Diffusion-Based Adaptive Distributed Detection: Steady-State Performance in the Slow Adaptation Regime

Vincenzo Matta, Paolo Braca, Stefano Marano, Ali H. Sayed

***Abstract*—This work examines the close interplay between cooperation and adaptation for distributed detection schemes over fully decentralized networks. The combined attributes of cooperation and adaptation are necessary to enable networks of detectors to continually learn from streaming data and to continually track drifts in the state of nature when deciding in favor of one hypothesis or another. The results in the paper establish a fundamental scaling law for the steady-state probabilities of miss-detection and false-alarm in the slow adaptation regime, when the agents interact with each other according to distributed strategies that employ small constant step-sizes. The latter are critical to enable continuous adaptation and learning. The work establishes three key results. First, it is shown that the output of the collaborative process at each agent has a steady-state distribution. Second, it is shown that this distribution is asymptotically Gaussian in the slow adaptation regime of small step-sizes. And third, by carrying out a detailed large deviations analysis, closed-form expressions are derived for the decaying rates of the false-alarm and miss-detection probabilities. Interesting insights are gained from these expressions. In particular, it is verified that as the step-size μ decreases, the error probabilities are driven to zero exponentially fast as functions of $1/\mu$, and that the exponents governing the decay increase linearly in the number of agents. It is also verified that the scaling laws governing errors of detection and errors of estimation over networks behave very differently, with the former having an exponential decay proportional to $1/\mu$, while the latter scales linearly with decay proportional to μ. Moreover, and interestingly, it is shown that the cooperative strategy allows each agent to reach the same detection performance, in terms of detection error exponents, of a centralized stochastic-gradient solution. The results of the paper are illustrated by applying them to canonical distributed detection problems.**

***Index Terms*—Distributed detection, adaptive network, diffusion strategy, consensus strategy, false-alarm probability, miss-detection probability, large deviations analysis.**

I. Overview

RECENT advances in the field of distributed inference have produced several useful strategies aimed at exploiting local *cooperation* among network nodes to enhance the performance of each individual agent. However, the increasing availability of streaming data continuously flowing across the network has added the new and challenging requirement of online *adaptation* to track drifts in the data. In the adaptive mode of operation, the network agents must be able to enhance their learning abilities continually in order to produce reliable inference in the presence of drifting statistical conditions, drifting environmental conditions, and even changes in the network topology, among other possibilities. Therefore, concurrent adaptation (i.e., tracking) and learning (i.e., inference) are key components for the successful operation of distributed networks tasked to produce reliable inference under dynamically varying conditions and in response to streaming data.

Several useful distributed implementations based on consensus strategies [1]–[12] and diffusion strategies [13]–[18] have been developed for this purpose in the literature. The diffusion strategies have been shown to have superior stability ranges and mean-square performance when constant step-sizes are used to enable continuous adaptation and learning [19]. For example, while consensus strategies can lead to unstable growth in the state of adaptive networks even when all agents are individually stable, this behavior does not occur for diffusion strategies. In addition, diffusion schemes are robust, scalable, and fully decentralized. Since in this work we focus on studying *adaptive* distributed inference strategies, we shall therefore focus on diffusion schemes due to their enhanced mean-square stability properties over adaptive networks.

Now, the interplay between the two fundamental aspects of cooperation and adaptation has been investigated rather extensively in the context of *estimation* problems. Less explored in the literature is the same interplay in the context of *detection* problems. This is the main theme of the present work. Specifically, we shall address the problem of designing and characterizing the performance of diffusion strategies that reconcile both needs of adaptation and detection in decentralized systems. The following is a brief description of the scenario of interest.

A network of connected agents is assumed to monitor a certain phenomenon of interest. As time elapses, the agents collect an increasing amount of streaming data, whose statistical properties depend upon an *unknown* state of nature. The state is formally represented by a pair of hypotheses, say, $\mathcal{H}_0$ and $\mathcal{H}_1$. At each time instant, each agent is expected to produce a decision about the state of nature, based upon its own observations and the exchange of information with neighboring agents. The emphasis here is on *adaptation*: we allow the true hypothesis to drift over time, and the network must be

The work of A. H. Sayed was supported in part by NSF grants CCF-1011918, CCF-1524250, and ECCS-1407712. A short and limited version of this work appears in the conference publication [30].

V. Matta and S. Marano are with DIEM, University of Salerno, via Giovanni Paolo II 132, I-84084, Fisciano (SA), Italy (e-mail: vmatta@unisa.it; marano@unisa.it).

P. Braca is with NATO STO Centre for Maritime Research and Experimentation, La Spezia, Italy (e-mail: braca@cmre.nato.int).

A. H. Sayed is with the Electrical Engineering Department, University of California, Los Angeles, CA 90095 USA (e-mail: sayed@ee.ucla.edu).

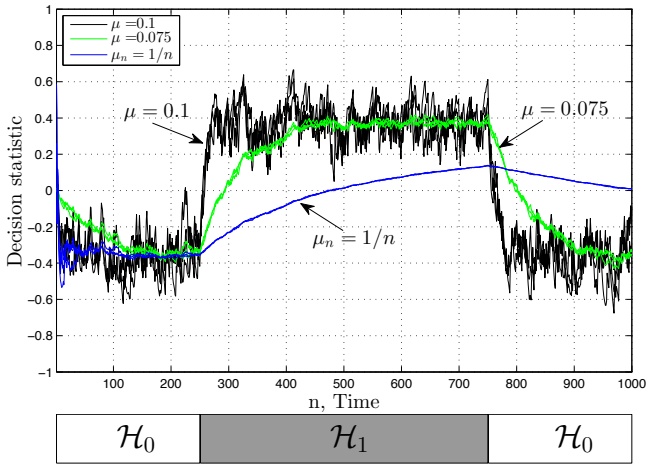

Fig. 1. The top panel illustrates the time-evolution of the decision statistics at three generic local agents for two situations: (a) constant step-size adaptation using a diffusion strategy and (b) diminishing step-size updates using $\mu_n = 1/n$ and a running consensus strategy. The actual variation of the true hypothesis is depicted in the bottom panel from $\mathcal{H}_0$ to $\mathcal{H}_1$ to $\mathcal{H}_0$.

able to track the drifting state. This framework is illustrated in Fig. 1, where we show the time-evolution of the actual realization of the decision statistics computed by three generic network agents. Two situations are considered. In the first case, the agents run a constant-step size diffusion strategy [15], [20] and in the second case, the agents run a consensus strategy with a diminishing step-size of the form $\mu_n = 1/n$ [1]–[6]. Note from the curves in the figure that the statistics computed by different sensors are hardly distinguishable, emphasizing a certain equivalence in performance among distinct agents, an important feature that will be extensively commented on in the forthcoming analysis.

Assume that high (positive) values of the statistic correspond to deciding for $\mathcal{H}_1$, while low (negative) values correspond to deciding for $\mathcal{H}_0$. The bottom panel in the figure shows how the true (unknown) hypothesis changes at certain (unknown) epochs following the sequence $\mathcal{H}_0 \rightarrow \mathcal{H}_1 \rightarrow \mathcal{H}_0$. It is seen in the figure that the adaptive diffusion strategy is more apt in tracking the drifting state of nature. It is also seen that the diminishing step-size consensus implementation is unable to track the changing conditions. Moreover, the inability to track the drift degrades further as time progresses since the step-size sequence $\mu_n = 1/n$ decays to zero as $n \rightarrow \infty$. For this reason, in this work we shall set the step-sizes to constant values to enable continuous adaptation and learning by the distributed network of detectors. In order to evaluate how well these adaptive networks perform, we need to be able to assess the goodness of the inference performance (reliability of the decisions), so as to exploit the trade-off between adaptation and learning capabilities. This will be the main focus of the paper.

A. Related Work

The literature on distributed detection is definitely rich, see, e.g., [21]–[28] as useful entry points on the topic. A distinguishing feature of our approach is its emphasis on *adaptive* distributed detection techniques that respond to streaming data in real-time. We address this challenging problem with reference to the *fully decentralized* setting, where no fusion center is admitted, and the agents cooperate through local interaction and consultation steps.

For several useful formulations of distributed point estimation and detection, the use of stochastic approximation consensus-based solutions with *diminishing* step-sizes leads to asymptotically optimal performance, either in the sense of asymptotic variance in point estimation [12], in the sense of error exponents [4]–[6], or in the sense of asymptotic relative efficiency in the locally optimum detection framework [2]. Optimality in these works is formulated in reference to the centralized solution, and the qualification "asymptotic" is used to refer either to a large number of observations or a large time window. The error performance (e.g., mean-square error for estimation or error probabilities for detection) is shown in these works to decay with optimal rates as time elapses, provided that some conditions on the network structure are met. For these results to hold, it is critical for the statistical properties of the data to remain invariant and for the algorithms to rely on a recursive test statistics with a *diminishing* step-size.

In some other distributed inference applications, however, the statistical properties of the data can vary over time. For instance, in a detection problem, the actual hypothesis in force, and/or some parameters of the pertinent distributions, might change at certain moments. Therefore, the adaptation aspect, i.e., the capability of persistently tracking dynamic scenarios, becomes important. In such scenarios, the diffusion algorithms (with non-diminshing, constant step-size) provide effective mechanisms for continuous adaptation and learning. Similar to the consensus-based algorithms with diminishing step-sizes, they are easy to implement, since they involve linear operations, and are naturally suited to a fully distributed implementation. However, differently from the consensus algorithms with diminishing step-size, the strategies with constant step-size are inherently able to work under dynamically changing conditions and offer enhanced tracking capability.

B. Inherent Tracking Mechanism

It is well-known in the adaptation and learning literature that using *constant step-sizes* in the update relations automatically infuses the algorithms with a tracking mechanism that enables them to track variations in the underlying models. This is because constant step-sizes keep adaptation alive, forever. This is in contrast to decaying step-sizes, which tend to zero and ultimately stop adapting. With a constant step-size, learning is always active. When the hypothesis changes, an algorithm with a constant step-size will continue learning from that point onwards and given sufficient time to learn, the steady-state analysis in this article will show that the probabilities of error will indeed decay exponentially as functions of the inverse of the step-size.

The key challenge in these scenarios is that a constant step-size keeps the update active, which then causes gradient noise to seep continuously into the operation of the algorithm. This

effect does not happen for decaying step-sizes because the diminishing step-size annihilates the gradient noise term in the limit. However, a decaying step-size cannot track changing hypotheses due to the vanishing step-size. The difficulty in the constant step-size case is therefore to show that despite the presence of gradient noise, the dynamics of the learning algorithm is such that it can keep this effect under check and is capable to learn. The more it learns, the more it reduces the size of the gradient noise and this feedback mechanism leads to effective learning. This is one of the key conclusions in this work, namely, showing that indeed the probabilities of error decay exponentially with the inverse of the step-size. This result is non-trivial and the derivations will take some effort before arriving at the insightful scaling laws that we are presenting in this work.

C. Analysis of Detection Performance

The aforementioned properties of the diffusion strategies used in this work explain their widespread utilization in the context of adaptive estimation [17], and motivate their use in the context of adaptive distributed detection [29]–[31]. With reference to this class of algorithms, while several results have been obtained for the mean-square-error (MSE) *estimation* performance of adaptive networks [15], [20], less is known about the performance of distributed detection networks. In particular, in [29], the miss-detection and false-alarm probabilities have been evaluated with reference to Gaussian observations. However, a detailed analytical characterization of the detection performance (i.e., false-alarm and detection probabilities), with reference to a general observational model, is still missing. This is mainly due to the fact that results on the asymptotic distribution of the error quantities under constant step-size adaptation over networks are largely unavailable in the literature.

While reference [32] argues that the error in single-agent least-mean-squares (LMS) adaptation converges in distribution, the resulting distribution is not characterized. These questions are considered in [33], [34] in the context of distributed estimation over adaptive networks. Nevertheless, these results on the asymptotic distribution of the errors are still insufficient to characterize the rate of decay of the probability of error over networks of distributed detectors. The main purpose of this work is to fill this gap. To do so, it is necessary to pursue a large deviations analysis in the constant step-size regime. Motivated by these remarks, we therefore provide a thorough statistical characterization of the diffusion network in a manner that enables detector design and analysis.

Notation. We use boldface letters to denote random variables, and normal font letters for their realizations. Capital letters refer to matrices, small letters to both vectors and scalars. Sometimes we violate this latter convention, for instance, we denote the total number of sensors by S. The symbols $\mathbb{P}$ and $\mathbb{E}$ are used to denote the probability and expectation operators, respectively. The notation $\mathbb{P}_h$ and $\mathbb{E}_h$, with $h = 0, 1$, means that the pertinent statistical distribution corresponds to hypothesis $\mathcal{H}_0$ or $\mathcal{H}_1$.

II. Preliminaries and Main Results

Consider a connected network of S agents. The scalar observation collected by the k-th sensor at time n will be denoted by $\boldsymbol{x}_k(n)$, $k = 1, 2, \ldots, S$. Data are assumed to be spatially and temporally independent and identically distributed (i.i.d.), *conditioned* on the hypothesis that gives rise to them. The distributed network is interested in making an inference about the true state of nature (i.e., the underlying hypothesis), which is allowed to vary over time. Since in this work we focus on a steady-state analysis, it is unnecessary at this stage to introduce an explicit dependence of the datum $\boldsymbol{x}_k(n)$ on the particular hypothesis giving rise to it.

Remark. When dealing with i.i.d. observations across sensors, the important issue of local versus aggregate distinguishability is bypassed. In most practical scenarios, sensors observe different aspects of a field, so local distinguishability is hard to achieve but the collective observation model may still be globally informative. The issue when local information is not sufficient for discrimination has been studied in several works before, including [35]–[37], and in other related references on diffusion strategies. In the context of multi-agent processing, the distinguishability condition essentially amounts to a positivity condition on the global Gramian (Hessian) matrix while allowing the individual Gramians to be non-negative definite. Learning is still possible in these cases, as shown, for example, in [17], [38], [39]. □

As it is well-known, for the i.i.d. data model, an optimal centralized (and non-adaptive) detection statistic is the sum of the log-likelihoods. When these are not available, alternative detection statistics obtained as the sum of some suitably chosen functions of the observations are often employed, as happens in some specific frameworks, e.g., in locally optimum detection [45] and in universal hypothesis testing [46]. Accordingly, each sensor in the network will try to compute, as its own detection statistic, a weighted combination of some function of the local observations. We assume the symbol $\boldsymbol{x}_k(n)$ represents the local statistic that is available at time n at sensor k.

Since we are interested in an adaptive inferential scheme, and given the idea of relying on weighted averages, we resort to the class of diffusion strategies for adaptation over networks [15], [29]. These strategies admit various forms. We consider the ATC form due to some inherent advantages in terms of a slightly improved mean-square-error performance relative to other forms [15]. In the ATC diffusion implementation, each node k updates its state from $\boldsymbol{y}_k(n-1)$ to $\boldsymbol{y}_k(n)$ through local cooperation with its neighbors as follows:

$$\boldsymbol{v}_k(n) = \boldsymbol{y}_k(n-1) + \mu[\boldsymbol{x}_k(n) - \boldsymbol{y}_k(n-1)], \quad (1)$$

$$\boldsymbol{y}_k(n) = \sum_{\ell=1}^{S} a_{k,\ell}\boldsymbol{v}_\ell(n) \quad (2)$$

where $0 < \mu \ll 1$ is a small step-size parameter. In this construction, node k first uses its local statistic, $\boldsymbol{x}_k(n)$, to update its state from $\boldsymbol{y}_k(n-1)$ to an intermediate value $\boldsymbol{v}_k(n)$. All other nodes in the network perform similar updates

simultaneously using their local statistics. Subsequently, node k aggregates the intermediate states of its neighbors using nonnegative convex combination weights $\{a_{k,\ell}\}$ that add up to one. Again, all other nodes in the network perform a similar calculation. If we collect the combination coefficients into a matrix $A = [a_{k,\ell}]$, then A is a right-stochastic matrix in that the entries on each of its rows add up to one:

$$a_{k,\ell} \geq 0, \quad A\mathbb{1} = \mathbb{1}, \tag{3}$$

with $\mathbb{1}$ being a column-vector with all entries equal to 1.

A. Performance and Convergence Analyses

At time n, the k-th sensor needs to produce a decision based upon its state value $\boldsymbol{y}_k(n)$. To this aim, a decision rule must be designed, by choosing appropriate decision regions. The performance of the test will be measured according to the Type-I (false-alarm) and Type-II (miss-detection) error probabilities defined, respectively, as

$$\alpha_k(n) \triangleq \mathbb{P}\left[\begin{array}{c}\text{agent } k \text{ decides } \mathcal{H}_1 \text{ at time } n \\ \text{while } \mathcal{H}_0 \text{ is true at time } n\end{array}\right], \tag{4}$$

$$\beta_k(n) \triangleq \mathbb{P}\left[\begin{array}{c}\text{agent } k \text{ decides } \mathcal{H}_0 \text{ at time } n \\ \text{while } \mathcal{H}_1 \text{ is true at time } n\end{array}\right]. \tag{5}$$

Note that these probabilities depend upon the statistical properties of the *whole* set of data used in the diffusion algorithm up to current time n. In particular, the error probabilities depend upon the different variations of the statistical distributions may have occurred during the evolution of the algorithm, and not only upon the particular hypothesis in force at time n.

Therefore, a rigorous analytical characterization of the system in terms of its overall inference performance at each time instant, and under general operation modalities (i.e., for arbitrarily varying statistical conditions) is generally not viable. This implies, among other difficulties, that the structure of the optimal, or even a reasonable test, is unknown. A standard approach in the adaptation literature to get useful performance metrics and meaningful insights, consists of splitting the analysis in two parts:

$i)$ A *transient* analysis where, starting from a given state, some variations in the statistical conditions occur and the time to track such variations is evaluated. It is possible to carry out studies that focus on the transient phase of the learning algorithm, and to clarify its behavior during this stage of operation, as is done in [38], [39].

$ii)$ A *steady-state* analysis, where the inference performance is evaluated with reference to an infinitely long period of stationarity. Even in the steady-state regime, an exact analytical characterization of the inference performance is seldom affordable. Therefore, closed-form results are usually obtained working in the regime of slow adaptation, i.e., of small step-sizes.

These two views are complementary. Typically, for a given value of the step-size μ, the diffusion algorithm exhibits the following features:

$i)$ The convergence rate towards the steady-state regime is known to occur at an exponential rate in the order of $O(c^n)$ for some $c \in (0,1)$; this is a faster rate than $O(1/n)$ that is afforded, for example, by diminishing step-sizes. Nevertheless, in the constant step-size case, the smaller the value of μ is, the closer the value of c gets to one.

