Improved Performance Properties of the CISPRT Algorithm for Distributed Sequential Detection

Kun Liu and Yajun Mei

Abstract

In distributed sequential detection problems, local sensors observe raw local observations over time, and are allowed to communicate local information with their immediate neighborhood at each time step so that the sensors can work together to make a quick but accurate decision when testing binary hypotheses on the true raw sensor distributions. One interesting algorithm is the Consensus-Innovation Sequential Probability Ratio Test (CISPRT) algorithm proposed by Sahu and Kar (IEEE Trans. Signal Process., 2016). In this article, we present improved finite-sample properties on error probabilities and expected sample sizes of the CISPRT algorithm for Gaussian data in term of network connectivity, and derive its sharp first-order asymptotic properties in the asymptotic regime when Type I and II error probabilities go to 0. The usefulness of our theoretical results are validated through numerical simulations.

Keywords: CISPRT, distributed learning, network connectivity, sequential detection, oracle properties.

I. Introduction

Distributed online learning becomes increasingly important in many real-world applications such as cognitive radio networks [1], [2], social recommender systems [3], [4], natural language processing [5]. Under a general setting, there are N sensors or agents taking raw observations over time in a system, and each local sensor can only communicate its local information with the immediate neighbors at each time. Such local information communication can be conducted adaptively or sequentially over time so that sensors can work together to reach consensus quickly. The advantages of distributed settings are to protect intrinsic privacy of sensitive data [4], increase computational capacity [6]–[8], and mitigate collection and storage burden of modern large datasets [9], [10].

There are many important distributed online learning problems in engineering and statistics, and one of them is the distributed sequential detection, see [11]–[13], where the distributed sensors work together to quickly and correctly decide which is the true underlying probability measure or model for raw sensor observations. Had the local sensors been able to send local information to a central location, often called the fusion center, for further analysis, extensive research has been done along two distinct directions. The first one is when the fusion center has access to all raw sensor observations, which is the centralized sequential detection problem. This is well studied in the classical subfield of sequential analysis in statistics [14]–[17]. In particular, the optimal centralized procedure is the well-known Sequential Probability Ratio Test (SPRT), see [14]. The other direction is the decentralized sequential detection, where the local sensors send quantized sensor messages to the fusion center to make a global decision, see [18]–[21].

Manuscript received March 28, 2019.

K. Liu is with Wells Fargo & Company, San Francisco, CA, USA (email: liukun0924@gmail.com), and Y. Mei is with The H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA, USA (ymei@isye.gatech.edu). This is part of Kun Liu's PhD dissertation at the Georgia Institute Technology, and the research is partially supported by the National Science Foundation [CMMI-1362876 and ATD-1830344].

Research is rather limited in the distributed sequential detection where there is no fusion center and the local sensors need to work together to make a decision: very few efficient algorithms have been proposed partly because it involves complicated communication strategies between local sensors and their neighborhood sensors. One exception is the Consensus-Innovation SPRT (CISPRT) algorithm developed in [22] that is based on the weighted average of local log-likelihood ratio tests, see [23]–[27] for the motivation and more background. The CISPRT algorithm is novel and interesting, as each local sensor utilizes local information not only from itself and its immediate neighbor sensors, but also from remote connected sensors that are 2-hop or more hops away from itself. Also see [28] for an interesting generalization of the CISPRT algorithm under the fixed q-round message passing protocol and see [29] for the extension in composite hypothesis testing problems.

Intuitively, the performance of distributed algorithms including the CISPRT will depend on the neighborhood structure of local senors, or the network connectivity. For any pre-specified neighborhood structure of local senors, Sahu and Kar [22] characterized the various performance properties of the CISPRT for Gaussian data. In particular, for the CISPRT satisfying the error probability constraint ϵ, explicit lower and upper bounds were derived on its performance properties $h(\epsilon)$ such as the expected sample sizes and the information loss with respect to the optimal centralized SPRT: $L(\epsilon) < h(\epsilon) < U(\epsilon)$, which holds non-asymptotically for *any* network structure. These are remarkable non-asymptotic results on sequential detection in the high-dimensional setting, as the explicit bounds $L(\epsilon)$ and $U(\epsilon)$ clearly characterize the effects of the network structure and the dimension (or the number of sensors). Unfortunately, these bounds are too loose in the special centralized setting when each local sensor is connected to all other sensors and the CISPRT becomes the well-known centralized SPRT: the ratio $U(\epsilon)/L(\epsilon)$ converges to $5/4$ and $10/7$ for the expected sample size and information loss, respectively, under the asymptotic regime when the error probability constraint $\epsilon \to 0$. This led us to raise an open problem whether one can derive non-asymptotic lower and/or upper bounds that are also asymptotically sharp.

