathcal{M} \times \mathcal{M}$ transition probability matrix P, where $\mathcal{M}$ is the number of possible realizations of G_t. We characterize the rate $\mathcal{J}$ as a function of the transition probability matrix P. We show that the rate $\mathcal{J}$ is determined by the most likely way in which the Markov chain stays in a subset of states (graphs) whose union is disconnected. The probability of this event is determined by the spectral radius of the block in the transition matrix P that corresponds to this most likely subset of states, and this spectral radius determines the rate $\mathcal{J}$.

We study temporally i.i.d. asymmetric matrices W_k, and we characterize the following large deviation limit: $\mathcal{J}_{\text{dir}} = \lim_{k\to+\infty} -\frac{1}{k} \log \mathbb{P}\left(|\lambda_2(W_k \cdots W_1)| \geq \epsilon\right), \ \epsilon \in (0,1]$, which is a natural generalization of the quantity $\mathcal{J}$ to directed networks. We show that the limit $\mathcal{J}_{\text{dir}}$ depends on the distribution of matrices only through the support graphs: $\mathcal{J}_{\text{dir}}$ is determined by the probability of the most likely set of support graphs whose union does not contain a directed spanning tree. We illustrate our results on a commonly used broadcast gossip protocol, where (only one) node u activates at a time with probability p_u, and broadcasts its state to all single-hop neighbors. We show that the rate $\mathcal{J}_{\text{dir}} = |\log 1 - p_{\min}|$, where $p_{\min}$ is the probability of activation of the node that activates least frequently.

Chapter 4: Large deviations for distributed inference

We consider linear distributed inference algorithms with vector innovations $Z_{i,k}$ and generic distributions. We study the large deviation rates $I(E)$ for generic sets $E \subset \mathbb{R}^d$. For spatio-temporally i.i.d. observations and asymmetric matrices W_k, we show that performance $I(E)$ of distributed inference is at least as good as the performance of the performance of isolated inference. Likewise, distributed inference is always worse, or at best equal, to the performance of the centralized, ideal inference. Although very intuitive, the result is challenging to prove, and requires modification of the arguments of the Gartner-Ellis Theorem. When

the W_k's are *symmetric*, distributed inference guarantees much larger gains over isolated inference. The results reveal a very interesting interplay between the underlying network and the distribution of the agents' innovations. Distributed inferenceis close to the centralized performance for very high required accuracies, but it can become much worse from the centralized performance for very coarse precisions. Further, for regular networks, we establish the full large deviations principle for distributed inference. Finally, for *spatially different innovations* and symmetric W_k's, we show that the relative performance of distributed inference over the centralized inference is still a highly nonlinear function of the target accuracy.

Chapter 5: Distributed detection

We establish the large deviations performance of distributed detection algorithm in [1], hence establishing a counterpart result to the (centralized detection's) Chernoff lemma [8]. We show that distributed detection exhibits a phase transition behavior with respect to the large deviations rate of consensus $\mathcal{J}$ (network connectivity). If $\mathcal{J}$ is above a threshold, then distributed detector' error exponent equals the Chernoff information–the best possible error exponent of the optimal centralized detector. For $\mathcal{J}$ below the threshold, we quantify the achievable fraction of the centralized detector's performance with the distributed detector. We discover a very interesting interplay between the distribution of the agents' measurements (e.g., Gaussian or Laplace) and the network connectivity (the value of $\mathcal{J}$). For example, for the same network connectivity (same $\mathcal{J}$), a distributed detector with Laplace observations may achieve the optimal asymptotic performance, while the distributed detector for Gaussian observations may be suboptimal, even though the corresponding centralized detectors are asymptotically equivalent. Finally, we address the problem of allocating the agents' transmission powers such that they achieve asymptotically optimal detection while minimizing the invested transmission power.

Bibliography

[1] P. Braca, S. Marano, V. Matta, and P. Willet, "Asymptotic optimality of running consensus in testing binary hypothesis," *IEEE Transactions on Signal Processing*, vol. 58, pp. 814–825, February 2010.

[2] S. Kar, J. M. F. Moura, and K. Ramanan, "Distributed parameter estimation in sensor networks: Nonlinear observation models and imperfect communication," *Accepted for publication in IEEE Transactions on Information Theory, 51 pages*, August 2008.

[3] J. Li and A. H. Sayed, "Modeling bee swarming behavior through diffusion adaptation with asymmetric information sharing," *EURASIP Journal on Advances in Signal Processing*, Jan. 2012. doi:10.1186/1687-6180-2012-18.