$ii)$ The steady-state inference performance is a decreasing function of the step-size. Therefore, the lower μ is, the lower the steady-state error.

In this article, we address in some detail the steady-state performance of diffusion strategies for distributed detection over adaptive networks. Our main interest is in showing that the multi-agent network is able to learn well, with error probabilities exhibiting an exponential decay as functions of $1/\mu$. In particular, our analysis will be conducted with reference to the steady-state properties (as $n \to \infty$), and for small values of the step-size ($\mu \to 0$). Throughout the paper, the term steady-state will refer to the limit as the time-index n goes to infinity, while the term asymptotic will be used to refer to the slow adaptation regime where $\mu \to 0$. Specifically, we will follow these steps:

- We show that, in the stationary, steady-state regime, $\boldsymbol{y}_k(n)$ has a *limiting distribution* as n goes to infinity (Theorem 1).
- For small step-sizes, the steady-state distribution of $\boldsymbol{y}_k(n)$ approaches a Gaussian, i.e., it is *asymptotically normal* (Theorem 2).
- We characterize the *large deviations* of the steady-state output $\boldsymbol{y}_k(n)$ in the slow adaptation regime when $\mu \to 0$ (Theorem 3).
- The results of the above steps will provide a series of tools for designing the detector and characterizing its performance (Theorem 4).

B. Comparison with Decaying Step-Size Solutions

It is useful to contrast the above results with those pertaining to distributed detection algorithms with diminishing step-size [4]–[6]. The result in Theorem 1 reveals that, under stationary conditions, the detection statistic (i.e., the diffusion output $\boldsymbol{y}_k(n)$) converges to a limiting distribution, and the results in Theorem 2 add that such limiting distribution is approximately Gaussian in the slow adaptation regime. In contrast, in the diminishing step-size case, the detection statistic will collapse, as time elapses, into a *deterministic* value (e.g., the Kullback-Leibler divergence). Such convergence to a deterministic value reflects the continuously improving performance as time elapses, with diminishing step-sizes. In particular, under stationary conditions, the error probabilities for diminishing step-size algorithms decay exponentially *as functions of the time index* n — see, e.g. [4]–[6]. The latter feature must be contrasted with the results of our Theorems 3 and 4, where the exponential decay of the error probabilities does *not* refer to the time index n. Instead, we find the new result that the error probabilities decay exponentially *as functions of the (inverse of the) step-size* μ.

Finally, we would like to mention that the detailed statistical characterization offered by Theorems 1-3 is not confined to the specific detection problems we are dealing with. As a matter of fact, these results are of independent interest, and might be

useful for the application of adaptive diffusion strategies in broader contexts.

C. Main Results

As explained in the previous section, we focus on a connected network of S sensors, performing distributed detection by means of adaptive diffusion strategies. The adaptive nature of the solution allows the network to track variations in the hypotheses being tested over time. In order to enable continuous adaptation and learning, we shall employ distributed diffusion strategies with a *constant* step-size parameter μ. Now, let $\alpha_{k,\mu}$ and $\beta_{k,\mu}$ represent the steady-state (as $n \to \infty$) Type-I and Type-II error probabilities at the k-th sensor. One of the main conclusions established in this paper can be summarized by the following scaling laws:

$$\boxed{\alpha_{k,\mu} \doteq e^{-(1/\mu)\, S\, \mathcal{E}_0}, \qquad \beta_{k,\mu} \doteq e^{-(1/\mu)\, S\, \mathcal{E}_1}} \tag{6}$$

where the notation $\doteq$ means equality to the leading exponential order as μ goes to zero [40]. In the above expressions, the parameters $\mathcal{E}_0$ and $\mathcal{E}_1$ are solely dependent on the moment generating function of the single-sensor data $\boldsymbol{x}$, and of the decision regions. These parameters are *independent* of the step-size μ, the number of sensors S, and the network connectivity. Result (6) has at least four important and insightful ramifications about the performance of adaptive schemes for distributed detection over networks.

To begin with, Eq. (6) reveals a fundamental scaling law for distributed detection with diffusion adaptation, namely, it asserts that as the step-size decreases, the error probabilities are driven to zero exponentially as functions of $1/\mu$, and that the error exponents governing such a decay increase linearly in the number of sensors. These implications are even more revealing if examined in conjunction with the known results concerning the scaling law of the Mean-Square-Error (MSE) for adaptive distributed estimation over diffusion networks [15], [20]. Assuming a connected network with S sensors, and using sufficiently small step-sizes $\mu \approx 0$, the MSE that is attained by sensor k obeys (see expression (32) in [15]):

$$\text{MSE}_k \;\propto\; \frac{\mu}{S}, \tag{7}$$

where the symbol $\propto$ denotes proportionality. Some interesting symmetries are observed. In the estimation context, the MSE decreases as μ goes to zero, and the scaling rate improves linearly in the number of sensors. Recalling that smaller values of μ mean a lower degree of adaptation, we observe that reaching a better inference quality costs in terms of adaptation speed. This is a well-known trade-off in the adaptive estimation literature between tracking speed and estimation accuracy.

Second, we observe from (6) and (7) that the scaling laws governing errors of detection and estimation over distributed networks behave very differently, the former exhibiting an exponential decay proportional to $1/\mu$, while the latter is linear with decay proportional to μ. The significance and elegance of this result for adaptive distributed networks lie in revealing an intriguing analogy with other more traditional inferential schemes. As a first example, consider the standard case of a centralized, non-adaptive inferential system with N i.i.d. data points. It is known that the error probabilities of the best detector decay exponentially fast to zero with N, while the optimal estimation error decays as $1/N$ [41], [42]. Another important case is that of rate-constrained multi-terminal inference [43], [44]. In this case the detection performance scales exponentially with the bit-rate R while, again, the squared estimation error vanishes as $1/R$. Thus, at an abstract level, reducing the step-size corresponds to increasing the number of independent observations in the first system, or increasing the bit-rate in the second system. The above comparisons furnish an interesting interpretation for the step-size μ as the basic parameter quantifying the cost of information used by the network for inference purposes, much as the number of data N or the bit-rate R in the considered examples.

A third aspect pertaining to the performance of the distributed network relates to the potential benefits of cooperation. These are already encoded into (6), and we have already implicitly commented on them. Indeed, note that the error exponents increase linearly in the number of sensors. This implies that cooperation offers *exponential gains* in terms of detection performance.

The fourth and final ramification we would like to highlight relates to how much performance is lost by the *distributed* solution in comparison to a centralized stochastic gradient solution. Again, the answer is contained in (6). Specifically, the centralized solution is equivalent to a fully connected network, so that (6) applies to the centralized case as well. As already mentioned, the parameters $\mathcal{E}_0$ and $\mathcal{E}_1$ do not depend on the network connectivity, which therefore implies that, as the step-size μ decreases, the distributed diffusion solution of the inference problem exhibits a detection performance governed by the *same error exponents* of the centralized system. This is a remarkable conclusion and it is also consistent with results in the context of adaptive distributed estimation over diffusion networks [15].

We now move on to describe the adaptive distributed solution and to establish result (6) and the aforementioned properties.

III. Existence of Steady-State Distribution

Let $\boldsymbol{y}_n$ denote the $S \times 1$ vector that collects the state variables from across the network at time n, i.e.,

$$\boldsymbol{y}_n = \text{col}\{\boldsymbol{y}_1(n),\, \boldsymbol{y}_2(n),\, \ldots,\, \boldsymbol{y}_S(n)\}. \tag{8}$$

Likewise, we collect the local statistics $\{\boldsymbol{x}_k(n)\}$ at time n into the vector $\boldsymbol{x}_n$. It is then straightforward to verify from the diffusion strategy (1)–(2) that the vector $\boldsymbol{y}_n$ is given by:

$$\boxed{\boldsymbol{y}_n = (1-\mu)^n\, A^n \boldsymbol{y}_0 + \frac{\mu}{1-\mu} \sum_{i=1}^{n} (1-\mu)^{n-i+1} A^{n-i+1} \boldsymbol{x}_i} \tag{9}$$

We are concerned here with a *steady-state* analysis. Accordingly, we must examine the situation where the data are possibly nonstationary up to a certain time instant, after

which they are drawn from the same stationary distribution for infinitely long time. This implies that, when performing the steady-state analysis, it suffices to assume that the data, for all $n \geq 1$, arise from one and the same distribution. The past history (including possible drifts occurred in the statistical conditions) that influences the overall algorithm evolution, is reflected in the initial state vector $\boldsymbol{y}_0$. In addition, since, for $n \geq 1$, we only need to specify the particular distribution from which data are drawn, in the forthcoming derivations we shall conduct our study with reference to a sequence of i.i.d. data with a given distribution. Later on, when applying the main findings to the detection problem, we shall use a subscript $h \in \{0, 1\}$ to denote that data follow the distribution corresponding to a particular hypothesis.

We are now ready to show the existence and the specific shape of the limiting distribution. By making the change of variables $i \leftarrow n - i + 1$, Eq. (9) can be written as

$$\boldsymbol{y}_n = (1-\mu)^n A^n \boldsymbol{y}_0 + \frac{\mu}{1-\mu} \sum_{i=1}^{n} (1-\mu)^i A^i \boldsymbol{x}_{n-i+1}. \quad (10)$$

It follows that the state of the k-th sensor is given by:

$$\begin{aligned} \boldsymbol{y}_k(n) &= \underbrace{(1-\mu)^n \sum_{\ell=1}^{S} b_{k,\ell}(n) \boldsymbol{y}_\ell(0)}_{\text{transient}} \\ &+ \underbrace{\frac{\mu}{1-\mu} \sum_{i=1}^{n} (1-\mu)^i \sum_{\ell=1}^{S} b_{k,\ell}(i) \boldsymbol{x}_\ell(n-i+1)}_{\text{steady-state}}, \end{aligned} \quad (11)$$

where the scalars $b_{k,\ell}(n)$ are the entries of the matrix power:

$$B_n \triangleq A^n. \quad (12)$$

Since we are interested in reaching a *balanced* fusion of the observations, we shall assume that A is *doubly stochastic* with second largest eigenvalue magnitude strictly less than one, which yields [8], [16], [48]:

$$B_n \overset{n\to\infty}{\longrightarrow} \frac{1}{S} \mathbb{1}\mathbb{1}^T. \quad (13)$$

Now, we notice that the first term on the RHS of (11) vanishes almost surely (a.s.) (and, hence, in probability [41]) with n, since, for any initial state vector $\boldsymbol{y}_0$, we have:

$$\left| (1-\mu)^n \sum_{\ell=1}^{S} b_{k,\ell}(n) \boldsymbol{y}_\ell(0) \right| \leq (1-\mu)^n \sum_{\ell=1}^{S} |\boldsymbol{y}_\ell(0)|. \quad (14)$$

Accordingly, if we are able to show that the second term on the RHS of (11) converges to a certain limiting distribution, we can then conclude that the variable $\boldsymbol{y}_k(n)$ converges as well to the same limiting distribution, as a direct application of Slutsky's Theorem [41].

In order to reveal the steady-state behavior of $\boldsymbol{y}_k(n)$, it suffices to focus on the last summation in (11). We observe preliminarily that the term $\boldsymbol{x}_{n-i+1}$ in (10) depends on the time index n in such a way that the most recent datum $\boldsymbol{x}_n$ is assigned the highest scaling weight, in compliance with the adaptive nature of the algorithm. However, since the vectors $\boldsymbol{x}_i$ are i.i.d. across time, and since we shall be only concerned with the distribution of partial sums involving these terms, the statistical properties of the summation in (10) are left unchanged if we replace $\boldsymbol{x}_{n-i+1}$ with a random vector $\boldsymbol{x}'_i$, where $\{\boldsymbol{x}'_i\}$ is a sequence of i.i.d. random vectors distributed similarly to the $\{\boldsymbol{x}_{n-i+1}\}$. Formally, as regards the steady-state term on the RHS of (11), we can write:

$$\begin{aligned} &\frac{\mu}{1-\mu} \sum_{i=1}^{n} (1-\mu)^i \sum_{\ell=1}^{S} b_{k,\ell}(i) \boldsymbol{x}_\ell(n-i+1) \\ &\quad \overset{d}{=} \frac{\mu}{1-\mu} \sum_{i=1}^{n} (1-\mu)^i \sum_{\ell=1}^{S} b_{k,\ell}(i) \boldsymbol{x}'_\ell(i) \triangleq \sum_{i=1}^{n} \boldsymbol{z}_k(i), \end{aligned} \quad (15)$$

where $\overset{d}{=}$ denotes equality *in distribution*, and where the definition of $\boldsymbol{z}_k(i)$ should be clear. As a result, we are faced with a sum of independent, but *not* identically distributed, random variables. Let us evaluate the first two moments of the sum:

$$\mathbb{E}\left(\sum_{i=1}^{n} \boldsymbol{z}_k(i) \right) = \mathbb{E}\boldsymbol{x} \sum_{i=1}^{n} \mu(1-\mu)^{i-1} \underbrace{\sum_{\ell=1}^{S} b_{k,\ell}(i)}_{=1} \overset{n\to\infty}{\longrightarrow} \mathbb{E}\boldsymbol{x}, \quad (16)$$

and

$$\begin{aligned} \text{VAR}\left(\sum_{i=1}^{n} \boldsymbol{z}_k(i) \right) &= \sigma_x^2 \sum_{i=1}^{n} \mu^2 (1-\mu)^{2(i-1)} \underbrace{\sum_{\ell=1}^{S} b_{k,\ell}^2(i)}_{\leq 1} \\ &\leq \frac{\sigma_x^2 \mu}{2-\mu} < \infty, \end{aligned} \quad (17)$$

where VAR denotes the variance operator, and $\sigma_x^2 \triangleq \text{VAR}(\boldsymbol{x})$. We have thus shown that the expectation of the sum expression from (15) converges to $\mathbb{E}\boldsymbol{x}$, and that its variance converges to a finite value. In view of the Infinite Convolution Theorem — see [49, p. 266], these two conditions are sufficient to conclude that the RHS of (15), i.e., the sum of random variables $\boldsymbol{z}_k(i)$, converges in distribution as $n \to \infty$, and the first two moments of the limiting distribution are equal to $\mathbb{E}\boldsymbol{x}$ and $\sum_{i=1}^{\infty} \text{VAR}(\boldsymbol{z}_k(i))$. The random variable characterized by the limiting distribution will be denoted by $\boldsymbol{y}^\star_{k,\mu}$, where we make explicit the dependence upon the step-size μ for later use.

The above statement can be sharpened to ascertain that the sum of random variables $\boldsymbol{z}_k(i)$ actually converges almost surely. This conclusion can be obtained by applying Kolmogorov's Two Series Theorem [49]. In view of the a.s. convergence, it makes sense to define the limiting random variable $\boldsymbol{y}^\star_{k,\mu}$ as:

$$\boxed{\boldsymbol{y}^\star_{k,\mu} \triangleq \sum_{i=1}^{\infty} \sum_{\ell=1}^{S} \mu (1-\mu)^{i-1} b_{k,\ell}(i) \boldsymbol{x}'_\ell(i)} \quad (18)$$

We wish to avoid confusion here. We are not stating that the actual diffusion output $\boldsymbol{y}_k(n)$ converges almost surely

(a behavior that would go against the adaptive nature of the diffusion algorithm). We are instead claiming that $\boldsymbol{y}_k(n)$ converges in distribution to a random variable $\boldsymbol{y}^\star_{k,\mu}$ that can be conveniently defined in terms of the a.s. limit (18).

The main result about the steady-state behavior of the diffusion output is summarized below (the symbol $\rightsquigarrow$ means convergence in distribution).

THEOREM 1: *(Steady-state distribution of $\boldsymbol{y}_k(n)$). The state variable $\boldsymbol{y}_k(n)$ that is generated by the diffusion strategy (1)–(2) is asymptotically stable in distribution, namely,*

$$\boxed{\boldsymbol{y}_k(n) \overset{n\to\infty}{\rightsquigarrow} \boldsymbol{y}^\star_{k,\mu}} \tag{19}$$

□

It is useful to make explicit the meaning of Theorem 1. By definition of convergence in distribution (or weak convergence), the result (19) can be formally stated as [42], [50]:

$$\lim_{n\to\infty} \mathbb{P}[\boldsymbol{y}_k(n) \in \Gamma] = \mathbb{P}[\boldsymbol{y}^\star_{k,\mu} \in \Gamma], \tag{20}$$

for any set Γ such that $\mathbb{P}[\boldsymbol{y}^\star_{k,\mu} \in \partial\Gamma] = 0$, where $\partial\Gamma$ denotes the boundary of Γ. It is thus seen that the properties of the steady-state variable $\boldsymbol{y}^\star_{k,\mu}$ will play a key role in determining the steady-state performance of the diffusion output. Accordingly, we state two useful properties of $\boldsymbol{y}^\star_{k,\mu}$.

First, when the local statistic $\boldsymbol{x}_k(n)$ has an *absolutely continuous* distribution (where the reference measure is the Lebesgue measure over the real line), it is easily verified that the distribution of $\boldsymbol{y}^\star_{k,\mu}$ is *absolutely continuous as well*. Indeed, note that we can write $\boldsymbol{y}^\star_{k,\mu} = \boldsymbol{z}_k(1) + \sum_{i=2}^{\infty} \boldsymbol{z}_k(i)$. Now observe that $\boldsymbol{z}_k(1)$, which has an absolutely continuous distribution by assumption, is independent of the other term. The result follows by the properties of convolution and from the fact that the distribution of the sum of two independent variables is the convolution of their respective distributions.

Second, when the local statistic $\boldsymbol{x}_k(n)$ is a *discrete* random variable, by the Jessen-Wintner law [51], [52], we can only conclude that $\boldsymbol{y}^\star_{k,\mu}$ is of *pure type*, namely, its distribution is pure: absolutely continuous, or discrete, or continuous but singular.

An intriguing case is that of the so-called *Bernoulli convolutions*, i.e., random variables of the form $\sum_{i=1}^{\infty}(1-\mu)^{i-1}\boldsymbol{x}(i)$, where $\boldsymbol{x}(i)$ are equiprobable ± 1. For this case, it is known that if $1/2 < \mu < 1$, then the limiting distribution is a *Cantor* distribution [53]. This is an example of a distribution that is neither discrete nor absolutely continuous. When $\mu < 1/2$, which is relevant for our discussion since we shall be concerned with small step-sizes, the situation is markedly different, and the distribution is absolutely continuous for almost all values of μ.

Before proceeding, we stress that we have proved that a steady-state distribution for $\boldsymbol{y}_k(n)$ exists, but its form is not known. Accordingly, even in steady-state, the structure of the optimal test is still unknown. In tackling this issue, and recalling that the regime of interest is that of slow adaptation, we now focus on the case $\mu \ll 1$.