The objective of this article is to provide a positive answer to this open problem, and our main focus is to present improved finite-sample upper bounds on error probabilities and expected sample sizes of the CISPRT algorithm, and derive its sharp first-order asymptotic properties when the error probability constraint ϵ goes to 0. Note that it is mathematically challenging to provide an accurate analysis on the performance properties of procedures in sequential detection or sequential hypothesis testing. The standard techniques are renewal theory and overshoot analysis [14]–[17], but they are designed for the fixed dimensional setting and are inappropriate in our context for any network structure regardless of the dimension N (or the number N of sensors), since the corresponding results implicitly involve overshoot constants that cannot be computed explicitly for a given dimension N and will be exponentially increasing as a function of the dimension N.

Our main scientific contributions are two-fold. From the technique viewpoint, we develop a tail probability analysis technique that is able to derive sharp information bounds that are not only comparable to the classical techniques in the one or low dimensional setting (i.e., when there are one or very few sensors), but also much better in the high-dimensional setting (i.e., when there are a large number of sensors). Our proposed technique is to extend the finite sum of tail probabilities in [22] to the infinite sum, and provides a new and useful tool that is able to provide accurate performance analysis in the sequential detection context. From the network application viewpoint, we derive refined, non-asymptotic upper bounds on the performance properties of the CISPRT algorithm for Gaussian data under any pre-specified neighborhood structure of local senors regardless of the number N of sensors, as compared to those in [22]. Our results indicate that the more the number of sensors or the sparser the network neighborhood connectivity is, the larger the information loss is, i.e., the larger expected sample size is needed to achieve the desired Type I and II error probabilities.

The remainder of this article is organized as follows. In Section II, we present the formulation of the distributed sequential

detection problem, the CISPRT algorithm proposed in [22], and the background materials for spectral graph theory. In Section III, we present our main theoretical results on the refined performance properties of the CISPRT. Simulation studies results are presented in Section IV, and the detailed proof of our main theorem, Theorem 1, is provided in Section V. Some conclusion remarks are included in Section VI.

II. Preliminaries and Background

A. Distributed Sequential Detection Problems

Consider a network system of N sensors that takes observations over time. At each time step $t = 1, 2, \cdots$, the i-th sensor observes an observation $y_i(t)$ for $i = 1, \ldots, N$. There are two hypotheses on the distributions of the local sensor observations $y_i(t)$'s. Under the null hypothesis H_0, the $y_i(t)$'s are $N(-\mu, \sigma^2)$ and under the alternative hypothesis H_1, the $y_i(t)$'s are $N(\mu, \sigma^2)$. Here the sensor observations $y_i(t)$'s are assumed to be independent and identically distributed (iid) over time and across sensors, conditional on each hypothesis. Note that Guassian distributions are one of the most widely used models in signal processing and many other applications, and by the linear additive properties of Gaussian models, our problem is equivalent to a more familiar problem in the additive white Gaussian noise channel of utilizing the observations $y_i^*(t) = y_i(t) + \mu$ to test hypotheses $H_0 : N(0, \sigma^2)$ (i.e., white noises) against $H_1 : N(2\mu, \sigma^2)$ (i.e., a signal exists). Here we follow [22] to adopt the current notation so as to simplify the technical presentations and proofs.

Under the distributed sequential detection setting, the objective is for each local sensor to work with its neighborhood sensors to make a quick and accurate decision on which of these two hypotheses is true. In particular, each local sensor can only communicate its local information with its (one-hop) neighborhood sensors. Here we assume that the neighborhood structure of sensors is pre-specified, and can be represented as an undirected graph $G = (V, E)$: the i-th vertex in V represents the i-th sensor, and there is an edge between the i-th vertex/sensor and the j-th vertex/sensor, i.e., $(i, j) \in E$, if and only if the corresponding sensors are neighbors and can communicate local information with each other. Here we assume that the graph $G = (V, E)$ is simple, i.e., without self loops and multiple edges. For the i-th sensor, its neighborhood is given by $\Omega_i = \{j \in V | (i, j) \in E\}$, and its degree is given by the cardinality $d_i = |\Omega_i|$. See [30] for more graph theoretic methods in network systems.