[4] S.-Y. Tu and A. H. Sayed, "Mobile adaptive networks," *IEEE J. Selected Topics on Signal Processing*, vol. 5, pp. 649–664, Aug. 2011.

[5] F. Cattivelli and A. Sayed, "Distributed detection over adaptive networks using diffusion adaptation," *IEEE Transactions on Signal Processing*, vol. 59, pp. 1917–1932, May 2011.

[6] L. Xie, D. Choi, S. Kar, and H. Poor, "Fully distributed state estimation for wide-area monitoring systems," *to appear in the IEEE Transactions on Smart Grid*, May 2011.

[7] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: John Wiley and Sons, 1991.

[8] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*. Boston, MA: Jones and Barlett, 1993.

[9] M. Arcones, "Large deviations for m-estimators," *Annals of the Institute of Statistical Mathematics*, vol. 58, no. 1, pp. 21–52, 2006.

[10] D. Bajović, J. Xavier, J. M. F. Moura, and B. Sinopoli, "Consensus and products of random stochastic matrices: Exact rate for convergence in probability," *IEEE Transactions on Signal Processing*, vol. 61, pp. 2557–2571, May 2013.

[11] D. Bajović, J. Xavier, J. M. F. Moura, and B. Sinopoli, "Exact rate for convergence in probability of averaging processes via generalized min-cut," in *CDC '12, 51st IEEE Conference on Decision and Control*, (Hawaii, USA), December 2012.

[12] A. Leizarowitz, "On infinite products of stochastic matrices," *Linear Algebra and its Applications*, vol. 168, pp. 189–219, April 1992.

[13] L. Bruneau, A. Joye, and M. Merkli, "Infinite products of random matrices and repeated interaction dynamics," *Annales de l'Institut Henri Poincar, Probabilits et Statistiques*, vol. 46, no. 2, pp. 442–464, 2010.

[14] E. Seneta, *Nonnegative Matrices and Markov Chains*. New York: Springer, 1981.

[15] R. Olfati-Saber and R. M. Murray, "Consensus problems in networks of agents with switching topology and time-delays," *IEEE Transactions on Automatic Control*, vol. 49, pp. 1520–1533, Sept. 2004.

[16] R. D. Carr, G. Konjevod, G. Little, V. Natarajan, and O. Parekh, "Compacting cuts: a new linear formulation for minimum cut," *ACM Transactions on Algorithms*, vol. 5, July 2009. DOI: 10.1145/1541885.1541888.

[17] B. Johansson, M. Rabi, and M. Johansson, "A randomized incremental subgradient method for distributed optimization in networked systems," *SIAM Journal on Optimization*, vol. 20, no. 3, pp. 1157–1170, 2009.

[18] F. d. Hollander, *Large deviations*. Fields Institute Monographs, American Mathematical Society, 2000.

[19] T. C. Aysal, M. E. Yildiz, A. D. Sarwate, and A. Scaglione, "Broadcast gossip algorithms for consensus," *IEEE Transactions on Signal Processing*, vol. 57, pp. 2748–2761, July 2009.

[20] J. N. Tsitsiklis, *Problems in decentralized decision making and computation*. Ph.d., MIT, Cambridge, MA, 1984.

[21] M. H. DeGroot, "Reaching a consensus," *Journal of the American Statistical Association*, vol. 69, pp. 118–121, 1974.

[22] A. Jadbabaie, J. Lin, and A. S. Morse, "Coordination of groups of mobile autonomous agents using nearest neighbor rules," *IEEE Trans. Automat. Contr*, vol. AC-48, pp. 988–1001, June 2003.

[23] S. Boyd, A. Ghosh, B. Prabhakar, and D. Shah, "Randomized gossip algorithms," *IEEE Transactions on Information Theory*, vol. 52, pp. 2508–2530, June 2006.

[24] A. Dimakis, A. Sarwate, and M. Wainwright, "Geographic gossip: Efficient averaging for sensor networks," *IEEE Transactions on Signal Processing*, vol. 56, no. 3, pp. 1205–1216, 2008.

[25] D. Üstebay, B. Oreshkin, M. Coates, and M. Rabbat, "Greedy gossip with eavesdropping," *IEEE Transactions on Signal Processing*, vol. 58, no. 7, pp. 3765–3776, 2010.

[26] A. G. Dimakis, S. Kar, J. M. F. Moura, M. G. Rabbat, and A. Scaglione, "Gossip algorithms for distributed signal processing," *Proceedings of the IEEE*, vol. 98, pp. 1847–1864, November 2010. Digital Object Identifier: 10.1109/JPROC.2010.2052531.