IV. THE SMALL-μ REGIME.

While the exact form of the steady-state distribution is generally impossible to evaluate, it is nevertheless possible to approximate it well for small values of the step-size parameter. Indeed, in this section we prove two results concerning the statistical characterization of the steady-state distribution for $\mu \to 0$. The first one is a result of *asymptotic normality*, stating that $\boldsymbol{y}^\star_{k,\mu}$ approaches a Gaussian random variable with known moments as μ goes to zero (Theorem 2). The second finding (Theorem 3) provides the complete characterization for the *large deviations* of $\boldsymbol{y}^\star_{k,\mu}$. In the following, $\mathcal{N}(a,b)$ is a shortcut for a Gaussian distribution with mean a and variance b, and the symbol $\sim$ means "distributed as".

THEOREM 2: *(Asymptotic normality of $\boldsymbol{y}^\star_{k,\mu}$ as $\mu \to 0$). Under the assumption $\mathbb{E}|\boldsymbol{x}_k(n)|^3 < \infty$, the variable $\boldsymbol{y}^\star_{k,\mu}$ fulfills, for all $k = 1, 2, \ldots, S$:*

$$\boxed{\frac{\boldsymbol{y}^\star_{k,\mu} - \mathbb{E}\boldsymbol{x}}{\sqrt{\mu}} \overset{\mu\to 0}{\rightsquigarrow} \mathcal{N}\left(0, \frac{\sigma_x^2}{2\,S}\right)} \tag{21}$$

Proof: The argument requires dealing with independent but non-identically distributed random variables, as done in the Lindeberg-Feller CLT (Central Limit Theorem) [49]. This theorem, however, does not apply to our setting since the asymptotic parameter is *not* the number of samples, but rather the step-size. Some additional effort is needed, and the detailed technical derivation is deferred to Appendix A. □

A. Implications of Asymptotic Normality

Let us now briefly comment on several useful implications that follow from the above theorem:

1) First, note that *all sensors* share, for μ small enough, the *same* distribution, namely, the inferential diffusion strategy equalizes the statistical behavior of the agents. This finding complements well results from [15], [20], [34] where the asymptotic equivalence among the sensors has been proven in the context of mean-square-error estimation. One of the main differences between the estimation context and the detection context studied in this article is that in the latter case, the regression data is deterministic and the randomness arises from the stochastic nature of the statistics $\{\boldsymbol{x}_k(n)\}$. For this reason, the steady-state distribution in (21) is characterized in terms of the moments of these statistics and not in terms of the moments of regression data, as is the case in the estimation context.
2) The result of Theorem 2 is valid provided that the connectivity matrix fulfills (13). This condition is satisfied when the network topology is strongly-connected, i.e., there exists a path with nonzero weights connecting any two arbitrary nodes and at least one node has $a_{k,k} > 0$ [16]. Obviously, condition (13) is also satisfied in the fully connected case when $a_{k,\ell} = b_{k,\ell} = 1/S$ for all $k, \ell = 1, 2, \ldots, S$. This latter situation would correspond to a representation of the centralized stochastic gradient

algorithm, namely, an implementation of the form

$$\boldsymbol{y}^{(c)}(n) = \boldsymbol{y}^{(c)}(n-1) + \frac{\mu}{S}\sum_{\ell=1}^{S}[\boldsymbol{x}_\ell(n) - \boldsymbol{y}^{(c)}(n-1)], \tag{22}$$

where $\boldsymbol{y}^{(c)}(n)$ denotes the output by the centralized solution at time n. The above algorithm can be deduced from (1)–(2) by defining

$$\boldsymbol{y}^{(c)}(n) \triangleq \frac{1}{S}\sum_{\ell=1}^{S}\boldsymbol{y}_\ell(n). \tag{23}$$

Now, since the moments of the limiting Gaussian distribution in (21) are independent of the particular connectivity matrix, the net effect is that each agent of the *distributed* network acts, asymptotically, as the *centralized* system. This result again complements well results in the estimation context where the role of the statistics variables $\{\boldsymbol{x}_k(n)\}$ is replaced by that of stochastic regression data [54].

3) The asymptotic normality result is powerful in approximating the steady-state distribution for relatively small step-sizes, thus enabling the analysis and design of inferential diffusion networks in many different contexts. With specific reference to the detection application that is the main focus here, Eq. (21) can be exploited for an accurate threshold setting when one desires to keep under control one of the two errors, say, the false-alarm probability, as happens, e.g., in the Neyman-Pearson setting [42]. To show a concrete example on how this can be done, let us assume that, without loss of generality, $\mathbb{E}_0\boldsymbol{x} < \mathbb{E}_1\boldsymbol{x}$, and consider a single-threshold detector for which:

$$\Gamma_0 = \{\gamma \in \mathbb{R} : \gamma \leq \eta_\mu\}, \qquad \Gamma_1 = \mathbb{R} \setminus \Gamma_0, \tag{24}$$

where the threshold is set as

$$\eta_\mu = \mathbb{E}_0\boldsymbol{x} + \sqrt{\frac{\mu\sigma_{x,0}^2}{2S}}\,Q^{-1}(\bar{\alpha}). \tag{25}$$

Here, $\sigma_{x,0}^2$ is the variance of $\boldsymbol{x}$ under $\mathcal{H}_0$, $Q(\cdot)$ denotes the complementary CDF for a standard normal distribution, and $\bar{\alpha}$ is the prescribed false-alarm level. By (21), it is straightforward to check that this threshold choice ensures

$$\lim_{\mu\to 0}\mathbb{P}_0[\boldsymbol{y}_{k,\mu}^\star > \eta_\mu] = \bar{\alpha}. \tag{26}$$

In summary, Theorem 2 provides an approximation of the diffusion output distribution for small step-sizes. At first glance, this may seem enough to obtain a complete characterization of the detection problem. A closer inspection reveals that this is not the case. A good example to understand why Theorem 2 alone is insufficient for characterizing the detection performance is obtained by examining the Neyman-Pearson threshold setting just described in (25)–(26) above. While we have seen that the asymptotic behavior of the false-alarm probability in (26) is completely determined by the application of Theorem 2, the situation is markedly different as regards the miss-detection probability $\mathbb{P}_1[\boldsymbol{y}_{k,\mu}^\star \leq \eta_\mu]$. Indeed, by using (25) we can write:

$$\begin{aligned}\mathbb{P}_1[\boldsymbol{y}_{k,\mu}^\star \leq \eta_\mu] &= \mathbb{P}_1\left[\frac{\boldsymbol{y}_{k,\mu}^\star - \mathbb{E}_1\boldsymbol{x}}{\sqrt{\mu}} \leq \frac{\eta_\mu - \mathbb{E}_1\boldsymbol{x}}{\sqrt{\mu}}\right] \\ &= \mathbb{P}_1\left[\frac{\boldsymbol{y}_{k,\mu}^\star - \mathbb{E}_1\boldsymbol{x}}{\sqrt{\mu}} \leq \frac{\mathbb{E}_0\boldsymbol{x} - \mathbb{E}_1\boldsymbol{x}}{\sqrt{\mu}} + \sqrt{\frac{\sigma_{x,0}^2}{2S}}\,Q^{-1}(\bar{\alpha})\right].\end{aligned} \tag{27}$$

Since $\mathbb{E}_0\boldsymbol{x} < \mathbb{E}_1\boldsymbol{x}$, the quantity $\frac{\mathbb{E}_0\boldsymbol{x}-\mathbb{E}_1\boldsymbol{x}}{\sqrt{\mu}}$ diverges to $-\infty$ as $\mu \to 0$. As a consequence, the fact that $\frac{\boldsymbol{y}_{k,\mu}^\star-\mathbb{E}_1\boldsymbol{x}}{\sqrt{\mu}}$ is asymptotically normal does not provide much more insight than revealing that the miss-detection probability converges to zero as $\mu \to 0$. A meaningful asymptotic analysis would instead require to examine the way this convergence takes place (i.e., the error exponent). The same kind of problem is found when one lets *both* error probabilities vanish exponentially, such that the Type-I and Type-II detection error exponents furnish a meaningful asymptotic characterization of the detector. In order to fill these gaps, the study of the *large* deviations of $\boldsymbol{y}_{k,\mu}^\star$ is needed.

B. Large Deviations of $\boldsymbol{y}_{k,\mu}^\star$.

From (21) we learn that, as $\mu \to 0$, the diffusion output shrinks down to its limiting expectation $\mathbb{E}\boldsymbol{x}$ and that the *small* (of order $\sqrt{\mu}$) deviations around this value have a Gaussian shape. But this conclusion is not helpful when working with *large* deviations, namely, with terms like:

$$\mathbb{P}[|\boldsymbol{y}_{k,\mu}^\star - \mathbb{E}\boldsymbol{x}| > \delta] \overset{\mu\to 0}{\longrightarrow} 0, \quad \delta > 0, \tag{28}$$

which play a significant role in detection applications. While the above convergence to zero can be inferred from (21), it is well known that (21) is not sufficient in general to obtain the rate at which the above probability vanishes. In order to perform accurate design and characterization of reliable inference systems [55], [56] it is critical to assess this rate of convergence, which turns out to be the main purpose of a large deviations analysis.

Accordingly, we will be showing in the sequel that the process $\boldsymbol{y}_{k,\mu}^\star$ obeys a Large Deviations Principle (LDP), namely, that the following limit exists [55], [56]:

$$\lim_{\mu\to 0}\mu\,\ln\mathbb{P}[\boldsymbol{y}_{k,\mu}^\star \in \Gamma] = -\inf_{\gamma\in\Gamma} I(\gamma) \triangleq -I_\Gamma, \tag{29}$$

for some $I(\gamma)$ that is called the *rate function*. Equivalently:

$$\mathbb{P}[\boldsymbol{y}_{k,\mu}^\star \in \Gamma] = e^{-(1/\mu)\,I_\Gamma + o(1/\mu)} \doteq e^{-(1/\mu)\,I_\Gamma}, \tag{30}$$

where $o(1/\mu)$ stands for any correction term growing slower than $1/\mu$, namely, such that $\mu\,o(1/\mu) \to 0$ as $\mu \to 0$, and the notation $\doteq$ was introduced in (6). From (30) we see that, in the large deviations framework, only the dominant exponential term is retained, while discarding any sub-exponential terms. It is also interesting to note that, according to (30), the probability that $\boldsymbol{y}_{k,\mu}^\star$ belongs to a given region Γ is dominated by the infimum I_Γ of the rate function $I(\gamma)$ within the region Γ. In other words, the smallest exponent ($\Rightarrow$ highest

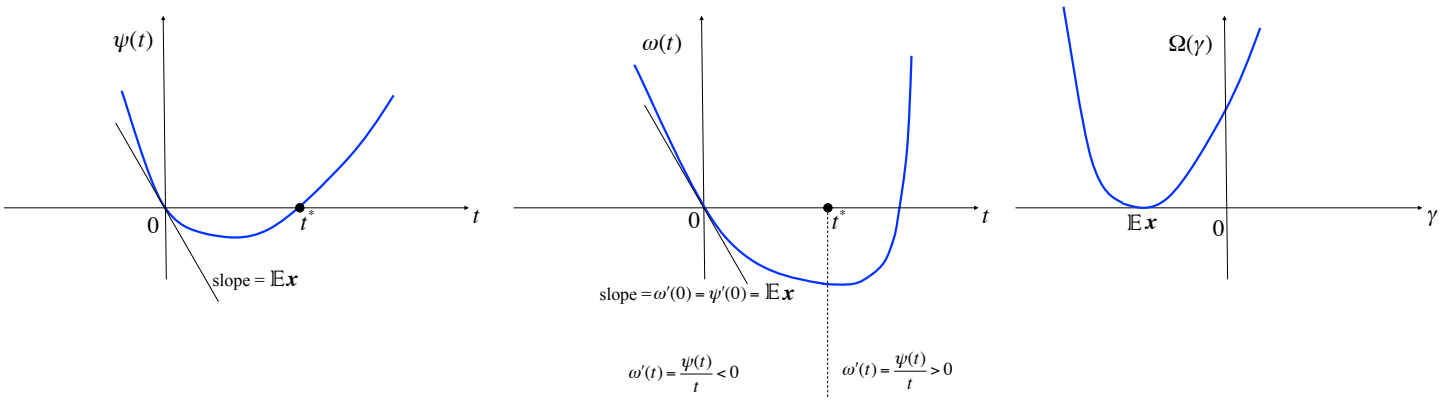

Fig. 2. Leftmost panel: The LMGF $\psi(t)$ of the original data $\boldsymbol{x}_k(n)$; its slope at the origin is $\mathbb{E}\boldsymbol{x}$. Middle panel: The function $\omega(t)$ defined by (36) is strictly convex; its slope at the orgin is also equal to $\mathbb{E}\boldsymbol{x}$. The labels underneath the plot illustrate the intervals over which $\omega'(t)$ is negative and positive for the LMGF $\psi(t)$ shown in the leftmost plot. Rightmost panel: The Fenchel-Legendre transform, $\Omega(\gamma)$, which is relevant for the evaluation of the rate function, attains the minimum value of zero at $\gamma = \mathbb{E}\boldsymbol{x}$.

probability) dominates, which is well explained in [56] through the statement: "*any large deviation is done in the least unlikely of all the unlikely ways*".

In summary, the LDP generally implies an exponential scaling law for probabilities, with an exponent governed by the rate function. Therefore, knowledge of the rate function is enough to characterize the exponent in (30). We shall determine the expression for $I(\gamma)$ pertinent to our problem in Theorem 3 further ahead — see Eq. (37).

In the traditional case where the statistic under consideration is the arithmetic average of i.i.d. data, the asymptotic parameter is the number of samples and the usual tool for determining the rate function in the LDP is Cramér's Theorem [55], [56]. Unfortunately, in our adaptive and distributed setting, we are dealing with a more general statistic $\boldsymbol{y}_{k,\mu}^{\star}$, whose dependence is on the step-size parameter and not on the number of samples. Cramér's Theorem is not applicable in this case, and we must resort to a more powerful tool, known as the Gärtner-Ellis Theorem [55], [56], stated below in a form that uses directly the set of assumptions relevant for our purposes.

GÄRTNER-ELLIS THEOREM [56]. *Let $\boldsymbol{z}_\mu$ be a family of random variables with Logarithmic Moment Generating Function (LMGF) $\phi_\mu(t) = \ln \mathbb{E}\exp\{t\boldsymbol{z}_\mu\}$. If*

$$\phi(t) \triangleq \lim_{\mu\to 0} \mu\,\phi_\mu(t/\mu) \tag{31}$$

exists, with $\phi(t) < \infty$ for all $t \in \mathbb{R}$, and $\phi(t)$ is differentiable in $\mathbb{R}$, then $\boldsymbol{z}_\mu$ satisfies the LDP property (29) with rate function given by the Fenchel-Legendre transform of $\phi(t)$, namely:

$$\Phi(\gamma) \triangleq \sup_{t\in\mathbb{R}}[\gamma t - \phi(t)]. \tag{32}$$

□

In what follows, we shall use capital letters to denote Fenchel-Legendre transforms, as done in (32).

We now show how the result allows us to assess the asymptotic performance of the diffusion output in the inferential network. Let us introduce the LMGF of the data $\boldsymbol{x}_k(n)$, and that of the steady-state variable $\boldsymbol{y}_{k,\mu}^{\star}$, respectively:

$$\psi(t) \triangleq \ln \mathbb{E}\exp\{t\boldsymbol{x}_k(n)\}, \tag{33}$$

$$\phi_{k,\mu}(t) \triangleq \ln \mathbb{E}\exp\{t\boldsymbol{y}_{k,\mu}^{\star}\}. \tag{34}$$

THEOREM 3: *(Large deviations of $\boldsymbol{y}_{k,\mu}^{\star}$ as $\mu \to 0$). Assume that $\psi(t) < \infty$ for all $t \in \mathbb{R}$. Then, for all $k = 1, 2, \ldots, S$:*

i)

$$\boxed{\phi(t) \triangleq \lim_{\mu\to 0} \mu\,\phi_{k,\mu}(t/\mu) = S\,\omega(t/S)} \tag{35}$$

where

$$\boxed{\omega(t) \triangleq \int_0^t \frac{\psi(\tau)}{\tau} d\tau} \tag{36}$$

ii) The steady-state variable $\boldsymbol{y}_{k,\mu}^{\star}$ obeys the LDP with a rate function given by:

$$\boxed{I(\gamma) = S\,\Omega(\gamma)} \tag{37}$$

that is, by the Fenchel-Legendre transform of $\omega(t)$ multiplied by the number of sensors S.

Proof: See Appendix B. □

C. Main Implications of Theorem 3

From Theorem 3, a number of interesting conclusions can be drawn:

- The function $\omega(t)$ in (36) depends only upon the LMGF $\psi(t)$ of the original statistic $\boldsymbol{x}_k(n)$, and does *not* depend on the number of sensors.
- As a consequence of the above observation, part $ii)$ implies that the rate function (and, therefore, the large deviations exponent) of the diffusion output depends *linearly on the number of sensors*. Moreover, the rate can be determined by knowing only the statistical distribution of the input data $\boldsymbol{x}_k(n)$.
- The rate function does not depend on the particular sensor k. This implies that *all sensors are asymptotically equivalent also in terms of large deviations*, thus strengthening

what we have already found in terms of asymptotic normality — see Theorem 2 and the subsequent discussion.
- Theorem 3 can be applied to the centralized stochastic algorithm (22) as well, and, again, the diffusion strategy is able to match, asymptotically, the *centralized* solution.

Before ending this section, it is useful to comment on some essential features of the rate function $\Omega(\gamma)$, which will provide insights on its usage in connection with the distributed detection problem. To this aim, we refer to the following convexity properties shown in Appendix C (see also [55], Ex. 2.2.24, and [56], Ex. I.16):

i) $\omega''(t) > 0$ for all $t \in \mathbb{R}$, implying that $\omega(t)$ is strictly convex.

ii) $\Omega(\gamma)$ is strictly convex in the interior of the set:

$$\mathcal{D}_\Omega = \{\gamma \in \mathbb{R} : \ \Omega(\gamma) < \infty\}. \tag{38}$$

iii) $\Omega(\gamma)$ attains its unique minimum at $\gamma = \mathbb{E}\boldsymbol{x}$, with

$$\Omega(\mathbb{E}\boldsymbol{x}) = 0. \tag{39}$$

In light of these properties, it is possible to provide a geometric interpretation for the main quantities in Theorem 3, as illustrated in Fig. 2. The leftmost panel shows a typical behavior of the LMGF of the original data $\boldsymbol{x}_k(n)$. Using the result $\omega'(t) = \psi(t)/t$, and examining the sign of $\psi(t)/t$, it is possible to deduce the corresponding typical behavior of $\omega(t)$, depicted in the middle panel. As it can be seen, the slope at the origin is preserved, and is still equal to the expectation of the original data, $\mathbb{E}\boldsymbol{x}$. The intersection with the t-axis is changed, and moves further to the right in the considered example. Starting from $\omega(t)$, it is possible to draw a sketch of its Fenchel-Legendre transform $\Omega(\gamma)$ (rightmost panel), which illustrates its convexity properties, and the fact that the minimum value of zero is attained only at $\gamma = \mathbb{E}\boldsymbol{x}$.