For a distributed sequential procedure D, it consists of $(T_i, \delta_i)_{i=1}^N$, where T_i is the number of time steps the i-th local sensor needs to make a local decision $\delta_i \in \{0, 1\}$. Here T_i is a local stopping time in the sense that $\{T_i = t\}$ depends on the information from the i-th local sensor as well as its neighborhood up to time t. The local decision $\delta_i = 0$ or 1 means that the i-th local sensor accepts H_0 or H_1, respectively.

The performance of a distributed sequential procedure $D = (T_i, \delta_i)_{i=1}^N$ is evaluated by its local expected sample sizes, $\boldsymbol{E}_1[T_i]$ and $\boldsymbol{E}_0[T_i]$, and its local error probabilities, $\boldsymbol{P}_0[\delta_i = 1]$ and $\boldsymbol{P}_1[\delta_i = 0]$. Ideally one would like all these four local quantities to be simultaneously as small as possible for all local sensors, which is impossible. As mentioned in [22], one useful formulation is to find a distributed sequential procedure $D = (T_i, \delta_i)_{i=1}^N$ that (asymptotically) minimizes

$$\max_{i=1,2,\cdots,N} \boldsymbol{E}_1[T_i] \tag{1}$$

subject to the local false alarm and missed detection constraints:

$$\boldsymbol{P}_0[\delta_i = 1] \leq \alpha \text{ and } \boldsymbol{P}_1[\delta_i = 0] \leq \beta \tag{2}$$

for all $i = 1, 2, \cdots, N$, where $0 < \alpha, \beta \leq 1/2$ are the pre-specified false alarm bounds.

Note that the objective function in (1) can also be replaced by other functions if one wants. For instance, the local criterion $\boldsymbol{E}_1[T_i]$ can be replaced by $\boldsymbol{E}_0[T_i]$, or more generally, the Bayesian-type criterion $\pi_0\boldsymbol{E}_0[T_i] + (1-\pi_0)\boldsymbol{E}_1[T_i]$. Fortunately, in the context of sequential tests, Wald's (optimal centralized) SPRT can minimize each and every of these criteria, and thus we consider the criterion of $\boldsymbol{E}_1[T_i]$ here. Moreover, we can also consider other kind of criteria at the global level such as $\min_{i=1,2,\cdots,N} \boldsymbol{E}_1[T_i]$. Here we do not discuss the appropriateness of different formulations or the corresponding optimality theories, and our focus is to investigate the performance of a specific distributed sequential procedure. For that reason, our results below deal with the local expected sample sizes $\boldsymbol{E}_1[T_i]$'s themselves, since it is straightforward to extend these local results to the global level such as that in (1).

B. SPRT and CISPRT

Let us first consider the centralized setup when the graph of the network neighborhood structure is complete in the sense that at each time step each local sensor has access to all raw observations over the graph. This is equivalent to the scenario with the fusion center, as each and every local sensor can be regarded as the fusion center. In such scenario, Wald's SPRT is the optimal centralized sequential test under the formulation of (1) and (2). To define the SPRT, denote the local log-likelihood ratio of the i-th sensor at time step t by

$$\eta_i(t) = \log \frac{f_1(y_i(t))}{f_0(y_i(t))} = \frac{2\mu}{\sigma^2} y_i(t), \tag{3}$$

and denote the centralized likelihood ratio statistic up to time t by

$$S_c(t) = \sum_{s=1}^{t}\sum_{i=1}^{N} \eta_i(s) = S_c(t-1) + \sum_{i=1}^{N} \eta_i(t) \tag{4}$$

for all $t \geq 1$. The centralized SPRT is then defined by the stopping time

$$T_c = \inf\{t : S_c(t) \notin [\gamma_c^l, \gamma_c^h]\}, \tag{5}$$

for some pre-specified constants $\gamma_c^l < 0 < \gamma_c^h$. When stopping, the centralized SPRT makes a decision $\delta_c = 0$ or 1 depending on whether lower bound γ_c^l or upper bound γ_c^h is crossed at time T_c. It is well-known (Ch 2 of [15]) that

$$\begin{aligned} \boldsymbol{P}_0[\delta_c = 1] &\leq \exp(-\gamma_c^h)(1 - \boldsymbol{P}_1[\delta_c = 0]), \\ \boldsymbol{P}_1[\delta_c = 0] &\leq \exp(\gamma_c^l)(1 - \boldsymbol{P}_0[\delta_c = 1]). \end{aligned} \tag{6}$$