[27] A. Tahbaz-Salehi and A. Jadbabaie, "Consensus over ergodic stationary graph processes," *IEEE Transactions on Automatic Control*, vol. 55, pp. 225–230, January 2010.

[28] A. Nedić and A. Ozdaglar, "Convergence rate for consensus with delays," *Journal of Global Optimization*, vol. 47, no. 3, pp. 437–456, 2008.

[29] A. Nedić, A. Olshevsky, A. Ozdaglar, and J. N. Tsitsiklis, "On distributed averaging algorithms and quantization effects," *IEEE Transactions on Automatic Control*, vol. 54, no. 11, pp. 2506–2517, 2009.

[30] Y. Mo and B. Sinopoli, "Communication complexity and energy efficient consensus algorithm," in *2nd IFAC Workshop on Distributed Estimation and Control in Networked Systems*, (France), Sep. 2010. DOI: 10.3182/20100913-2-FR-4014.00057.

[31] A. Olshevsky and J. N. Tsitsiklis, "Convergence speed in distributed consensus and averaging," *SIAM Rev.*, vol. 53, pp. 747–772, November 2011.

[32] A. Tahbaz-Salehi and A. Jadbabaie, "On consensus over random networks," in *44th Annual Allerton Conference on Communication, Control, and Computing*, (Monticello, Il), pp. 1315–1321, September 2006.

[33] F. Fagnani and S. Zampieri, "Randomized consensus algorithms over large scale networks," *IEEE Journal on Selected Areas in Communications*, vol. 26, pp. 634–649, May 2008.

[34] B. Touri and A. Nedić, "Product of random stochastic matrices," *Submitted to the Annals of Probability*, 2011. available at: http://arxiv.org/pdf/1110.1751v1.pdf.

[35] B. Touri and A. Nedić, "On backward product of stochastic matrices," 2011. available at: http://arxiv.org/abs/1102.0244.

[36] P. Diaconis and P. M. Wood, "Random doubly stochastic tridiagonal matrices," 2011. available at: http://stat.stanford.edu/ cgates/PERSI/papers/TriDiag1.pdf.

[37] V. N. Tutubalin, "On limit theorems for products of random matrices," *Theory of Probability and its Applications*, vol. 10, pp. 15–27, 1965.

[38] Y. Guivarc'h and A. Raugi, "Products of random matrices: convergence theorems," *Contemporary mathematics*, vol. 50, pp. 31–54, 1986.

[39] É. Le Page, "Théorèmes limites pour les produits de matrices aléatoires," *Probability Measures on Groups (Oberwolfach, 1981), Lecture Notes in Mathematics*, vol. 928, pp. 258–303, 1982.

[40] H. Hennion, "Limit theorems for products of positive random matrices," *The Annals of Probability*, vol. 25, no. 4, pp. 1545–1587, 1997.

[41] V. Kargin, "Products of random matrices: dimension and growth in norm," *The Annals of Probability*, vol. 20, no. 3, pp. 890–906, 2010.

[42] S. A. Aldosari and J. M. F. Moura, "Detection in sensor networks: the saddlepoint approximation," *IEEE Transactions on Signal Processing*, vol. 55, pp. 327–340, January 2007.

[43] R. Viswanatan and P. R. Varshney, "Decentralized detection with multiple sensors: Part I–fundamentals," *Proc. IEEE*, vol. 85, pp. 54–63, January 1997.

[44] R. S. Blum, S. A. Kassam, and H. V. Poor, "Decentralized detection with multiple sensors: Part II–advanced topics," *Proc. IEEE*, vol. 85, pp. 64–79, January 1997.

[45] J. F. Chamberland and V. V. Veeravalli, "Wireless sensors in distributed detection applications," *IEEE Signal Processing Magazine*, vol. 24, pp. 16–25, May 2007.

[46] J. F. Chamberland and V. Veeravalli, "Decentralized dectection in sensor networks," *IEEE Transactions on Signal Processing*, vol. 51, pp. 407–416, February 2003.

[47] J. F. Chamberland and V. Veeravalli, "Asymptotic results for decentralized detection in power constrained wireless sensor networks," *IEEE Journal on Selected Areas in Communications*, vol. 22, pp. 1007–1015, August 2004.

[48] D. Bajović, B. Sinopoli, and J. Xavier, "Sensor selection for event detection in wireless sensor networks," *IEEE Transactions on Signal Processing*, vol. 59, pp. 4938–4953, Oct. 2011.

[49] S. Kar, S. A. Aldosari, and J. M. F. Moura, "Topology for distributed inference on graphs," *IEEE Transactions on Signal Processing*, vol. 56, pp. 2609–2613, June 2008.