V. The Distributed Detection Problem

The tools and results developed so far allow us to address in some detail the detection problem we are interested in. Let us denote the decision regions in favor of $\mathcal{H}_0$ and $\mathcal{H}_1$ by Γ_0 and Γ_1, respectively. We assume that they are the same at all sensors because, in view of the asymptotic equivalence among sensors proved in the previous section, there is no particular interest in making a different choice. Note, however, that all the subsequent development does not rely on this assumption and applies, *mutatis mutandis*, to the case of distinct decision regions used by distinct agents.

The Type-I and Type-II error probabilities at the k-th sensor at time n are defined in (4) and (5), respectively. Since we are interested in their *steady-state* behavior, namely, for an increasingly large interval where a certain hypothesis stays in force, the only distribution that matters is that corresponding to such hypothesis. Therefore, it is legitimate to write:

$$\lim_{n\to\infty} \alpha_k(n) \ = \ \lim_{n\to\infty} \mathbb{P}_0[\boldsymbol{y}_k(n) \in \Gamma_1], \tag{40}$$

$$\lim_{n\to\infty} \beta_k(n) \ = \ \lim_{n\to\infty} \mathbb{P}_1[\boldsymbol{y}_k(n) \in \Gamma_0], \tag{41}$$

where the subscripts 0 and 1 denote here the (stationary) situation where the data collected for all $n \geq 1$ come from one and the same distribution. As already observed, this simply corresponds to saying that the stationarity period used to compute the steady-state distribution starts at time $n = 1$. Some questions arise. Do these limits exist? Do these probabilities vanish as n approaches infinity? Theorem 1 provides the answers. Indeed, we found that $\boldsymbol{y}_k(n)$ stabilizes in distribution as n goes to infinity. In the sequel, in order to avoid dealing with pathological cases, we shall assume that $\mathbb{P}_0[\boldsymbol{y}^\star_{k,\mu} \in \partial\Gamma_1] = 0$ and that $\mathbb{P}_1[\boldsymbol{y}^\star_{k,\mu} \in \partial\Gamma_0] = 0$. This is a mild assumption, which is verified, for instance, when the limiting random variable $\boldsymbol{y}^\star_{k,\mu}$ has an absolutely continuous distribution, and the decision regions are not so convoluted to have boundaries with strictly positive measure. Accordingly, by invoking the weak convergence result of Theorem 1, and in view of (20) we can write:

$$\alpha_{k,\mu} \ \triangleq \ \lim_{n\to\infty} \alpha_k(n) = \mathbb{P}_0[\boldsymbol{y}^\star_{k,\mu} \in \Gamma_1], \tag{42}$$

$$\beta_{k,\mu} \ \triangleq \ \lim_{n\to\infty} \beta_k(n) = \mathbb{P}_1[\boldsymbol{y}^\star_{k,\mu} \in \Gamma_0], \tag{43}$$

where the dependence upon μ has been made explicit for later use. We notice that, in the above, we work with decision regions that do not depend on n, which corresponds exactly to the setup of Theorem 1. Generalizations where the regions are allowed to change with n can be handled by resorting to known results from asymptotic statistics. To give an example, consider the meaningful case of a detector with a sequence of thresholds $\eta(n)$ that converges to a value η as $n \to \infty$. Here,

$$\lim_{n\to\infty} \mathbb{P}_h[\boldsymbol{y}_k(n) > \eta(n)] = \mathbb{P}_h[\boldsymbol{y}^\star_{k,\mu} > \eta], \tag{44}$$

which can be seen, e.g., as an application of Slutsky's Theorem [41], [42].

From (42)–(43), it turns out that, as time elapses, the error probablities do not vanish exponentially. As a matter of fact, they do not vanish at all. This situation is in contrast to what happens in the case of running consensus strategies with diminishing step-size studied in the literature [1]–[6]. We wish to avoid confusion here. In the diminishing step-size case, one does need to examine the effect of large deviations [4]–[6] for large n, quantifying the rate of decay to zero of the error probabilities *as time progresses*. In the adaptive context, on the other hand, where *constant* step-sizes are used to enable continuous adaptation and learning, the large deviations analysis is totally different, in that it is aimed at characterizing the decaying rate of the error probabilities *as the step-size μ approaches zero.*

Returning to the detection performance evaluation (42)–(43), we stress that the steady-state values of these error probabilities are unknown, since the distribution of $\boldsymbol{y}^\star_{k,\mu}$ is generally unknown. However, the large deviations result offered by Theorem 3 allows us to characterize the *error exponents* in the regime of small step-sizes.

Theorem 3 can be tailored to our detection setup as follows (subscripts 0 and 1 are used to indicate that the statistical quantities are evaluated under $\mathcal{H}_0$ and $\mathcal{H}_1$, respectively):

THEOREM 4: *(Detection error exponents). For $h \in \{0, 1\}$, let Γ_h be the decision regions –independent of μ– and assume*

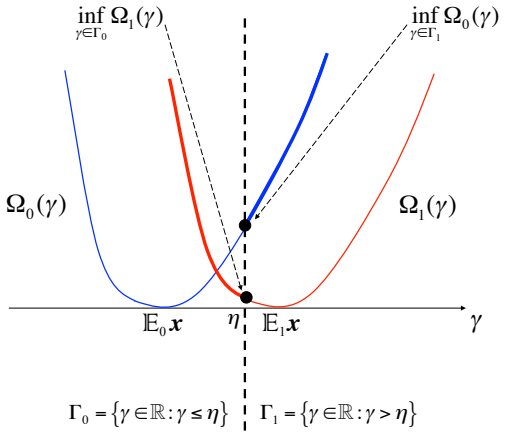

Fig. 3. A geometric view of Theorem 4.

that $\psi_h(t) < \infty$ *for all* $t \in \mathbb{R}$*, and define:*

$$\omega_h(t) \triangleq \int_0^t \frac{\psi_h(\tau)}{\tau} d\tau. \tag{45}$$

Then, for all $k = 1, 2, \ldots, S$*, Eq.(6) holds true, namely,*

$$\boxed{\lim_{\mu \to 0} \mu \, \ln \alpha_{k,\mu} = -S \, \mathcal{E}_0, \qquad \lim_{\mu \to 0} \mu \, \ln \beta_{k,\mu} = -S \, \mathcal{E}_1} \tag{46}$$

with

$$\boxed{\mathcal{E}_0 = \inf_{\gamma \in \Gamma_1} \Omega_0(\gamma), \qquad \mathcal{E}_1 = \inf_{\gamma \in \Gamma_0} \Omega_1(\gamma)} \tag{47}$$

where $\Omega_h(\gamma)$ *is the Fenchel-Legendre transform of* $\omega_h(t)$. □

REMARK I. The technical requirement that the LMGFs $\psi_0(t)$ and $\psi_1(t)$ are finite is met in many practical detection problems, as already shown in [5]. In particular, the assumption is clearly verified when the observations have (the same, under the two hypotheses) compact support, a special interesting case being that of discrete variables supported on a finite alphabet; and for shift-in-mean detection problems where the data distributions fulfill mild regularity conditions — see Remark II in [5] for a detailed list.

REMARK II. As typical in large deviations analysis, we have worked with regions Γ_0 and Γ_1 that do not depend on the step-size μ. Generalizations are possible to the case in which these regions depend on μ. A relevant case where this might be useful is the Neyman-Pearson setup, where one needs to work with a fixed (non-vanishing) value of the false-alarm probability. An example of this scenario is provided in Sec. VI-C — see the discussion following (78) — along with the detailed procedure for the required generalization.

In Fig. 3, we provide a geometric interpretation that can be useful to visualize the main message conveyed by Theorem 4. In order to rule out trivial cases, we assume that $\mathbb{E}_0\boldsymbol{x} \neq \mathbb{E}_1\boldsymbol{x}$, as happens, e.g., in the standard situation where the local statistic $\boldsymbol{x}_k(n)$ is a log-likelihood ratio and the detection problem is identifiable [42]. Without loss of generality, we take $\mathbb{E}_0\boldsymbol{x} < \mathbb{E}_1\boldsymbol{x}$, and, for the sake of concreteness, we consider a detector with threshold η, amounting to the following form for the decision regions:

$$\Gamma_0 = \{\gamma \in \mathbb{R} : \gamma \leq \eta\}, \qquad \Gamma_1 = \mathbb{R} \setminus \Gamma_0. \tag{48}$$

Let us set $\mathbb{E}_0\boldsymbol{x} < \eta < \mathbb{E}_1\boldsymbol{x}$ since, as will be clear soon, choosing a threshold outside the range $(\mathbb{E}_0\boldsymbol{x}, \mathbb{E}_1\boldsymbol{x})$ will lead to trivial performance for one of the error exponents. According to Theorem 4, to evaluate the exponent $\mathcal{E}_0$ (resp., $\mathcal{E}_1$), one must consider the worst-case, i.e., the smallest value of the function $\Omega_0(\gamma)$ (resp., $\Omega_1(\gamma)$), within the corresponding *error* region Γ_1 (resp., Γ_0). In view of the convexity properties discussed at the end of Sec. IV-C, and reported in Appendix C, we see that, for the threshold detector, both minima are attained only at $\gamma = \eta$. Certainly, this shape turns out to be of great interest in practical applications where, inspired by the optimality properties of a log-likelihood ratio test in the centralized case, a threshold detector is often an appealing and reasonable choice. On the other hand, we would like to stress that different, arbitrary decision regions can be in general chosen, and that the minima of $\Omega_0(\cdot)$ and $\Omega_1(\cdot)$ in Fig. 3 might be correspondingly located at two different points.

In summary, Theorem 4 allows us to compute the exponents $\mathcal{E}_0$ and $\mathcal{E}_1$ as functions of i) the kind of statistic $\boldsymbol{x}$ employed by the sensors, which determines the shape of the LMGFs $\psi_h(t)$ to be used in (45); and ii) of the employed decision regions relevant for the minimizations in (47). Once $\mathcal{E}_0$ and $\mathcal{E}_1$ have been found, the error probabilities $\alpha_{k,\mu}$ and $\beta_{k,\mu}$ can be approximated using Eq. (6). This result is then key for both detector design and analysis, so that we are now ready to illustrate the operation of the adaptive distributed network of detectors.

VI. EXAMPLES OF APPLICATION

In this section, we apply the developed theory to four relevant detection problems. We start with the classical Gaussian shift-in-mean problem. Then, we consider a scenario of specific relevance for sensor network applications, namely, detection with hardly (one-bit) quantized measurements. This case amounts to testing two Bernoulli distributions with different parameters under the different hypotheses. Both the Gaussian and the finite-alphabet assumptions are removed in the subsequent example, where a problem of relevance to radar applications is addressed, that is, shift-in-mean with additive noise sampled from a Laplace (double-exponential) distribution. Finally, we examine a case where the agents have limited knowledge of the underlying data model, and agree to employ a simple sample-mean detector, in the presence of noise distributed as a Gaussian mixture.

Before dwelling on the presentation of the numerical experiments, we provide some essential details on the strategy that has been implemented for obtaining them:

- The network used for our experiments consists of ten sensors, arranged so as to form the topology in Fig. 4, with combination weights $a_{k,\ell}$ following the Laplacian rule [8], [16].
- The decision rule for the detectors is based on comparing

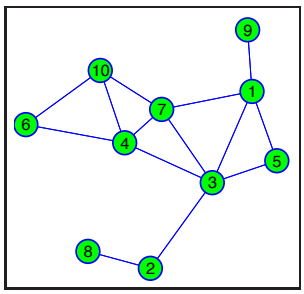

Fig. 4. Network skeleton used for the numerical simulations.

the diffusion output $\boldsymbol{y}_k(n)$ to some threshold η, namely,

$$\boldsymbol{y}_k(n) \underset{\mathcal{H}_1}{\overset{\mathcal{H}_0}{\lessgtr}} \eta, \tag{49}$$

where the decision regions are the same as in (48).

- Selecting the threshold η in (49) is a critical stage of detector design and implementation. This choice can be guided by different criteria, which would lead to different threshold settings. In the following examples, we present three relevant cases, namely: i) a threshold setting that is suited to the Bayesian and the max-min criteria (Sec. VI-B); ii) a Neyman-Pearson threshold setting (Sec. VI-C); iii) and a threshold setting in the presence of insufficient information about the underlying statistical models (Sec. VI-D). We would like to stress that using different threshold setting rules for different statistical models has no particular meaning. These choices are just meant to illustrate different rules and different models while avoiding repetition of similar results.
- The diffusion output is obtained after consultation steps involving the exchange of some local statistics $\boldsymbol{x}_k(n)$. The particular kind of statistic used in the different examples will be detailed when needed.

A. Shift-in-mean Gaussian Problem

The first hypothesis testing problem we consider is the following:

$$\mathcal{H}_0 \quad : \quad \boldsymbol{d}_k(n) \sim \mathcal{N}(0, \sigma^2), \tag{50}$$

$$\mathcal{H}_1 \quad : \quad \boldsymbol{d}_k(n) \sim \mathcal{N}(\theta, \sigma^2), \tag{51}$$

where $\boldsymbol{d}_k(n)$ denotes the local datum collected by sensor k at time n. We assume the local statistic $\boldsymbol{x}_k(n)$ to be shared during the diffusion process is the log-likelihood ratio of the measurement $\boldsymbol{d}_k(n)$:

$$\boldsymbol{x}_k(n) = \frac{\theta}{\sigma^2}\left(\boldsymbol{d}_k(n) - \frac{\theta}{2}\right). \tag{52}$$

Note that in the Gaussian case the log-likelihood ratio is simply a shifted and scaled version of the collected observation $\boldsymbol{d}_k(n)$, such that no substantial differences are expected if the agents share directly the observations.

In the specific case that $\boldsymbol{x}_k(n)$ is the log-likelihood ratio, the expectations $\mathbb{E}_0\boldsymbol{x}$ and $\mathbb{E}_1\boldsymbol{x}$ assume a peculiar meaning. Indeed, they can be conveniently represented as:

$$\mathbb{E}_0\boldsymbol{x} = -\mathcal{D}(\mathcal{H}_0||\mathcal{H}_1), \quad \mathbb{E}_1\boldsymbol{x} = \mathcal{D}(\mathcal{H}_1||\mathcal{H}_0), \tag{53}$$

where $\mathcal{D}(\mathcal{H}_i||\mathcal{H}_j)$, with $i,j \in \{0,1\}$, is the Kullback-Leibler (KL) divergence between hypotheses i and j — see [40]. In particular, for the Gaussian shift-in-mean problem the distribution of the log-likelihood ratio can be expressed in terms of the KL divergences as follows:

$$\boldsymbol{x}_k(n) \overset{\mathcal{H}_0}{\sim} \mathcal{N}(-\mathcal{D}, 2\mathcal{D}), \qquad \boldsymbol{x}_k(n) \overset{\mathcal{H}_1}{\sim} \mathcal{N}(\mathcal{D}, 2\mathcal{D}), \tag{54}$$

where

$$\mathcal{D} \triangleq \mathcal{D}(\mathcal{H}_0||\mathcal{H}_1) = \mathcal{D}(\mathcal{H}_1||\mathcal{H}_0) = \frac{\theta^2}{2\sigma^2}, \tag{55}$$

is the KL divergence for the Gaussian shift-in-mean case [40].

Since the LMGF of a Gaussian random variable $\mathcal{N}(a,b)$ is $at + bt^2/2$ [42], we deduce from (54) that

$$\psi_0(t) = \mathcal{D}t(t-1), \quad \psi_1(t) = \mathcal{D}t(t+1). \tag{56}$$

Note that $\psi_1(t) = \psi_0(t+1)$, a relationship that holds true more generally when working with the LMGFs of the log-likelihood ratio — see, e.g., [55]. Now, applying (45) to (56) readily gives

$$\omega_0(t) = \mathcal{D}t\left(\frac{t}{2} - 1\right), \quad \omega_1(t) = \mathcal{D}t\left(\frac{t}{2} + 1\right). \tag{57}$$

According to its definition (32), in order to find the Fenchel-Legendre transform we should maximize, with respect to t, the function $\gamma t - \omega(t)$. In view of the convexity properties proved in Appendix C, this can be done by taking the first derivative and equating it to zero, which is equivalent to writing

$$\gamma \quad = \quad \omega_0'(t_0) = \frac{\psi_0(t_0)}{t_0} \Rightarrow t_0 = \frac{\gamma}{\mathcal{D}} + 1, \tag{58}$$

$$\gamma \quad = \quad \omega_1'(t_1) = \frac{\psi_1(t_1)}{t_1} \Rightarrow t_1 = \frac{\gamma}{\mathcal{D}} - 1. \tag{59}$$

These expressions lead to

$$\Omega_0(\gamma) = \frac{(\gamma+\mathcal{D})^2}{2\mathcal{D}}, \qquad \Omega_1(\gamma) = \frac{(\gamma-\mathcal{D})^2}{2\mathcal{D}}. \tag{60}$$

Selecting the threshold η within the interval $(-\mathcal{D}, \mathcal{D})$, the minimization in (47) is easily performed — refer to Fig. 3 and the related discussion. The final result is:

$$\boxed{\alpha_{k,\mu} \doteq e^{-(1/\mu)\, S\, \frac{(\eta+\mathcal{D})^2}{2\mathcal{D}}}, \qquad \beta_{k,\mu} \doteq e^{-(1/\mu)\, S\, \frac{(\eta-\mathcal{D})^2}{2\mathcal{D}}}} \tag{61}$$

These expressions provide the complete asymptotic characterization to the leading exponential order (i.e., they furnish the detection error exponents) of the adaptive distributed network of detectors for the Gaussian shift-in-mean problem, and for any choice of the threshold η within the interval $(-\mathcal{D}, \mathcal{D})$.

We have run a number of numerical simulations to check the validity of the results. Clearly, in order to show the generality of our methods, it is desirable to test them on non-Gaussian data as well. Since the interpretation of the results for both Gaussian and non-Gaussian data is essentially similar, we shall skip the numerical results for the Gaussian case to avoid unnecessary repetitions and focus on other cases. Accordingly, also the discussion on how to make a careful selection of the detection threshold η is postponed to the forthcoming sections.