Hence, in order to satisfy the false alarm constraints in (2), by (6), a commonly used choice of the threshold is

$$\gamma_c^l = \log \frac{\beta}{1-\alpha} \quad \text{and} \quad \gamma_c^h = \log \frac{1-\beta}{\alpha}. \tag{7}$$

Moreover, since $y_i(t) \sim N(-\mu, \sigma^2)$ under H_0 or $N(\mu, \sigma^2)$ under H_1, the Kullback-Leibler divergence at each local sensor is

$$m = \boldsymbol{E}_1(\eta_i(t)) = \frac{2\mu^2}{\sigma^2} \tag{8}$$

and thus the centralized Kullback-Leibler divergence of the joint observation $\mathbf{Y}(t) = (Y_1(t), \cdots, Y_N(t))$ is Nm. Furthermore, as shown in [14], subject to the false alarm constraint in (2), for any sequential test T, distributed or centralized, $\boldsymbol{E}_1(T) \geq \mathcal{M}(\alpha, \beta)$, where the universal lower bound is given by

$$\mathcal{M}(\alpha, \beta) = \frac{1}{Nm}\Big[(1-\beta)\log\frac{1-\beta}{\alpha} + \beta \log \frac{\beta}{1-\alpha}\Big]. \tag{9}$$

Also the centralized SPRT T_c in (5) attains this lower bound asymptotically for fixed N and m as $\alpha, \beta \to 0$.

Now let us switch to the distributed setup for a general neighborhood structure where each local sensor can only communicate with its neighborhood sensors. In [22], the authors proposed an interesting CISPRT algorithm where each local sensor makes a local decision based on the weighted average of the local likelihood ratio statistics from itself and its neighborhood sensors. Specifically, at time step t, each i-th local sensor computes its local test statistic recursively:

$$\begin{aligned} S_i(t) &= w_{ii}S_i(t-1) + \sum_{j\in\Omega_i} w_{ij}S_j(t-1) \\ &\quad + w_{ii}\eta_i(t) + \sum_{j\in\Omega_i} w_{ij}\eta_j(t), \end{aligned} \tag{10}$$

for $t = 1, 2, \cdots$, where the initial value $S_i(0) = 0$ and Ω_i is the (one-hop) neighborhood of the i-th sensor. Here the w_{ij}'s are pre-specified weights satisfying

$$w_{ii} + \sum_{j\in\Omega_i} w_{ij} = 1, \text{ and } w_{ij} \geq 0, \ \forall i, j \tag{11}$$

and the discussion on the choices of the weights w_{ij}'s will be postponed a little bit.

Under the matrix notation, let us collect the weights w_{ij}'s in an $N \times N$ matrix $\mathbf{W}$, where $w_{ij} = 0$ if $(i, j) \notin E$. Denote by $\mathbf{S}(t)$ and $\boldsymbol{\eta}(t)$ as the $N \times 1$ vectors $(S_1(t), \cdots, S_N(t))^T$ and $(\eta_1(t), \cdots, \eta_N(t))^T$. The local test statistics can be updated recursively as

$$\mathbf{S}(t) = \mathbf{W}\Big(\mathbf{S}(t-1) + \boldsymbol{\eta}(t)\Big). \tag{12}$$

for $t \geq 1$.

For the CISPRT, each i-th sensor makes a local decision at time

$$T_i = \inf\{t \geq 1 : S_i(t) \notin [\gamma_i^l, \gamma_i^h]\}, \tag{13}$$

for some pre-specific thresholds $\gamma_i^l < 0 < \gamma_i^h$. When stopping, the i-th sensor makes a local decision

$$\delta_i = \begin{cases} 0, & \text{if } S_i(T_i) \leq \gamma_i^l; \\ 1, & \text{if } S_i(T_i) \geq \gamma_i^h. \end{cases} \tag{14}$$

From the pure mathematical viewpoint, the stopping times T_i's in (13) depend on the properties of $S_i(t)$'s, which is a component of the N-dimensional random walks $\mathbf{S}(t)$. One may be able to apply the classical renewal theory to analyze the "asymptotic" properties of the stopping times T_i's in (13), but unfortunately such an approach will involve "constant" terms for overshoot analysis that are exponentially increasing as the dimension N increases. In particular, when the number N of sensors is large, such constant terms can be huge, and thus the corresponding asymptotic analysis can be meaningless under any reasonable practical setting of distributed detection. Here we provide an alternative approach that yields the same first-order asymptotic result as in the classical renewal theory when the number N of sensors is 1 or very small, but has a potential to derive useful oracle properties of stopping times under the setting of high dimension N for Gaussian data.