[50] M. Alanyali, S. Venkatesh, O. Savas, and S. Aeron, "Distributed detection in sensor networks with packet losses and finite capacity links," *IEEE Transactions on Signal Processing*, vol. 54, pp. 4118–4132, July 2004.

[51] S. A. Aldosari and J. M. F. Moura, "Topology of sensor networks in distributed detection," in *ICASSP'06, IEEE International Conference on Signal Processing*, vol. 5, (Toulouse, France), pp. 1061 – 1064, May 2006.

[52] S. Kar, S. Aldosari, and J. M. F. Moura, "Topology for distributed inference on graphs," *IEEE Transactions on Signal Processing*, vol. 56, pp. 2609–2613, June 2008.

[53] C. G. Lopes and A. H. Sayed, "Diffusion least-mean squares over adaptive networks: formulation and performance analysis," *IEEE Transactions on Signal Processing*, vol. 56, pp. 3122–3136, July 2008.

[54] I. D. Schizas, G. Mateos, and G. B. Giannakis, "Distributed LMS for consensus-based in-network adaptive processing," *IEEE Trans. on Signal Processing*, vol. 57, pp. 2365–2381, June 2009.

[55] G. Mateos, I. D. Schizas, and G. B. Giannakis, "Distributed recursive least-squares for consensus-based in-network adaptive estimation," *IEEE Trans. on Signal Processing*, vol. 57, pp. 4583–4588, November 2009.

[56] F. S. Cattivelli and A. H. Sayed, "Diffusion LMS strategies for distributed estimation," *IEEE Transactions on Signal Processing*, vol. 58, pp. 1035–1048, March 2010.

[57] F. S. Cattivelli and A. H. Sayed, "Distributed detection over adaptive networks based on diffusion estimation schemes," in *Proc. IEEE SPAWC '09, 10th IEEE International Workshop on Signal Processing Advances in Wireless Communications*, (Perugia, Italy), pp. 61–65, June 2009.

[58] F. S. Cattivelli and A. H. Sayed, "Diffusion LMS-based detection over adaptive networks," in *Proc. Asilomar Conf. Signals, Systems and Computers*, (Pacific Grove, CA), pp. 171–175, October 2009.

[59] S. Kar, R. Tandon, H. V. Poor, and S. Cui, "Distributed detection in noisy sensor networks," in *Proc. ISIT 2011, International Symposium on Information Theory*, (Saint Petersburgh, Russia), pp. 2856–2860, August 2011.

[60] S. S. Stanković, N. Ilić, M. S. Stanković, and K. H. Johansson, "Distributed change detection based on a consensus algorithm," *to appear in the IEEE Transactions on Signal Processing*, Digital Object Identifier: 10.1109/TSP.2011.2168219 2011.

[61] S. S. Stanković, N. Ilić, M. S. Stanković, and K. H. Johansson, "Distributed change detection based on a randomized consensus algorithm," in *ECCSC '10, 5th European Conference on Circuits and Systems for Communications*, (Belgrade, Serbia), pp. 51–54, March 2010.

[62] D. Bajović, D. Jakovetić, J. Xavier, B. Sinopoli, and J. M. F. Moura, "Distributed detection over time varying networks: large deviations analysis," in *48th Allerton Conference on Communication, Control, and Computing*, (Monticello, IL), pp. 302–309, Oct. 2010.

[63] D. Bajović, D. Jakovetić, J. Xavier, B. Sinopoli, and J. M. F. Moura, "Distributed detection via Gaussian running consensus: Large deviations asymptotic analysis," *IEEE Transactions on Signal Processing*, vol. 59, pp. 4381–4396, Sep. 2011.

[64] D. Bajović, D. Jakovetić, J. M. F. Moura, J. Xavier, and B. Sinopoli, "Large deviations performance of consensus+innovations distributed detection with non-Gaussian observations," *IEEE Transactions on Signal Processing*, vol. 60, pp. 5987–6002, Nov. 2012.

[65] D. Jakovetić, J. M. F. Moura, and J. Xavier, "Distributed detection over noisy networks: large deviations analysis." Submitted for publication, August 2011.

[66] S. A. Kassam, *Signal Detection in Non-Gaussian Noise*. New York: Springer-Verlag, 1987.

[67] H. Furstenberg and H. Kesten, "Products of random matrices," *Annals of mathematical statistics*, vol. 31, pp. 457–469, June 1960.

[68] P. Denantes, F. Benezit, P. Thiran, and M. Vetterli, "Which distributed averaging algorithm should I choose for my sensor network?," in *INFOCOM 2008, The 27th IEEE Conference on Computer Communications*, (Phoenix, Arizona), pp. 986–994, March 2008.