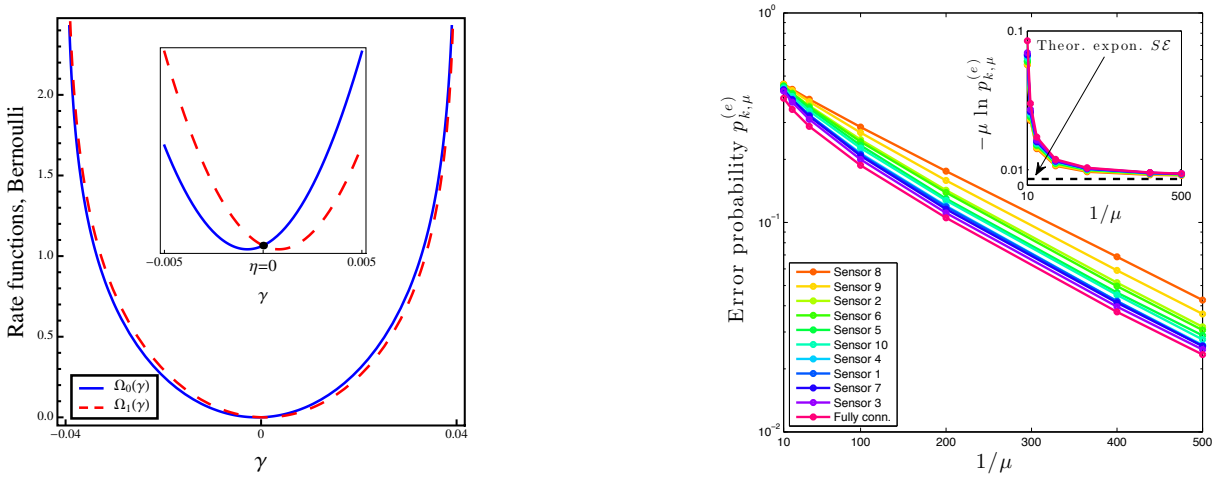

Fig. 5. Bernoulli example discussed in Sec. VI-B. We refer to the network in Fig. 4, and use detector (49) with $\eta = 0$. Leftmost panel: Rate functions. The dark circle in the close-up marks the employed detection threshold, which is relevant to error exponent evaluation. Rightmost panel: Steady-state error probabilities at different sensors, obtained via Monte Carlo simulation. For comparison purposes, the empirical error probabilities of the fully connected system are reported. The solid curves in the inset plot represent the empirical error exponents $-\mu \ln p_{k,\mu}^{(e)}$, for $k = 1, 2, \ldots, S$, while the dashed horizontal line is the exponent $S\,\mathcal{E}$ predicted by our large deviations analysis (Theorem 4). The parameters of the considered detection problem are $p_0 = 0.49$ and $p_1 = 0.51$. The number of Monte Carlo runs is 10^5.

B. Hardly (one-bit) Quantized Measurements

We now examine the example in which the measurements at the local sensors are hardly quantized. This situation can be formalized as the following hypothesis test:

$$\mathcal{H}_0 \quad : \quad \boldsymbol{d}_k(n) \sim \mathcal{B}(p_0), \tag{62}$$

$$\mathcal{H}_1 \quad : \quad \boldsymbol{d}_k(n) \sim \mathcal{B}(p_1), \tag{63}$$

with $\mathcal{B}(p)$ denoting a Bernoulli random variable with success probability p. As in the previous example, we assume that the local statistics $\boldsymbol{x}_k(n)$ employed by the sensors in the adaptation/combination stages are chosen as the local log-likelihood ratios that, in view of (62)–(63), can be written as:

$$\boldsymbol{x}_k(n) = \boldsymbol{d}_k(n) \ln\left(\frac{p_1}{p_0}\right) + (1 - \boldsymbol{d}_k(n)) \ln\left(\frac{q_1}{q_0}\right), \tag{64}$$

where $q_h = 1 - p_h$, with $h = 0, 1$. Since $\boldsymbol{d}_k(n) \in \{0, 1\}$, we see that $\boldsymbol{x}_k(n)$ is a binary random variable taking on the values $\ln(p_1/p_0)$ or $\ln(q_1/q_0)$. The distribution of $\boldsymbol{x}_k(n)$ is then characterized by:

$$\mathbb{P}_0\left[\boldsymbol{x}_k(n) = \ln\left(\frac{p_1}{p_0}\right)\right] = p_0, \; \mathbb{P}_1\left[\boldsymbol{x}_k(n) = \ln\left(\frac{p_1}{p_0}\right)\right] = p_1, \tag{65}$$

and, hence, the LMGFs for this example are readily computed:

$$\psi_0(t) = \ln\left(\frac{p_1^t}{p_0^{t-1}} + \frac{q_1^t}{q_0^{t-1}}\right), \tag{66}$$

$$\psi_1(t) = \ln\left(\frac{p_1^{t+1}}{p_0^t} + \frac{q_1^{t+1}}{q_0^t}\right). \tag{67}$$

According to the relationship (45) found in Theorem 4, these closed-form expressions are used for the evaluation of $\omega_0(t)$ and $\omega_1(t)$, which in turn are needed to compute the rate functions $\Omega_0(\gamma)$ and $\Omega_1(\gamma)$. Differently from the Gaussian example, here these tasks need to be performed numerically. The resulting rate functions are displayed in the leftmost panel of Fig. 5, and the observed behavior reproduces what is predicted by the general properties of the rate function — see also the explanation of Fig. 2.

Let us now examine the adaptive distributed network of detectors in operation. To do so, we must decide on how to set the detection threshold η in (49). As a method for selecting the threshold, in this section we illustrate the asymptotic Bayesian criterion that prescribes maximizing the exponent of the average error probability

$$p_{k,\mu}^{(e)} \triangleq \pi_0 \alpha_{k,\mu} + \pi_1 \beta_{k,\mu}, \tag{68}$$

where π_0 and π_1 are the prior probabilities of occurrence of hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$, respectively. It is easily envisaged that the exponent of the average error probability is determined by the worst one (slowest decay) between the Type-I and Type-II error exponents — see [56, Eq. (I.2), p. 4]. As a result, optimizing the Bayesian error exponent is equivalent to a max-min approach aimed at maximizing the minimum exponent. We now apply this criterion to the considered example. To this aim, a close inspection of the rate functions in Fig. 5 is beneficial. First, as it can be seen by the close-up shown in the inset plot, setting the threshold to $\eta = 0$ would imply

$$\mathcal{E}_0 = \inf_{\gamma > 0} \Omega_0(\gamma) = \Omega_0(0) = \Omega_1(0) = \inf_{\gamma \leq 0} \Omega_1(\gamma) = \mathcal{E}_1 \triangleq \mathcal{E}. \tag{69}$$

Moreover, any other choice of the threshold $\eta \neq 0$ makes one of the two exponents smaller than $\mathcal{E}$. This can be clearly visualized by varying the position of η in Fig. 3, and computing the infima over the pertinent decision regions. In summary, according to whether we adopt a Bayesian or a max-min criterion, an optimal choice for the threshold in this case is $\eta = 0$.

In the simulations, we refer to a sufficiently large time horizon, such that the steady-state assumption applies, and evaluate the error probabilities for different values of the step-size — see the rightmost panel in Fig. 5. In the considered

example, it is easily verified by symmetry arguments that the error probabilities (and not only the exponents) of first and second kind are equal, and therefore they equal the average error probability for any prior distribution of the hypotheses:

$$\alpha_{k,\mu} = \beta_{k,\mu} = p^{(e)}_{k,\mu}. \tag{70}$$

Accordingly, in the following description the terminologies "error probability" and "error exponent" can be equivalently and unambiguously referred to any of these errors.

In Fig. 5, rightmost panel, the performance of all the agents is displayed as a function of $1/\mu$, and different agents are marked with different colors. For comparison purposes, the performance of the fully connected system is also displayed. All these probability curves have been obtained by Monte Carlo simulation. Some remarkable features are observed.

First, all the different curves pertaining to different agents stay nearly parallel for sufficiently small values of the step-size μ. This is a way to visualize that $i)$ the detection error probabilities vanish exponentially at rate $1/\mu$; and $ii)$ the detection error *exponents* at different sensors are equal, and further equal to that of the fully-connected system corresponding to the centralized stochastic gradient solution. This is the basic message conveyed by the large deviations analysis. Indeed, the asymptotic relationships for the error probabilities in (6) express convergence *to the first leading order in the exponent*.

It remains to show that the *exponents* of the simulated error probabilities match the *exponents* predicted by Theorem 4. This is made in the inset plot of Fig. 5, rightmost panel, where the horizontal dashed line depicts the theoretical exponent $S\mathcal{E}$, with $\mathcal{E}$ computed using (69), while the solid curves represent the empirical error exponents seen at different sensors, namely the quantities $-\mu \ln p^{(e)}_{k,\mu}$, for $k = 1, 2, \ldots, S$. It is observed that, as the step-size decreases, the empirical error exponents converge towards the theoretical one $S\,\mathcal{E}$.

A further interesting evidence seems to emerge from the numerical experiments. The error probability curves in Fig. 5, rightmost panel, are basically ordered. Examining the relationship between this ordering and the sensor placement in Fig. 4, it is seen that the ordering reflects the degree of connectivity of each agent. For instance, sensor 3 has the highest number of neighbors (five), and its performance is the closest to the fully connected case. On the other hand, sensor 8 is the most isolated, and its error probability curve appears accordingly the highest one. Note that, since from the presented theory we learned that each agent reaches asymptotically the same detection exponent, these differences are related to higher order corrections (i.e., sub-exponential terms that are neglected in a large deviations analysis) and/or to non-asymptotic effects. A systematic and thorough analysis of the above features, as well as of their exact interplay with the network connectivity and more in general with the overall structure of the connectivity matrix A, requires a refined asymptotic estimate that goes beyond the large deviations analysis carried out here.

C. Shift-in-mean with Laplacian noise

In this section we consider another non-Gaussian example, namely, the case of a shift-in-mean detection problem with noise distributed according to a Laplace distribution. Denoting by $\mathcal{L}(a, b)$ a (shifted) Laplace distribution with shift parameter a and scale parameter b, i.e., having the probability density function:

$$f_L(\xi) = \frac{1}{2b} e^{-\frac{|\xi - a|}{b}}, \tag{71}$$

the hypothesis test we are now interested in is formulated as follows:

$$\mathcal{H}_0 \quad : \quad \boldsymbol{d}_k(n) \sim \mathcal{L}(0, \sigma), \tag{72}$$

$$\mathcal{H}_1 \quad : \quad \boldsymbol{d}_k(n) \sim \mathcal{L}(\theta, \sigma). \tag{73}$$

We assume again that the local statistics $\boldsymbol{x}_k(n)$ are chosen as the local log-likelihood ratios:

$$\boldsymbol{x}_k(n) = \frac{1}{\sigma}(|\boldsymbol{d}_k(n)| - |\boldsymbol{d}_k(n) - \theta|). \tag{74}$$

Then, the LMGFs for this case can be computed in closed form [5], and are given by:

$$\psi_0(t) \;=\; \ln\left(\frac{1-t}{1-2t}\, e^{-\rho\, t} - \frac{t}{1-2t} e^{-\rho\,(1-t)}\right), \tag{75}$$

$$\psi_1(t) \;=\; \ln\left(\frac{1+t}{1+2t}\, e^{\rho\, t} + \frac{t}{1+2t} e^{-\rho\,(1+t)}\right), \tag{76}$$

where we defined $\rho = \theta/\sigma$, and where, by limit arguments, we have $\psi_0(1/2) = \psi_1(-1/2) = -\rho/2 + \ln(1 + \rho/2)$.

As done before, we can use the above expressions in (45), for performing numerical evaluation of $\omega_0(t)$ and $\omega_1(t)$, and of their Fenchel-Legendre transforms $\Omega_0(\gamma)$ and $\Omega_1(\gamma)$, which are displayed in Fig. 6, leftmost panel.

Differently from the previous section, we now consider an alternative threshold setting, which is grounded on the well-known Neyman-Pearson criterion [42]. Its classical (asymptotic) formulation sets a maximum tolerable value for the false-alarm probability, and examines the decaying rate of the miss-detection probability (the role of the two errors can also be reversed). The main difference in relation to the setup considered so far is that we relax the condition that the Type-I error probability vanish exponentially, and this allows in general for a gain in terms of the Type-II error exponent. The procedure for the Neyman-Pearson threshold setting has been already described in Sec. IV — see (25)–(26). Accordingly, to achieve a false-alarm probability $\bar{\alpha}$, we need a threshold

$$\eta = \eta_\mu = \mathbb{E}_0 \boldsymbol{x} + \sqrt{\frac{\mu \sigma^2_{x,0}}{2S}}\, Q^{-1}(\bar{\alpha}). \tag{77}$$

It remains to evaluate the Type-II error probability

$$\beta_{k,\mu} = \mathbb{P}_1[\boldsymbol{y}^\star_{k,\mu} \leq \eta_\mu], \tag{78}$$

or, more precisely, the corresponding exponent $\mathcal{E}_1$. For this purpose, we must resort to Theorem 4. Note, however, that the threshold $\eta = \eta_\mu$ now depends on μ and, hence, Theorem 4 does not directly apply. As noted in Remark II, it is instructive to examine how the result of Theorem 4 can be generalized to manage similar situations. Indeed, we can work in terms of the shifted variables

$$\widehat{\boldsymbol{y}}^\star_{k,\mu} = \boldsymbol{y}^\star_{k,\mu} - \sqrt{\frac{\mu \sigma^2_{x,0}}{2S}}\, Q^{-1}(\bar{\alpha}), \tag{79}$$

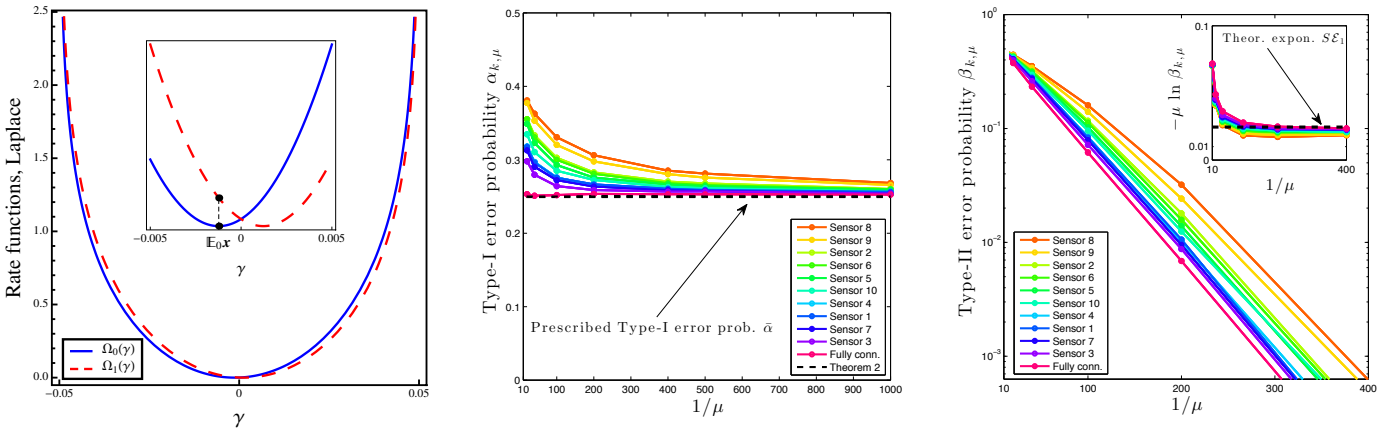

Fig. 6. Laplace example discussed in Sec. VI-C. We refer to the network in Fig. 4, and use the Neyman-Pearson detector with threshold (77), for two values of the desired false-alarm level $\bar{\alpha}$. Leftmost panel: Rate functions. The dark circle in the close-up marks the abscissa $\eta = \mathbb{E}_0\boldsymbol{x}$, which is relevant for computing the Type-II error exponent. Middle panel: Solid curves refer to the empirical steady-state Type-I error probabilities at different sensors, obtained via Monte Carlo simulation. For comparison purposes, the empirical error probabilities of the fully connected system are reported. The dashed horizontal lines pertain to the theoretical Type-I error probabilities obtained by the normal approximation (Theorem 2). Rightmost panel: Steady-state Type-II error probabilities at different sensors, along with the performance of the fully connected case. The solid curves in the inset plot represent the empirical Type-II error exponent $-\mu \ln \beta_{k,\mu}$, for $k = 1, 2, \ldots, S$, while the dashed horizontal line is the exponent predicted by our large deviations analysis (Theorem 4). The parameters of the considered detection problem are $\theta = 0.05$ and $\sigma = 1$. The number of Monte Carlo runs is 10^5.

yielding

$$\beta_{k,\mu} = \mathbb{P}_1[\widehat{\boldsymbol{y}}^{\star}_{k,\mu} \leq \mathbb{E}_0\boldsymbol{x}]. \tag{80}$$

By application of the Gärtner-Ellis Theorem to the shifted variables $\widehat{\boldsymbol{y}}^{\star}_{k,\mu}$, it is easy to see that the added deterministic term (vanishing with μ) does not alter the limiting function $\omega_1(t)$ in (45), and consequently the final rate function $\Omega_1(\gamma)$. Accordingly, and based on (80), the Type-II error exponent is

$$\mathcal{E}_1 = \inf_{\gamma \leq \mathbb{E}_0\boldsymbol{x}} \Omega_1(\gamma) = \Omega_1(\mathbb{E}_0\boldsymbol{x}). \tag{81}$$

The main implication of the above result can be understood, e.g., by examining the close-up in the leftmost panel of Fig. 6, where it is seen that:

$$\mathcal{E}_1 = \Omega_1(\mathbb{E}_0\boldsymbol{x}) > \Omega_1(0), \tag{82}$$

the latter value being the Type-II error exponent achieved by the max-min optimal detector with zero threshold previously described. This immediately shows the gain achieved by relaxing the constraint that *both* error probabilities must vanish exponentially.

We now present the numerical evidence for the Neyman-Pearson adaptive distributed detector. The middle panel in Fig. 6 shows the convergence of $\alpha_{k,\mu}$ towards the prescribed Type-I error probability $\bar{\alpha}$ as the step-size μ goes to zero. The rightmost panel refers instead to the corresponding Type-II error probability curves. The conclusions that can be drawn are similar to those discussed in the previous example, confirming the validity of the theoretical analysis. It is also interesting to note that the ordering of the different curves, for both error probabilities, is exactly the same obtained in the Bernoulli example. Since the network employed for the simulations is unchanged, this is another clue that the ordering may be related to the structure of the connectivity matrix A.