C. Spectral Graph Theory and Weight Matrix $\mathbf{W}$

In this subsection, let us present some basic materials for spectral graph theory that are related to the main assumption for our network structure and the design of the weight matrix $\mathbf{W}$ in (11). Also see [30] for a more complete introduction of graph theory.

Recall our network neighbor structure is characterized by the undirected simple graph $G = (V, E)$. In spectral graph theory, the degree matrix $\mathbf{D}$ is an $N \times N$ diagonal matrix with the i-th diagonal being d_i, the degree of the i-th vertex. The adjacency

matrix $\mathbf{A}$ is a $(0,1)$-matrix with zeros on its diagonal and $A_{ij} = 1$ if and only if the i-th vertex and the j-th vertex are connected for all $1 \leq i \neq j \leq N$. The Laplacian matrix is then defined as

$$\mathbf{L} = \mathbf{D} - \mathbf{A}. \tag{15}$$

Alternatively, for the Laplacian matrix $\mathbf{L}$, its elements are given by

$$L_{i,j} = \begin{cases} d_i & \text{if } i = j; \\ -1 & \text{if } i\text{-th and } j\text{-th vertex connected for } i \neq j; \\ 0 & \text{otherwise.} \end{cases}$$

The Laplacian matrix $\mathbf{L}$ is a positive semidefinite matrix, and thus has N non-negative eigenvalues:

$$0 = \lambda_1(\mathbf{L}) \leq \lambda_2(\mathbf{L}) \leq \cdots \leq \lambda_N(\mathbf{L}). \tag{16}$$

Moreover, the number of times 0 appears as an eigenvalue of the Laplacian matrix $\mathbf{L}$ is the number of connected components in the graph. Equivalently, a graph is connected if and only if $\lambda_2(\mathbf{L}) > 0$, see [31].

Our main assumption on the network neighborhood structure is as follows.

Assumption 1. *The graph $G = (V, E)$ is connected, or equivalently, the second smallest eigenvalue of the Laplacian matrix $\mathbf{L}$ is positive, i.e., $\lambda_2(\mathbf{L}) > 0$.*

Next, let us discuss the choices of the weight matrix $\mathbf{W}$ in (12). From the technical or algorithm viewpoint, the weight matrix $\mathbf{W}$ can be arbitrary as long as $\mathbf{W}$ is a stochastic matrix in the sense of satisfying (11). However, in the context of distributed sequential detection, the implicit assumption is that $w_{ij} > 0$ if and only if the i-th and j-th sensors are neighbors. There are still many reasonable choices for the weight matrix $\mathbf{W}$, and one useful one is to define

$$\mathbf{W} = \mathbf{I}_{N\times N} - \delta\mathbf{L}, \tag{17}$$

where $\mathbf{I}_{N\times N}$ is the $N \times N$ identity matrix and δ is a small positive constant so that all elements of $\mathbf{W}$ are positive and thus (11) holds. Under this choice of the weight matrix $\mathbf{W}$, for a given i-th sensor/vertex, it assigns a small but equal weight of $w_{ij} = \delta$ to all of its d_i neighbor sensors, and assigns most weight $w_{ii} = 1 - \delta d_i$ to itself.

An interesting fact of the weight matrix $\mathbf{W}$ in (17) is that it is symmetric ($w_{ij} = w_{ji}$) and irreducible, and the latter is due to the fact that the graph is connected under Assumption 1. In addition, it is straightforward to show that as a stochastic matrix satisfying (11), the matrix $\mathbf{W}$ has the largest eigenvalue 1, and the second largest eigenvalue, denote by r, is strictly less than 1. Recall that any $N \times N$ symmetric matrix can be written as $\mathbf{W} = \sum_{i=1}^{N} \lambda_i u_i v_i^T$, where u_i and v_i are singular vectors associated with the i-th largest eigenvalue λ_i. For the stochastic symmetric matrix $\mathbf{W}$ satisfying (11), we have $\lambda_1 = 1$ and $u_1 = v_1 = \frac{1}{\sqrt{N}}\mathbf{1}$, where $\mathbf{1}$ is an N-dimensional all one vector. Thus $\mathbf{W} - \mathbf{J} = \sum_{i=2}^{N} \lambda_i u_i v_i^T$, where $\mathbf{J} = \frac{1}{N}\mathbf{1}\mathbf{1}^T$ is an $N \times N$ matrix with all entries being the constant $1/N$. This eigenvalue decomposition representation shows that the largest eigenvalue of $\mathbf{W} - \mathbf{J}$ is simply the second largest eigenvalue of $\mathbf{W}$. Hence, under our notation, the second largest eigenvalue r of $\mathbf{W}$ can be characterized by the spectral norm (or the largest eigenvalue) of the matrix $\mathbf{W} - \mathbf{J}$, i.e.,