[69] A. F. Karr, *Probability*. New York: Springer-Verlag, 1993.

[70] M. Fiedler, "Algebraic connectivity of graphs," *Czechoslovak Mathematical Journal*, vol. 23, no. 98, pp. 298–305, 1973.

[71] A. Nedić and A. Ozdaglar, "Distributed subgradient methods for multi-agent optimization," *IEEE Transactions on Automatic Control*, vol. 54, no. 1, pp. 48–61, 2009.

[72] J.-B. Hiriart-Urruty and C. Lemarechal, *Fundamentals of Convex Analysis*. Grundlehren Text Editions, Berlin, Germany: Springer-Verlag, 2004.

[73] M. Stoer and F. Wagner, "A simple min-cut algorithm," *Journal of the ACM*, vol. 44, pp. 585–591, July 1997.

[74] K. Chan, A. Swami, Q. Zhao, and A. Scaglione, "Consensus algorithms over fading channels," in *Proc. MILCOM 2010, Military communications conference*, (San Jose, CA), pp. 549–554, October 2010.

[75] I. Matei, N. Martins, and J. S. Baras, "Almost sure convergence to consensus in markovian random graphs," in *CDC 2008. 47th IEEE Conference on Decision and Control*, (Cancun, Mexico), Dec. 2008.

[76] D. Bajović, J. Xavier, , and B. Sinopoli, "Products of stochastic matrices: large deviation rate for markov chain temporal dependencies," in *Allerton'12, 50th Allerton Conference on Communication, Control, and Computing*, (Monticello, Il), October 2012.

[77] D. Bajović, J. Xavier, , and B. Sinopoli, "Products of stochastic matrices: exact rate for convergence in probability for directed networks," in *TELFOR '12, 20th Telecommunications Forum*, (Belgrade, Serbia), November 2012.

[78] R. A. Horn and C. R. Johnson, *Matrix analysis*. Cambrige, United Kingdom: Cambrige Univesity Press, 1990.

[79] L. Xiao and S. Boyd, "Fast linear iterations for distributed averaging," *Syst. Contr. Lett.*, vol. 53, pp. 65–78, September 2004.

[80] A. Berman and R. J. Plemmons, *Nonnegative Matrices in the Mathematical Sciences*. New York, NY: Academic Press, INC., 1970.

[81] I. C. F. Ipsen and T. M. Selee, "Ergodicity coefficients defined by vector norms," *SIAM Journal on Matrix Analysis and Applications*, vol. 32, pp. 153–200, March 2011.

[82] C. W. Wu, "Synchronization and convergence of linear dynamics in random directed networks," *IEEE Transactions on Automatic Control*, vol. 51, pp. 1207–1210, July 2006.

[83] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambrige, United Kingdom: Cambridge University Press, 2004.

[84] J. M. Borwein and A. S. Lewis, *Convex Analysis and Nonlinear Optimization: Theory and Examples*. CMS Books in Mathematics, Canadian Mathematical Society, Springer, second ed., 2006.

[85] M. Sion, "On general minimax theorems," *Pacific Journal of Mathematics*, vol. 8, pp. 171–176, March 1958.

[86] D. Bajović, D. Jakovetić, J. M. F. Moura, J. Xavier, , and B. Sinopoli, "Large deviations analysis of consensus+innovations detection in random networks," in *Allerton'11, 49th Allerton Conference on Communication, Control, and Computing*, (Monticello, Il), October 2011.

[87] D. Bajović, J. Xavier, , and B. Sinopoli, "Robust linear dimensionality reduction for hypothesis testing with application to sensor selection," in *Allerton'09, 47th Allerton Conference on Communication, Control, and Computing*, (Monticello, Il), October 2009.

[88] D. Bajović, B. Sinopoli, and J. Xavier, "Sensor selection for hypothesis testing in wireless sensor networks: a Kullback-Leibler based approach," in *CDC '09, 48th IEEE Conference on Decision and Control*, (Shanghai, China), December 2009.

[89] J. Fang, H. Li, Z. Chen, and S. Li, "Optimal precoding design and power allocation for decentralized detection of deterministic signals," *IEEE Transactions on Signal Processing*, vol. 60, pp. 3149–3163, June 2012.

[90] M. J. Wainwright and M. I. Jordan, "Graphical models, exponential families, and variational inference," *Foundations and Trends in Machine Learning*, vol. 1, pp. 4938–4953, Oct. 2008.

[91] Y. Devir, "Matlab m-file for the min-cut algorithm," 2006. available at: http://www.mathworks.com/matlabcentral/fileexchange/13892-a-simple-min-cut-algorithm.