D. Shift-in-mean with Gaussian mixture noise

As a final example, we consider the case of a shift-in-mean detection problem with noise distributed according to a zero-mean Gaussian mixture, having the probability density function

$$f_{GM}(\xi) = \frac{1}{2}\left(\frac{1}{\sqrt{2\pi b_1}} e^{-\frac{(\xi-a_0)^2}{2b_1}} + \frac{1}{\sqrt{2\pi b_2}} e^{-\frac{(\xi+a_0)^2}{2b_2}}\right), \tag{83}$$

namely, a balanced mixture of normal random variables with different variances b_1 and b_2, and symmetric expectations $\pm a_0$. Denoting by $\mathcal{N}_{mix}(a, a_0, b_1, b_2)$ a *shifted* Gaussian mixture distribution with shift parameter a, we consider the following hypothesis test:

$$\mathcal{H}_0 \;:\; \boldsymbol{d}_k(n) \sim \mathcal{N}_{mix}(0, \theta_0, \sigma_1^2, \sigma_2^2), \tag{84}$$
$$\mathcal{H}_1 \;:\; \boldsymbol{d}_k(n) \sim \mathcal{N}_{mix}(\theta, \theta_0, \sigma_1^2, \sigma_2^2). \tag{85}$$

For this model, we do *not* assume that the local statistics $\boldsymbol{x}_k(n)$ are chosen as the local log-likelihood ratios. We assume instead that the agents of the network have scarce knowledge about the underlying statistical model. They know that it is a shift-in-mean problem, and possess a rough information about the value of θ. In these circumstances, the agents decide to implement a distributed sample-mean detector, namely, they exchange the local measurements *as they are, without any additional pre-processing*. This amounts to state that

$$\mathcal{H}_0 \;:\; \boldsymbol{x}_k(n) \sim \mathcal{N}_{mix}(0, \theta_0, \sigma_1^2, \sigma_2^2), \tag{86}$$
$$\mathcal{H}_1 \;:\; \boldsymbol{x}_k(n) \sim \mathcal{N}_{mix}(\theta, \theta_0, \sigma_1^2, \sigma_2^2). \tag{87}$$

Then, the LMGFs for this case can be computed in closed form [5], and are given by:

$$\psi_0(t) = \ln\left(\frac{1}{2}e^{\theta_0 t + \frac{\sigma_1^2 t^2}{2}} + \frac{1}{2}e^{-\theta_0 t + \frac{\sigma_2^2 t^2}{2}}\right), \tag{88}$$
$$\psi_1(t) = \theta t + \psi_0(t). \tag{89}$$

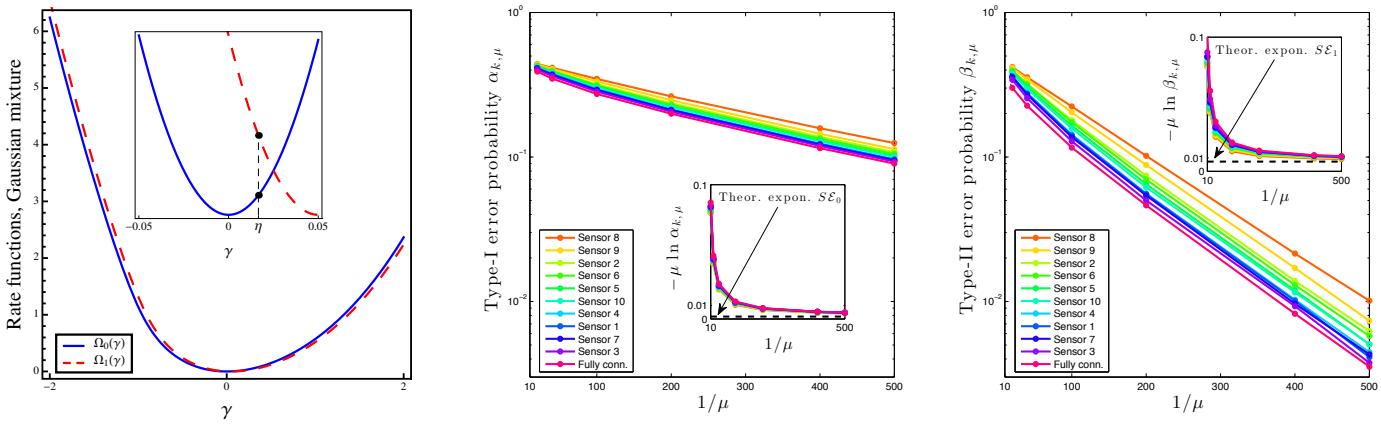

Fig. 7. Gaussian mixture example discussed in Sec. VI-D. We refer to the network in Fig. 4, and use detector (49) with $\eta = \theta/3$. Leftmost panel: Rate functions. The dark circle in the close-up marks the employed detection threshold, which is relevant for evaluating the error exponents. Middle panel: Steady-state Type-I error probabilities at different sensors, obtained via Monte Carlo simulation. For comparison purposes, the empirical error probabilities of the fully connected system are reported. The solid curves in the inset plot represent the empirical Type-I error exponent $-\mu \ln \alpha_{k,\mu}$, for $k = 1, 2, \ldots, S$, while the dashed horizontal line is the exponent predicted by our large deviations analysis (Theorem 4). Rightmost panel: Same of middle panel, but for the Type-II error. The parameters of the considered detection problem are $\theta = 0.05$, $\theta_0 = 1$, $\sigma_1 = 1$, and $\sigma_2 = 0.3$. The number of Monte Carlo runs is 10^5.

The above expressions are used in (45) for evaluating numerically $\omega_0(t)$ and $\omega_1(t)$, and then their Fenchel-Legendre transforms $\Omega_0(\gamma)$ and $\Omega_1(\gamma)$. These latter are depicted in the leftmost panel of Fig. 7. We assume the agents in the network are not able to optimize the choice of the detection threshold, due to their limited knowledge of the underlying statistical models. The particular value used in the simulations is $\eta = \theta/3$, which is marked in the close-up of Fig. 7, leftmost panel. It is seen that, differently from the previous examples, this choice does not correspond to a balancing of the detection error exponents, such that it is expected that the Type-I and Type-II error probabilities behave quite differently in this case. This is clearly observed in the middle (Type-I error) and rightmost (Type-II error) panels of Fig. 7. The numerical evidence confirms the theoretical predictions, as well as the essential features found in all the previous examples. Moreover, it is seen that the enhanced decaying rate of the Type-II error probability arising from the unbalanced threshold setting is paid in terms of a higher Type-I error probability.

E. Adaptation and detection

In the simulation results illustrated so far, we focused on the system performance at steady-state. It is of great interest to consider also the *time-evolution* of the system performance, and even more to show the system at work in a *dynamic* situation where the true hypothesis is changing over time, which is truly the main motivation for an *adaptive* framework.

To this aim, we return to the kind of situation described in Fig. 1, which is now re-examined in more quantitative terms by focusing on the actual error probabilities, rather than on the time-evolution of the detection statistics. Specifically, in Fig. 8 we display the performance of three generic agents of the network, for two values of the step-size. For comparison purposes, we show also the performance of the running consensus algorithm [1]–[6]. The underlying statistical model is the shift-in-mean with Laplacian noise detailed in Sec. VI-C, and we employ a zero-threshold detector.

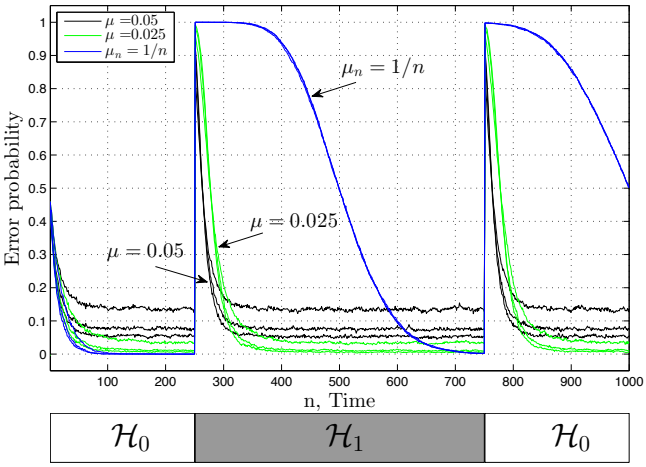

Fig. 8. Pictorial summary of adaptive diffusion for detection, with reference to the Laplace example discussed in Sec. VI-C. Top panel: time-evolution of the error probability at a local node with i) the diffusion strategy with different step-sizes $\mu = 0.025, 0.05$, and ii) the running consensus strategy (diminishing step-size $\mu_n = 1/n$). Actual variation of the true hypothesis is depicted in the bottom panel. The parameters of the considered detection problem are $\theta = 0.3$ and $\sigma = 1$.

First, the inference/adaptation trade-off is emphasized: smaller values of μ allow better inference (lower values of the steady-state error probabilities), at the cost of increasing the time for reliably learning that a change occurred. In this respect, the running consensus performance represents an extreme case: indeed, here the step-size is vanishing, i.e., $\mu_n = 1/n$, which explains the bad performance in terms of adaptation exhibited in Fig. 1.

VII. Concluding Remarks and Open Issues

The asymptotic tools developed in this paper allow designing and characterizing the performance of network detectors that are *adaptive and decentralized*. We show that the steady-state detection error probabilities of each individual agent

decrease exponentially with the inverse of the step-size and that cooperation among sensors makes the error exponents governing such decay equal to that of a centralized stochastic gradient solution. Closed-form expressions are derived, giving insights about the main scaling laws with respect to the fundamental system parameters.

In our treatment, we studied the detection performance of the diffusion strategy, *given a certain local statistic* $\boldsymbol{x}$. Our findings show that the steady-state observable, as well as its detection performance, in general depend upon the kind of transmitted data $\boldsymbol{x}$. A plausible, though heuristic, choice for $\boldsymbol{x}$ is that of the log-likelihood ratio of the measured data. However, the problem of choosing the *best* statistic is open, and we feel that the obtained results can assist in exploring the relationship between the asymptotic performance and the choice of an optimal statistic $\boldsymbol{x}$.

We would like to finally note that in order to avoid a prohibitive number of Monte-Carlo runs, the simulations in the previous section were run in the small signal-to-noise ratio regime, where the error probabilities need not be too small. In this regime, the exact rate functions could in principle be replaced by parabolic approximations (see, e.g., the leftmost plot in Fig. 5) and a parabolic approximation is basically a Gaussian approximation. To avoid confusion, we note that the results of this work do not require any small signal-to-noise ratio assumption; they hold in greater generality. Moreover, using a Gaussian approximation will generally lead to a wrong error exponent. For the same reason of avoiding prohibitive simulation runs in the convergence analysis of the Type-II error exponent, the Type-I error probability for the Neyman-Pearson setting of Fig. 6 was set to $\bar{\alpha} = 1/4$ (rather than to much smaller values) and used to illustrate the theoretical findings against the simulated curves.

Appendix A: Proof of Theorem 2

Since the transient term in (11) does not affect the limiting behavior of $\boldsymbol{y}_k(n)$, it suffices to focus on the limiting behavior of the summations in (15). We introduce accordingly the following finite-horizon variable:

$$\boldsymbol{y}_k^\star(n) \triangleq \sum_{i=1}^{n} \boldsymbol{z}_k(i) \;=\; \sum_{i=1}^{n}\sum_{\ell=1}^{S} \mu(1-\mu)^{i-1} b_{k,\ell}(i) \boldsymbol{x}'_\ell(i). \quad (90)$$

Since $\boldsymbol{y}_k^\star(n)$ converges in distribution to $\boldsymbol{y}_{k,\mu}^\star$ as $n \to \infty$, by Lévy's continuity Theorem [49], the corresponding characteristic functions must converge as well. It is convenient to work in terms of the normalized variable:

$$\widetilde{\boldsymbol{y}}_{k,\mu}^\star = \frac{\boldsymbol{y}_{k,\mu}^\star - \mathbb{E}\boldsymbol{x}}{\sqrt{\mu\,\sigma_x^2/(2S)}}. \quad (91)$$

Denoting by $\varphi_{k,\mu}(t)$ the characteristic function of $\widetilde{\boldsymbol{y}}_{k,\mu}^\star$, using (90) and (91), and taking the limit as $n \to \infty$, we have:

$$\varphi_{k,\mu}(t) = \mathbb{E}e^{jt\widetilde{\boldsymbol{y}}_{k,\mu}^\star} = \prod_{i=1}^{\infty}\prod_{\ell=1}^{S} \mathbb{E}e^{jt\widetilde{\boldsymbol{x}}'_\ell(i)\zeta_{i,\ell}}, \quad (92)$$

defined in terms of the non-random variable

$$\zeta_{i,\ell} = \sqrt{2S\mu}(1-\mu)^{i-1} b_{k,\ell}(i), \quad (93)$$

and the centered and normalized random variable

$$\widetilde{\boldsymbol{x}}'_\ell(i) = \frac{\boldsymbol{x}'_\ell(i) - \mathbb{E}\boldsymbol{x}}{\sigma_x}. \quad (94)$$

Now, the claim of asymptotic normality in (21) can be proven by showing the convergence, as $\mu \to 0$, of $\varphi_{k,\mu}(t)$ towards the characteristic function of the standard normal distribution, $e^{-\frac{t^2}{2}}$. It suffices to work with $t > 0$ to verify the validity of this latter property. Formally, we would like to show that the quantity:

$$\left|\varphi_{k,\mu}(t) - e^{-\frac{t^2}{2}}\right| = \left|\prod_{i=1}^{\infty}\prod_{\ell=1}^{S} \mathbb{E}e^{jt\widetilde{\boldsymbol{x}}'_\ell(i)\zeta_{i,\ell}} - e^{-\frac{t^2}{2}}\right| \quad (95)$$

converges to zero as $\mu \to 0$. To this aim, we start by working with a finite n, and write:

$$\begin{aligned}
&\left|\prod_{i=1}^{n}\prod_{\ell=1}^{S} \mathbb{E}e^{jt\widetilde{\boldsymbol{x}}'_\ell(i)\zeta_{i,\ell}} - e^{-\frac{t^2}{2}}\right| \\
&\quad\le \left|\prod_{i=1}^{n}\prod_{\ell=1}^{S} \mathbb{E}e^{jt\widetilde{\boldsymbol{x}}'_\ell(i)\zeta_{i,\ell}} - \prod_{i=1}^{n}\prod_{\ell=1}^{S} e^{-\frac{t^2\zeta_{i,\ell}^2}{2}}\right| \\
&\quad+ \left|\prod_{i=1}^{n}\prod_{\ell=1}^{S} e^{-\frac{t^2\zeta_{i,\ell}^2}{2}} - e^{-\frac{t^2}{2}}\right|. \quad (96)
\end{aligned}$$

We first focus on the first term on the RHS of (96). For complex w_i, z_i, with $|w_i| \le 1$ and $|z_i| \le 1$, it is known that [49]:

$$\left|\prod_{i=1}^{n} w_i - \prod_{i=1}^{n} z_i\right| \le \sum_{i=1}^{n} |w_i - z_i|. \quad (97)$$

Since $\mathbb{E}e^{jt\widetilde{\boldsymbol{x}}'_\ell(i)\zeta_{i,\ell}}$ is a characteristic function, its magnitude is not greater than one [49], such that it is legitimate to write, in view of (97):

$$\begin{aligned}
&\left|\prod_{i=1}^{n}\prod_{\ell=1}^{S} \mathbb{E}e^{jt\widetilde{\boldsymbol{x}}'_\ell(i)\zeta_{i,\ell}} - \prod_{i=1}^{n}\prod_{\ell=1}^{S} e^{-\frac{t^2\zeta_{i,\ell}^2}{2}}\right| \\
&\qquad\le \sum_{i=1}^{n}\sum_{\ell=1}^{S}\left|\mathbb{E}e^{jt\widetilde{\boldsymbol{x}}'_\ell(i)\zeta_{i,\ell}} - e^{-\frac{t^2\zeta_{i,\ell}^2}{2}}\right|. \quad (98)
\end{aligned}$$

The single summand on the right-hand side of the above expression is upper bounded by

$$\left|\mathbb{E}e^{jt\widetilde{\boldsymbol{x}}'_\ell(i)\zeta_{i,\ell}} - 1 + \frac{t^2\zeta_{i,\ell}^2}{2}\right| + \left|e^{-\frac{t^2\zeta_{i,\ell}^2}{2}} - 1 + \frac{t^2\zeta_{i,\ell}^2}{2}\right|. \quad (99)$$

Using the fact that $\mathbb{E}\widetilde{\boldsymbol{x}}'_\ell(i) = 0$ and $\mathbb{E}[\widetilde{\boldsymbol{x}}'_\ell(i)]^2 = 1$, we can further bound the first term in the above expression as

$$\begin{aligned}
&\left|\mathbb{E}e^{jt\widetilde{\boldsymbol{x}}'_\ell(i)\zeta_{i,\ell}} - 1 + \frac{t^2\zeta_{i,\ell}^2}{2}\right| \\
&\quad= \left|\mathbb{E}\left(e^{jt\widetilde{\boldsymbol{x}}'_\ell(i)\zeta_{i,\ell}} - 1 - j\widetilde{\boldsymbol{x}}'_\ell(i)t\zeta_{i,\ell} + [\widetilde{\boldsymbol{x}}'_\ell(i)]^2\frac{t^2\zeta_{i,\ell}^2}{2}\right)\right| \\
&\quad\le \mathbb{E}|\widetilde{\boldsymbol{x}}'_\ell(i)|^3\,\frac{t^3\zeta_{i,\ell}^3}{6}, \quad (100)
\end{aligned}$$

where the last inequality follows from upper bounding the remainder of the Taylor expansion of the complex exponential:

$$\left|e^{jt} - 1 - \frac{jt}{1!} - \cdots - \frac{(jt)^{n-1}}{(n-1)!}\right| \le \frac{|t|^n}{n!}. \quad (101)$$

Likewise, the second term in (99) is upper bounded by $\frac{t^4\zeta_{i,\ell}^4}{8}$ since $|e^{-s}-1+s| \leq s^2/2$ for any $s \geq 0$.