$$r = ||\mathbf{W} - \mathbf{J}||. \tag{18}$$

III. Improved Properties of CISPRT

In this section, we derive our main theoretical properties of the CISPRT procedure in (13) and (14) under Assumption 1 when the network is connected. Note that there are two thresholds, γ_i^l and γ_i^h, in the CISPRT procedure in (13). At the high

level, the upper threshold γ_i^h is closely related to Type I error probability and the expected sample size under H_1, whereas the lower threshold γ_i^l is closely related to Type II error probability and the expected sample size under H_0. For simplicity of the technical proofs, below our main theorem will focus on the effects of the upper bound γ_i^h on the Type I error probability and the expected sample size under H_1 of the CISPRT. The usefulness of the our main theorem is illustrated in several corollaries which consider the symmetric scenario when the lower and upper thresholds satisfy $\gamma_i^l = -\gamma_i^h$.

Let us first summarize the main theoretical results of [22] that are closely related to our paper. For a given weight matrix $\mathbf{W}$, denote by

$$\rho = 1 - \exp\left(-\frac{Nm}{4(Nr^2+1)}\right), \tag{19}$$

where m is the Kullback-Leibler divergence in (8) and r is the second largest eigenvalue of $\mathbf{W}$ and can be rewritten as in (18). It was shown in [22] that the Type I error of the CISPRT algorithm satisfies

$$\mathbb{P}_0(\delta_i = 1) \leq 2\rho^{-1} \exp\left(-\frac{7}{8}D\right), \tag{20}$$

with the term D in the exponent being

$$D = \frac{N}{Nr^2+1}\gamma_i^h, \tag{21}$$

and its expected sample size satisfies

$$\begin{aligned} & \frac{1}{m}[\gamma_i^h - \mathbb{P}_1(\delta_i = 0)(\gamma_i^h - \gamma_i^l) - c] \\ \leq \quad & \boldsymbol{E}_1(T_i) \leq \frac{5}{4}\frac{1}{m}\gamma_i^h + \rho^{-1} + 1, \end{aligned} \tag{22}$$

see Theorem 4.1, Theorem 4.7, and equation (49) of [22]. Here the constant $c > 0$ in the lower bound in (22) is independent of the thresholds γ_i^h, γ_i^l and is a complicated function of the network topology and the Gaussian model statistics, see equations (47)-(49) of [22]. Note that the original upper bound does not have the constant 1 in the right-hand side of (22), but we found out that the original proof in [22] contains a minor mistake to count the number of integers in the interval $0 \leq t \ \leq a$ as a, not $a+1$. Thus we add 1 here so that the results are mathematically rigorous.

Moreover, the authors in [22] considered the asymptotic properties of the symmetric CISPRT with $\gamma_i^l = -\gamma_i^h$ subject to the local false alarm constraints with $\alpha = \beta = \epsilon$ as $\epsilon \to 0$. By (20) and (21), a conservative choice of the thresholds $\gamma_i^h = -\gamma_i^l$ is

$$\gamma_i^{h,0} = -\gamma_i^{l,0} = \frac{8(Nr^2+1)}{7N}\left(\log(2\rho^{-1}) + \log\epsilon^{-1}\right), \tag{23}$$

see equation (19) in Theorem 4.1 of [22]. For the CISPRT with the thresholds in (23), the authors in [22] then compared its expected stopping time with the universal lower bound in (9):

$$1 \leq \limsup_{\epsilon\to 0} \frac{\boldsymbol{E}_1(T_i)}{\mathcal{M}(\epsilon)} \leq \frac{10}{7}(Nr^2+1), \tag{24}$$

where the lower bound is trivial since $\mathcal{M}(\epsilon) = \mathcal{M}(\alpha = \epsilon, \beta = \epsilon)$ is the universal lower bound in (9) for any stopping times satisfying the local false alarm constraints when $\alpha = \beta = \epsilon$, and the factor $\frac{10}{7}$ in the upper bound are based on the factors $\frac{8}{7} \times \frac{5}{4}$ from (23) and (22).