We can accordingly rewrite (96) as:

$$\begin{aligned}\left|\prod_{i=1}^{n}\prod_{\ell=1}^{S}\mathbb{E}e^{jt\widetilde{\boldsymbol{x}}'_\ell(i)\zeta_{i,\ell}} - e^{-\frac{t^2}{2}}\right| &\leq \mathbb{E}|\widetilde{\boldsymbol{x}}'_\ell(i)|^3\,\frac{t^3}{6}\sum_{i=1}^{n}\sum_{\ell=1}^{S}\zeta_{i,\ell}^3 \\ &+ \frac{t^4}{8}\sum_{i=1}^{n}\sum_{\ell=1}^{S}\zeta_{i,\ell}^4 \\ &+ \left|e^{-\frac{t^2}{2}\sum_{i=1}^{n}\sum_{\ell=1}^{S}\zeta_{i,\ell}^2} - e^{-\frac{t^2}{2}}\right|.\end{aligned} \tag{102}$$

We now take the limit as $n \to \infty$ in the above expression. To this aim, observe that, by the definition (93), the summation:

$$\sum_{i=1}^{n}\sum_{\ell=1}^{S}\zeta_{i,\ell}^m, \qquad m = 1, 2, \ldots \tag{103}$$

is made of nonnegative terms, and is upper bounded by a convergent geometric series, since $b_{k,\ell} \leq 1$. This implies the convergence of the series (103) as $n \to \infty$. Accordingly, taking the limit as $n \to \infty$ in (102), and using (95), we have:

$$\begin{aligned}\left|\varphi_{k,\mu}(t) - e^{-\frac{t^2}{2}}\right| &= \left|\prod_{i=1}^{\infty}\prod_{\ell=1}^{S}\mathbb{E}e^{jt\widetilde{\boldsymbol{x}}'_\ell(i)\zeta_{i,\ell}} - e^{-\frac{t^2}{2}}\right| \\ &\leq \mathbb{E}|\widetilde{\boldsymbol{x}}'_\ell(i)|^3\,\frac{t^3}{6}\sum_{i=1}^{\infty}\sum_{\ell=1}^{S}\zeta_{i,\ell}^3 \\ &+ \frac{t^4}{8}\sum_{i=1}^{\infty}\sum_{\ell=1}^{S}\zeta_{i,\ell}^4 \\ &+ \left|e^{-\frac{t^2}{2}\sum_{i=1}^{\infty}\sum_{\ell=1}^{S}\zeta_{i,\ell}^2} - e^{-\frac{t^2}{2}}\right|.\end{aligned} \tag{104}$$

According to the latter relationships, in order to show that $\left|\varphi_{k,\mu}(t) - e^{-\frac{t^2}{2}}\right|$ converges to zero as $\mu \to 0$, it suffices to verify that:

$$\sum_{i=1}^{\infty}\sum_{\ell=1}^{S}\zeta_{i,\ell}^m \overset{\mu\to 0}{\longrightarrow} 0, \qquad m = 3, 4, \tag{105}$$

$$\sum_{i=1}^{\infty}\sum_{\ell=1}^{S}\zeta_{i,\ell}^2 \overset{\mu\to 0}{\longrightarrow} 1. \tag{106}$$

A technical remark is useful at this stage. Given the assumption of finite absolute third moment, there exists a simpler way to prove our claim, relying on the celebrated Berry-Esseen theorems [49, p. 542]. Such technique would directly reduce our proof to the verification of properties such as (105) and (106), without the preliminary work with characteristic functions. However, we prefer to offer here a more general proof, which might be useful to obtain future generalizations where the condition about the third moment could be weakened.

The key for proving (105) and (106) is Perron's Theorem, which provides a uniform bound on the convergence rate of the matrix $B_n = A^n$ — see [48, Th. 8.5.1]. Let λ_2 be the second largest magnitude eigenvalue of A. For any positive λ such that $|\lambda_2| < \lambda < 1$, there exists a positive constant $\mathcal{C} = \mathcal{C}(\lambda, A)$, ensuring for all i, k and ℓ:

$$\left|b_{k,\ell}(i) - \frac{1}{S}\right| \leq \mathcal{C}\lambda^i. \tag{107}$$

The above result follows by noting that the largest magnitude eigenvalue of the difference matrix $B_n - (1/S)\,\mathbb{1}\mathbb{1}^T$ is λ_2, and by applying the result on the convergence rate in [48, Corollary 5.6.13].

According to the above discussion, let us modify the variables $\zeta_{i,\ell}$ by replacing the matrix entries $b_{k,\ell}(i)$ with their limit $1/S$, namely,

$$\widetilde{\zeta}_{i,\ell} = \sqrt{2S\mu}(1-\mu)^{i-1}\frac{1}{S}, \tag{108}$$

and introduce, for any integer $m \geq 2$, the absolute difference:

$$\begin{aligned}\left|\sum_{i=1}^{\infty}\sum_{\ell=1}^{S}\zeta_{i,\ell}^m - \sum_{i=1}^{\infty}\sum_{\ell=1}^{S}\widetilde{\zeta}_{i,\ell}^m\right| &\leq \sum_{i=1}^{\infty}\sum_{\ell=1}^{S}|\zeta_{i,\ell}^m - \widetilde{\zeta}_{i,\ell}^m| \\ &= (2S\mu)^{m/2}\sum_{i=1}^{\infty}\sum_{\ell=1}^{S}(1-\mu)^{m(i-1)}\left|b_{k,\ell}^m(i) - \frac{1}{S^m}\right|.\end{aligned} \tag{109}$$

Recalling the factorization

$$a^m - b^m = (a-b)\sum_{k=0}^{m-1}a^k b^{m-1-k}, \tag{110}$$

(which can be proved, for $a \neq b$, by using the geometric sum $\sum_{k=0}^{m-1} r^k = \frac{1-r^m}{1-r}$, and using $r = a/b$), along with the fact that $b_{k,\ell}(i)$ and $1/S$ are not greater than one, we conclude that

$$\left|b_{k,\ell}^m(i) - \frac{1}{S^m}\right| \leq m\left|b_{k,\ell}(i) - \frac{1}{S}\right|, \tag{111}$$

yielding

$$\begin{aligned}&\left|\sum_{i=1}^{\infty}\sum_{\ell=1}^{S}\zeta_{i,\ell}^m - \sum_{i=1}^{\infty}\sum_{\ell=1}^{S}\widetilde{\zeta}_{i,\ell}^m\right| \\ &\leq m(2S\mu)^{m/2}\sum_{i=1}^{\infty}\sum_{\ell=1}^{S}(1-\mu)^{m(i-1)}\left|b_{k,\ell}(i) - \frac{1}{S}\right| \\ &\leq \mathcal{C}\lambda m(2S\mu)^{m/2}\sum_{i=1}^{\infty}\sum_{\ell=1}^{S}(1-\mu)^{m(i-1)}\lambda^{i-1} \\ &= \mathcal{C}\lambda m 2^{m/2}S^{m/2+1}\frac{\mu^{m/2}}{1-\lambda(1-\mu)^m} \overset{\mu\to 0}{\longrightarrow} 0,\end{aligned} \tag{112}$$

where the second inequality follows from (107), and the limit holds because $\lambda < 1$. In view of the above result, in order to establish (105) and (106) it is enough to study the limiting behavior of the summation:

$$\begin{aligned}\sum_{i=1}^{\infty}\sum_{\ell=1}^{S}\widetilde{\zeta}_{i,\ell}^m &= \frac{(2\mu)^{m/2}}{S^{m/2-1}}\sum_{i=1}^{\infty}(1-\mu)^{m(i-1)} \\ &= \frac{2^{m/2}}{S^{m/2-1}}\,\frac{\mu^{m/2}}{1-(1-\mu)^m}.\end{aligned} \tag{113}$$

Applying L'Hospital's rule [57], the limit of the RHS as $\mu \to 0$ is:

$$\left(\frac{2}{S}\right)^{m/2-1} \lim_{\mu\to 0} \frac{\mu^{m/2-1}}{(1-\mu)^{m-1}}, \tag{114}$$

which converges to 1 for $m = 2$, and to 0 otherwise, completing the proof.

Appendix B: Proof of Theorem 3

We first list some regularity properties of $\psi(t)$ that will be applied in the subsequent analysis — see, e.g., [55], [56]:

1) By assumption, $\psi(t) < \infty$ for all $t \in \mathbb{R}$. Since it is a LMGF, it is infinitely differentiable in $\mathbb{R}$. Also, since $\boldsymbol{x}$ is a non-degenerate (i.e., non deterministic) random variable, we have
$$\psi''(t) > 0, \qquad \forall t \in \mathbb{R}, \tag{115}$$
and, hence, $\psi(t)$ is strictly convex in $\mathbb{R}$.
2) With reference to the function $\frac{\psi(t)}{t}$ appearing in (36), we note that
$$\lim_{t\to 0} \frac{\psi(t)}{t} = \psi'(0), \tag{116}$$
and, hence, $\frac{\psi(t)}{t}$ is continuous for all $t \in \mathbb{R}$, and the integral in (36) is well-posed.
3) For all $t \neq 0$, we have
$$\frac{d}{dt}\frac{\psi(t)}{t} = \frac{\psi'(t)\,t - \psi(t)}{t^2}, \tag{117}$$
with
$$\lim_{t\to 0} \frac{\psi'(t)\,t - \psi(t)}{t^2} = \frac{\psi''(0)}{2}, \tag{118}$$
implying that $\frac{d}{dt}\frac{\psi(t)}{t}$ is continuous for all $t \in \mathbb{R}$. In addition, we have:
$$\frac{d}{dt}\frac{\psi(t)}{t} > 0, \quad \forall t \in \mathbb{R}. \tag{119}$$
This is immediately verified for $t = 0$ by using (115) in (118). For $t \neq 0$, since $\psi(t)$ is strictly convex and differentiable in $\mathbb{R}$, we can apply the first-order condition for strict convexity — see Eq. (3.3) in [58]:
$$\psi(a) - \psi(b) > \psi'(b)(a-b), \quad \forall a, b \in \mathbb{R}, \quad a \neq b. \tag{120}$$
Setting $a = 0$, $b = t \neq 0$, and using $\psi(0) = 0$, result (119) now follows from (117).

In the following, we denote by $\phi_\mu^{(c)}(t)$ the LMGF of the steady-state variable $\boldsymbol{y}_{k,\mu}^\star$ that would correspond to a fully connected network with uniform weights, $a_{k,\ell} = b_{k,\ell} = 1/S$ for all $k, \ell = 1, 2, \ldots, S$. We start by stating two lemmas (their proofs are given in the sequel).

Lemma 1 Define an auxiliary function $f_1(t)$ whose values over the negative and positive ranges of time are scaled as follows:

$$f_1(t) = \frac{t^2}{2} \times \begin{cases} \max\limits_{\tau\in[0,t]} \left(\dfrac{d}{d\tau}\dfrac{\psi(\tau)}{\tau}\right), & t \geq 0, \\ \max\limits_{\tau\in[t,0]} \left(\dfrac{d}{d\tau}\dfrac{\psi(\tau)}{\tau}\right), & t < 0. \end{cases} \tag{121}$$

Then, the LMGF of $\boldsymbol{y}_{k,\mu}^\star$ for the fully connected solution with uniform weights is:

$$\boxed{\phi_\mu^{(c)}(t) = \frac{S}{\mu}\left[\int_0^{\frac{\mu}{S}t} \frac{\psi(\tau)}{\tau} d\tau + \sum_{i=1}^{\infty} c_i(t;\mu)\right]} \tag{122}$$

where the functions $c_i(t;\mu)$ are nonnegative and satisfy

$$\sum_{i=1}^{\infty} c_i(t;\mu) \leq f_1\left(\frac{\mu}{S}t\right) \times \frac{\mu^2}{1-(1-\mu)^2}. \tag{123}$$

□

Lemma 2 Let λ_2 be the second largest eigenvalue of A in magnitude, and let $|\lambda_2| < \lambda < 1$. Define another auxiliary function as:

$$f_2(t) = |t| \times \begin{cases} \max\limits_{\tau\in[0,t]} |\psi'(\tau)|, & t \geq 0, \\ \max\limits_{\tau\in[t,0]} |\psi'(\tau)|, & t < 0. \end{cases} \tag{124}$$

Then, the LMGF of the steady-state diffusion output $\boldsymbol{y}_{k,\mu}^\star$ defined by (19) is:

$$\boxed{\phi_{k,\mu}(t) = \phi_\mu^{(c)}(t) + \sum_{i=1}^{\infty}\sum_{\ell=1}^{S} c_{i,\ell}(t;\mu)} \tag{125}$$

where the functions $c_{i,\ell}(t;\mu)$ now satisfy

$$\sum_{i=1}^{\infty}\sum_{\ell=1}^{S} |c_{i.\ell}(t;\mu)| \leq (\mathcal{C}\lambda S)\,\frac{f_2(\mu t)}{1-\lambda(1-\mu)}, \tag{126}$$

for a positive constant $\mathcal{C}$ depending on λ and on the combination matrix A. □

We can easily show that:

$$0 \leq f_1(t) < \infty, \qquad 0 \leq f_2(t) < \infty, \qquad \forall t \in \mathbb{R}. \tag{127}$$

Indeed, $f_1(t) \geq 0$ from (119), while $f_2(t) \geq 0$ by definition. Finiteness of both functions follows from Weierstrass extreme value theorem [57] since, by the properties of $\psi(t)$ discussed at the beginning of this appendix, the maxima appearing in (121) and (124) are maxima of continuous functions over compact sets for any finite t.

By using the above lemmas (whose proofs will be given soon), it is straightforward to prove Theorem 3.

Proof of Part i) of Theorem 3: we start by proving that

$$\lim_{\mu\to 0} \mu\,\phi_\mu^{(c)}(t/\mu) = S\int_0^{t/S} \frac{\psi(\tau)}{\tau} d\tau. \tag{128}$$

From the above Lemma 1 we have:

$$\left|\mu\,\phi_\mu^{(c)}(t/\mu) - S\int_0^{t/S} \frac{\psi(\tau)}{\tau} d\tau\right| = S\sum_{i=1}^{\infty} c_i(t/\mu;\mu) \leq S\,f_1(t/S) \times \frac{\mu^2}{1-(1-\mu)^2} \xrightarrow{\mu\to 0} 0\,. \tag{129}$$

On the other hand, using Lemma 2,

$$\mu\left|\phi_{k,\mu}(t/\mu)-\phi_{\mu}^{(c)}(t/\mu)\right| = \mu\left|\sum_{i=1}^{\infty}\sum_{\ell=1}^{S} c_{i,\ell}(t/\mu;\mu)\right| \leq (\mathcal{C}\lambda S)\, f_2(t)\,\frac{\mu}{1-\lambda(1-\mu)} \xrightarrow{\mu\to 0} 0, \tag{130}$$

and claim $i)$ is proven.

Proof of Part $ii)$ of Theorem 3: From the definition of $\omega(t)$ in (36) we have $\omega'(t)=\psi(t)/t$, which follows by continuity of $\psi(t)/t$ for all $t\in\mathbb{R}$ — see property 2) at the beginning of this appendix. Then, using the result proven in part $i)$, since $\omega(t)$ is differentiable in $\mathbb{R}$, the Gärtner-Ellis Theorem [56] stated in Sec. IV-B can be applied to conclude that $\boldsymbol{y}_{k,\mu}^{\star}$ must obey the LDP (29) with rate function given by the Fenchel-Legendre transform of the function $S\,\omega(t/S)$. It is straightforward to verify that the Fenchel-Legendre transform of a function scaled in this way is simply $S\,\Omega(\gamma)$.

We now prove the two lemmas.

Proof of Lemma 1. For the case of a fully connected network with uniform weights, the finite-horizon variable introduced in (90) reduces to

$$\boldsymbol{y}_k^{\star}(n) = \sum_{i=1}^{n}\sum_{\ell=1}^{S}\mu(1-\mu)^{i-1}\frac{1}{S}\,\boldsymbol{x}_{\ell}'(i). \tag{131}$$

Now since the LMGF is additive for sums of independent random variables, the LMGF of $\boldsymbol{y}_k^{\star}(n)$ defined above, for any fixed time instant n, is given by:

$$S\sum_{i=1}^{n}\psi\left((1-\mu)^{i-1}\frac{\mu}{S}\,t\right). \tag{132}$$

First we notice that, if we were able to show that this quantity converges as n goes to infinity, the limit will represent the LMGF, $\phi_{\mu}^{(c)}(t)$, of the steady-state random variable $\boldsymbol{y}_{k,\mu}^{\star}$ in the fully connected case, in view of the continuity theorem for the moment generating functions [59]. Define $g(t)=\psi(t)/t$ and let us focus initially on $t>0$. We introduce the countably infinite partition of the interval $[0,\frac{\mu}{S}\,t]$ with endpoints

$$\tau_i=(1-\mu)^{i-1}\frac{\mu}{S}\,t, \qquad i=1,2,\ldots,\infty. \tag{133}$$

A second-order Taylor expansion of the function $G(t)=\int_t^{\tau_i} g(\tau)d\tau$ around the point τ_i gives [57]:

$$\begin{aligned}\int_{\tau_{n+1}}^{\tau_1} g(\tau)d\tau &= \sum_{i=1}^{n}\int_{\tau_{i+1}}^{\tau_i} g(\tau)d\tau=\sum_{i=1}^{n}G(\tau_{i+1})\\ &= \sum_{i=1}^{n} g(\tau_i)\delta_i-\sum_{i=1}^{n} g'(\bar{t}_i)\frac{\delta_i^2}{2},\end{aligned} \tag{134}$$

for a certain $\bar{t}_i\in(\tau_{i+1},\tau_i)$, and with $\delta_i=\tau_i-\tau_{i+1}$. Using the explicit expressions for τ_i and $g(\cdot)$, we have

$$\begin{aligned}\sum_{i=1}^{n} g(\tau_i)\delta_i &= \sum_{i=1}^{n}\psi(\tau_i)\left(1-\frac{\tau_{i+1}}{\tau_i}\right)\\ &= \mu\sum_{i=1}^{n}\psi\left((1-\mu)^{i-1}\frac{\mu}{S}\,t\right),\end{aligned} \tag{135}$$

and we conclude that we can write

$$\mu\sum_{i=1}^{n}\psi\left((1-\mu)^{i-1}\frac{\mu}{S}\,t\right)=\int_{\tau_{n+1}}^{\tau_1} g(\tau)d\tau+\sum_{i=1}^{n} c_i(t;\mu), \tag{136}$$

where $c_i(t;\mu)$ is defined as:

$$c_i(t;\mu)=g'(\bar{t}_i)\frac{\delta_i^2}{2}>0. \tag{137}$$

Positiveness follows since $g'(t)>0$ for all $t\in\mathbb{R}$ in view of (119). Now note that

$$\sum_{i=1}^{n} c_i(t;\mu)\leq\sum_{i=1}^{\infty}\frac{\delta_i^2}{2}\max_{\tau\in[0,\mu t/S]} g'(\tau), \tag{138}$$

and recalling the definition of δ_i, we have

$$\begin{aligned}\sum_{i=1}^{\infty}\delta_i^2 &= \left(\frac{\mu}{S}\,t\right)^2\sum_{i=1}^{\infty}[(1-\mu)^{i-1}-(1-\mu)^{i}]^2\\ &= \left(\frac{\mu}{S}\,t\right)^2\frac{\mu^2}{1-(1-\mu)^2}.\end{aligned} \tag{139}$$

The proof for the case $t<0$ follows the same line of reasoning. We now obtain

$$\sum_{i=1}^{\infty} c_i(t;\mu)\leq f_1\left(\frac{\mu}{S}\,t\right)\times\frac{\mu^2}{1-(1-\mu)^2}, \tag{140}$$

where $f_1(\cdot)$ is defined in (121). As $n\to\infty$ in (136), the first term on the RHS converges to $\int_0^{\frac{\mu}{S}t} g(\tau)d\tau$ since the τ_i's define a countably infinite partition of $[0,\frac{\mu}{S}\,t]$. The second term is convergent from what was just proved. Using now (132), and letting $n\to\infty$, we finally get

$$\phi_{\mu}^{(c)}(t)=\frac{S}{\mu}\left[\int_0^{\frac{\mu}{S}t}\frac{\psi(\tau)}{\tau}d\tau+\sum_{i=1}^{\infty} c_i(t;\mu)\right]. \tag{141}$$

Proof of Lemma 2. Using a first-order Taylor expansion of the function $\psi(\cdot)$, the LMGF of the variable $\boldsymbol{y}_k^{\star}(n)$ defined earlier in (90) for diffusion networks using combination weights that are not necessarily uniform can be written as:

$$\begin{aligned}&\sum_{i=1}^{n}\sum_{\ell=1}^{S}\psi\left(\mu(1-\mu)^{i-1}b_{k,\ell}(i)t\right)\\ &= S\sum_{i=1}^{n}\psi\left((1-\mu)^{i-1}\,\frac{\mu}{S}\,t\right)\\ &+ \sum_{i=1}^{n}\sum_{\ell=1}^{S}\underbrace{\psi'(t_{i,\ell})\mu(1-\mu)^{i-1}\left[b_{k,\ell}(i)-\frac{1}{S}\right]}_{\triangleq c_{i,\ell}(t;\mu)}t,\end{aligned} \tag{142}$$

for an intermediate variable $t_{i,\ell}$ that, focusing first on the case $t>0$, must be certainly contained in the range $[0,\mu t]$, since $b_{k,\ell}\leq 1$. This yields:

$$\sum_{i=1}^{\infty}\sum_{\ell=1}^{S}|c_{i,\ell}(t;\mu)|\leq(\mathcal{C}\lambda S)\max_{\tau\in[0,\mu t]}|\psi'(\tau)|\,\frac{\mu\,t}{1-\lambda(1-\mu)}, \tag{143}$$

where we used Perron's Theorem (107). A similar argument holds for $t<0$.

APPENDIX C: CONVEXITY PROPERTIES OF $\omega(t)$ AND $\Omega(\gamma)$

The following properties hold.

i) $\omega''(t) > 0$ for all $t \in \mathbb{R}$, implying that $\omega(t)$ is strictly convex.

ii) $\Omega(\gamma)$ is strictly convex in the interior of the set:

$$\mathcal{D}_\Omega = \{\gamma \in \mathbb{R} : \ \Omega(\gamma) < \infty\}. \tag{144}$$

iii) $\Omega(\gamma)$ attains its unique minimum at $\gamma = \mathbb{E}\boldsymbol{x}$, with

$$\Omega(\mathbb{E}\boldsymbol{x}) = 0. \tag{145}$$

Proof.

i) In view of (36) we have $\omega'(t) = \psi(t)/t$. Positivity of $\omega''(t)$ follows now from (119).

ii) Consider first the following equation:

$$\gamma = \omega'(t). \tag{146}$$

Since $\omega'(t)$ is strictly increasing, it makes sense to define

$$\lim_{t\to+\infty} \omega'(t) = \omega_+, \qquad \lim_{t\to-\infty} \omega'(t) = \omega_-. \tag{147}$$

Clearly, if $\omega_+ = +\infty$ and $\omega_- = -\infty$, Eq. (146) will have a solution t for any $\gamma \in \mathbb{R}$. Consider the most restrictive situation that ω_- and ω_+ are both finite, and that $\gamma \notin [\omega_-, \omega_+]$. The case that only one of them is finite follows then in a straightforward manner.

Recall that the Fenchel-Legendre transform $\Omega(\gamma)$ of the function $\omega(t)$ is defined as:

$$\Omega(\gamma) = \sup_{t\in\mathbb{R}} [\gamma t - \omega(t)], \tag{148}$$

and let us introduce the function:

$$h(t) \triangleq \gamma t - \omega(t). \tag{149}$$

From the first-order condition for strict convexity (120) applied to the strictly convex function $\omega(t)$, we can write, for $t \neq 0$, $\omega'(t)t > \omega(t)$, which implies:

$$h(t) > [\gamma - \omega'(t)]\, t. \tag{150}$$

If $\gamma > \omega_+$, the term on the RHS diverges to $+\infty$ as $t \to +\infty$. Similarly, if $\gamma < \omega_-$, the term on the RHS diverges to $+\infty$ as $t \to -\infty$. This yields:

$$\sup_{t\in\mathbb{R}} h(t) = \infty, \tag{151}$$

showing, in view of (148) that the condition $\gamma \notin [\omega_-, \omega_+]$ implies $\gamma \notin \mathcal{D}_\Omega$.

The proof will be complete if we are able to show that $\Omega(\gamma) < \infty$ and $\Omega(\gamma)$ is strictly convex for $\gamma \in (\omega_-, \omega_+)$. Now, since $\omega(t)$ is differentiable and strictly convex in $\mathbb{R}$, we have that, for any γ, the function $h(t)$ in (149) is differentiable and strictly concave in $\mathbb{R}$, with

$$h'(t) = \gamma - \omega'(t). \tag{152}$$

Moreover, for $\gamma \in (\omega_-, \omega_+)$ the stationary-point equation

$$h'(t) = 0 \Leftrightarrow \gamma = \omega'(t) \tag{153}$$

admits a unique (since $\omega'(t)$ is strictly increasing) solution $t(\gamma)$. The strict concavity of $h(t)$ allows us to determine the supremum in (148) as follows:

$$\Omega(\gamma) = \gamma\, t(\gamma) - \omega(t(\gamma)) < \infty, \tag{154}$$

where finiteness of $\Omega(\gamma)$ follows by the fact that $t(\gamma) \in \mathbb{R}$, and by finiteness of $\omega(t)$. By further noting that $\omega'(t)$ is differentiable and $\omega''(t) > 0$, the theorem about differentiation of the inverse function [57, Ex. 2, p. 114] allows concluding that the derivative of the function $t(\gamma)$ can be computed as:

$$\frac{d}{d\gamma} t(\gamma) = \frac{1}{\omega''(t(\gamma))} > 0. \tag{155}$$

Then we can write

$$\frac{d}{d\gamma}\Omega(\gamma) = t(\gamma) + \gamma \frac{d}{d\gamma} t(\gamma) - \underbrace{\omega'(t(\gamma))}_{\gamma} \frac{d}{d\gamma} t(\gamma) = t(\gamma), \tag{156}$$

and

$$\frac{d^2}{d\gamma^2}\Omega(\gamma) = \frac{d}{d\gamma} t(\gamma) > 0, \tag{157}$$

which completes the proof.

iii) We have

$$\Omega(\gamma) = \sup_{t\in\mathbb{R}} [\gamma\, t - \omega(t)] \geq \gamma\, 0 - \omega(0) = 0. \tag{158}$$

Since $\omega'(0) = \mathbb{E}\boldsymbol{x}$, from (154) we conclude that

$$\Omega(\mathbb{E}\boldsymbol{x}) = (\mathbb{E}\boldsymbol{x})\, 0 - \omega(0) = 0, \tag{159}$$

and, hence, the minimum allowed value of zero is attained.

REFERENCES

[1] P. Braca, S. Marano, and V. Matta, "Enforcing consensus while monitoring the environment in wireless sensor networks," *IEEE Trans. Signal Process.*, vol. 56, no. 7, pp. 3375–3380, Jul. 2008.

[2] P. Braca, S. Marano, V. Matta, and P. Willett, "Asymptotic optimality of running consensus in testing statistical hypotheses," *IEEE Trans. Signal Process.*, vol. 58, no. 2, pp. 814–825, Feb. 2010.

[3] ——, "Consensus-based Page's test in sensor networks," *Signal Processing*, vol. 91, no. 4, pp. 919–930, Apr. 2011.

[4] D. Bajovic, D. Jakovetic, J. Xavier, B. Sinopoli, and J. M. F. Moura, "Distributed detection via Gaussian running consensus: Large deviations asymptotic analysis," *IEEE Trans. Signal Process.*, vol. 59, no. 9, pp. 4381–4396, Sep. 2011.

[5] D. Bajovic, D. Jakovetic, J. M. F. Moura, J. Xavier, and B. Sinopoli, "Large deviations performance of consensus+innovations distributed detection with non-Gaussian observations," *IEEE Trans. Signal Process.*, vol. 60, no. 11, pp. 5987–6002, Nov. 2012.

[6] D. Jakovetic, J. M. F. Moura, and J. Xavier, "Distributed detection over noisy networks: Large deviations analysis," *IEEE Trans. Signal Process.*, vol. 60, no. 8, pp. 4306–4320, Aug. 2012.

[7] J. Tsitsiklis, D. Bertsekas, and M. Athans, "Distributed asynchronous deterministic and stochastic gradient optimization algorithms," *IEEE Trans. Autom. Control*, vol. 31, no. 9, pp. 803–812, Sep. 1986.

[8] L. Xiao and S. Boyd, "Fast linear iterations for distributed averaging," *Systems and Control Letters*, vol. 53, no. 1, pp. 65–78, Sep. 2004.

[9] S. Boyd, A. Ghosh, B. Prabhakar, and D. Shah "Randomized gossip algorithms," *IEEE Trans. Inf. Theory*, vol. 52, no. 6, pp. 2508–2530, Jun. 2006.

[10] A. Nedic and A. Ozdaglar, "Distributed subgradient methods for multiagent optimization," *IEEE Trans. Autom. Control*, vol. 54, no. 1, pp. 48–61, Jan. 2009.

[11] A. G. Dimakis, S. Kar, J. M. F. Moura, M. G. Rabbat, and A. Scaglione, "Gossip algorithms for distributed signal processing," in *Proc. IEEE*, vol. 98, no. 11, pp. 1847–1864, Nov. 2010.

[12] S. Kar and J. M. F. Moura, "Convergence rate analysis of distributed gossip (linear parameter) estimation: Fundamental limits and tradeoffs," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 4, pp. 674–690, Aug. 2011.
[13] C. G. Lopes and A. H. Sayed, "Diffusion least-mean squares over adaptive networks: Formulation and performance analysis," *IEEE Trans. Signal Process.*, vol. 56, no. 7, pp. 3122–3136, Jul. 2008.
[14] F. S. Cattivelli and A. H. Sayed, "Diffusion LMS strategies for distributed estimation," *IEEE Trans. Signal Process.*, vol. 58, no. 3, pp. 1035–1048, Mar. 2010.
[15] A. H. Sayed, S.-Y. Tu, J. Chen, X. Zhao, and Z. J. Towfic, "Diffusion strategies for adaptation and learning over networks," *IEEE Signal Process. Mag.*, vol. 30, no. 3, pp. 155–171, May 2013.
[16] A. H. Sayed, "Diffusion adaptation over networks," in *Academic Press Library in Signal Processing*, vol. 3, R. Chellapa and S. Theodoridis, *Eds.*, pp. 323–454, Academic Press, Elsevier, 2014. Also available as arXiv:1205.4220 [cs.MA], May 2012.
[17] A. H. Sayed, "Adaptive networks," *Proceedings of the IEEE*, vol. 102, no. 4, pp. 460–497, Apr. 2014.
[18] J. Chen and A. H. Sayed, "Distributed Pareto optimization via diffusion strategies," *IEEE J. Sel. Topics Signal Process.*, vol. 7, no. 2, pp. 205–220, Apr. 2013.
[19] S.-Y. Tu and A. H. Sayed, "Diffusion strategies outperform consensus strategies for distributed estimation over adaptive networks," *IEEE Trans. Signal Process.*, vol. 60, no. 12, pp. 6217–6234, Dec. 2012.
[20] X. Zhao and A. H. Sayed, "Performance limits for distributed estimation over LMS adaptive networks," *IEEE Trans. Signal Process.*, vol. 60, no. 10, pp. 5107–5124, Oct. 2012.
[21] M. Longo, T. D. Lookabaugh, and R. M. Gray, "Quantization for decentralized hypothesis testing under communication constraints," *IEEE Trans. Inf. Theory*, vol. 36, no. 2, pp. 241–255, Mar. 1990.
[22] P. K. Varshney, *Distributed Detection and Data Fusion*. Springer-Verlag, New York, 1997.
[23] R. Viswanathan and P. K. Varshney, "Distributed detection with multiple sensors: Part I–Fundamentals," in *Proc. IEEE*, vol. 85, no. 1, pp. 54–63, Jan. 1997.
[24] R. S. Blum, S. A. Kassam, and H. V. Poor, "Distributed detection with multiple sensors: Part II–Advanced topics," in *Proc. IEEE*, vol. 85, no. 1, pp. 64–79, Jan. 1997.
[25] J. F. Chamberland and V. V. Veeravalli, "Decentralized detection in sensor networks," *IEEE Trans. Signal Process.*, vol. 51, no. 2, pp. 407–416, Feb. 2003.
[26] J. F. Chamberland and V. V. Veeravalli, "Wireless sensors in distributed detection applications," *IEEE Signal Process. Mag.*, vol. 24, no. 3, pp. 16–25, May 2007.
[27] B. Chen, L. Tong, and P. K. Varshney, "Channel-aware distributed detection in wireless sensor networks," *IEEE Signal Process. Mag.*, vol. 23, no. 4, pp. 16–26, Jul. 2006.
[28] V. Saligrama, M. Alanyali, and O. Savas, "Distributed detection in sensor networks with packet losses and finite capacity links," *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4118–4132, Nov. 2006.
[29] F. S. Cattivelli and A. H. Sayed, "Distributed detection over adaptive networks using diffusion adaptation," *IEEE Trans. Signal Process.*, vol. 59, no. 5, pp. 1917–1932, May 2011.
[30] P. Braca, S. Marano, V. Matta, and A. H. Sayed, "Large deviations analysis of adaptive distributed detection," in *Proc. IEEE ICASSP*, Florence, Italy, May 2014, pp. 6112–6116.
[31] V. Matta, P. Braca, S. Marano, and A. H. Sayed, "Exact asymptotics of distributed detection over adaptive networks," in *Proc. IEEE ICASSP*, Brisbane, Australia, April 2015, pp. 3377–3381.
[32] R. R. Bitmead, "Convergence in distribution of LMS-type adaptive parameter estimates," *IEEE Trans. Autom. Control*, vol. 28, no. 1, pp. 54–60, Jan. 1983.
[33] X. Zhao and A. H. Sayed, "Probability distribution of steady-state errors and adaptation over networks," in *Proc. IEEE International Workshop on Statistical Signal Processing (SSP)*, Nice, France, Jun. 2011, pp. 253–256.
[34] J. Chen and A. H. Sayed, "On the probability distribution of distributed optimization strategies," in *Proc. IEEE GlobalSIP*, Austin, TX, USA, Dec. 2013, pp. 1–5.
[35] K. R. Rad and A. Tahbaz-Salehi, "Distributed parameter estimation in networks," in *Proc. IEEE Conference on Decision and Control (CDC)*, Atlanta, GA, USA, Dec. 2010, pp. 5050–5055.
[36] S. Shahrampour and A. Jadbabaie, "Exponentially fast parameter estimation in networks using distributed dual averaging," in *Proc. IEEE Conference on Decision and Control (CDC)*, Florence, Italy, Dec. 2013, pp. 6196–6201.
[37] S. Kar, R. Tandon, H. V. Poor, and Shuguang Cui, "Distributed detection in noisy sensor networks," in *Proc. IEEE International Symposium on Information Theory (ISIT)*, St. Petersburg, Russia, Jul. 31–Aug. 5 2011, pp. 2856–2860.
[38] J. Chen and A. H. Sayed, "On the learning behavior of adaptive networks — Part I: Transient analysis," *IEEE Trans. Inf. Theory*, vol. 61, no. 6, pp. 3487-3517, Jun. 2015.
[39] J. Chen and A. H. Sayed, "On the learning behavior of adaptive networks — Part II: Performance analysis," *IEEE Trans. Inf. Theory*, vol. 61, no. 6, pp. 3518-3548, Jun. 2015.
[40] T. Cover and J. Thomas, *Elements of Information Theory*. John Wiley & Sons, NY, 1991.
[41] H. Shao, *Mathematical Statistics*, 2nd ed., Springer, 2003.
[42] E. L. Lehmann and J. P. Romano, *Testing Statistical Hypotheses*. Springer, 2005.
[43] T. Berger, Z. Zhang, and H. Viswanathan, "The CEO problem," *IEEE Trans. Inf. Theory*, vol. 42, no. 3, pp. 887–902, May 1996.
[44] H. Viswanathan and T. Berger, "The quadratic Gaussian CEO problem," *IEEE Trans. Inf. Theory*, vol. 43, no. 5, pp. 1549–1559, Sep. 1997.
[45] S. A. Kassam, *Signal Detection in Non-Gaussian Noise*. Springer-Verlag, 1987.
[46] H. V. Poor, *An Introduction to Signal Detection and Estimation*. Springer-Verlag, 1988.
[47] A. H. Sayed, *Adaptive Filters*. Wiley, NJ, 2008.
[48] R. Horn and C. Johnson, *Matrix Analysis*. Cambridge University Press, Cambridge, UK, 1985.
[49] W. Feller, *An Introduction to Probability and Its Applications,* vol. *2*, Wiley, NY, 1971.
[50] P. Billingsley, *Convergence of Probability Measures*, 2nd ed., Wiley, 1999.
[51] B. Jessen and A. Wintner, "Distribution functions and the Riemann Zeta function," *Trans. Amer. Math. Soc.*, vol. 38, pp. 48–88, 1935.
[52] L. Breiman, *Probability*. SIAM, PA, 1968.
[53] P. Erdös, "On a family of symmetric Bernoulli convolutions," *American Journal of Mathematics*, vol. 61, no. 4, pp. 974–976, 1939.
[54] X. Zhao and A. H. Sayed, "Attaining optimal batch performance via distributed processing over networks," in *Proc. IEEE ICASSP*, Vancouver, Canada, May 2013, pp. 5214–5218.
[55] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*. Springer, 1998.
[56] F. den Hollander, *Large Deviations*. American Mathematical Society, 2008.
[57] W. Rudin, *Principles of Mathematical Analysis*. McGraw-Hill, 1976.
[58] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, Cambridge, UK, 2004.
[59] J. H. Curtiss, "A note on the theory of moment generating functions," *The Annals of Statistics*, vol. 13, no. 4, pp. 430–433, 1942.