In this paper, we improve the constants $7/8$ and $5/4$ of the non-asymptotic upper bounds in (20)-(22) in the original CISPRT paper [22] to 1 and 1, respectively, which in turn allows us to improve the asymptotic inequality in (24) to find the exact asymptotic limit. This is because our non-asymptotic upper bounds turn out to be first-order tight under the asymptotic regime, and the price we pay is to add extra terms that can be thought of as the second-order terms under the asymptotic regime.

Our main new results under the non-asymptotic regime can be summarized in the following theorem, whose proof is very technical and will be postponed to Section V.

Theorem 1. *For the CISPRT, at any given i-th local sensor, the Type I error probability satisfies*

$$\begin{aligned}&\mathbb{P}_0(\delta_i=1)\le \rho^{-1}\exp\left(-D\right)\times\\&\times\Big\{1+\frac{1}{2}\min_{k=1,2,\cdots}\big[k\exp(\frac{1}{8k}D)\big]+\\&+\frac{1}{2}\min_{k=1,2,\cdots}\big[k\exp\big(\frac{1}{4k+4}D\big)\big]\Big\},\end{aligned}\tag{25}$$

where the constant $D>0$ is defined in (21). Moreover, its expected sample size under H_1 satisfies

$$\begin{aligned}\boldsymbol{E}_1(T_i) \;\le\; &\frac{1}{m}\gamma_i^h+1+\\&+\min_{k=1,2,\cdots}\left(\frac{\gamma_i^h}{2(k+1)m}+\frac{k+1}{2}\rho^{-1}\right),\end{aligned}\tag{26}$$

where ρ is a constant in (19), and m is the Kullback-Leibler divergence in (8).

Theorem 1 deals with the non-asymptotic properties of the CISPRT, and holds for any network structure regardless of how large the number N of sensors is. The original CISPRT upper bounds in (20) and (22) follows directly from Theorem 1 by letting $k=1$. When the upper threshold γ_i^h of the CISPRT or the exponent D in (21) is large, Theorem 1 can be simplified by the following corollary, which will be useful in our asymptotic analysis below.

Corollary 1. *Under the notation of Theorem 1, the CISPRT satisfies*

$$\mathbb{P}_0(\delta_i=1)\le\rho^{-1}\exp\left\{-D+\log\left(\frac{D}{4}+1\right)+1\right\},\tag{27}$$

and

$$\boldsymbol{E}_1(T_i)\le\begin{cases}\frac{5}{4m}\gamma_i^h+1+\frac{1}{\rho}, & \text{if } \gamma_i^h\le m/\rho;\\ \frac{1}{m}\gamma_i^h+1+\sqrt{\frac{1}{m\rho}\gamma_i^h}+\frac{1}{2\rho}, & \text{if } \gamma_i^h>m/\rho.\end{cases}\tag{28}$$

Proof of Corollary 1: In (26), fix a given $k=1,2,\cdots$, we have $1\le\exp(\frac{1}{8k}D)\le\exp(\frac{1}{4k+4}D)$ for $D>0$, and thus

$$\begin{aligned}\mathbb{P}_0(\delta_i=1) \;&\le\; \rho^{-1}\exp(-D)\big[1+\frac{k}{2}+\frac{k}{2}\big]\exp(\frac{1}{4k+4}D)\\&\le\; \rho^{-1}\exp(-D)\exp\big\{\log(u)+\frac{1}{4u}D\big\},\end{aligned}\tag{29}$$

where $u=k+1$. A simple calculus analysis shows that for any given $D>0$, the function $\log(u)+D/(4u)$ is minimized at $u_{opt}=D/4$ with the minimum value of $\log(D/4)+1$. However, a subtly here is that we should restrict u to be integers. The good news is that (29) holds for any integer $u=k+1$ and thus we can choose a specific integer $u^*=\lceil D/4\rceil$, the smallest integer $\ge D/4$. For this specific choice of u^*, we have

$$\begin{aligned}\log(u^*)+\frac{D}{4u^*} \;&\le\; \log(\frac{D}{4}+1)+\frac{D}{4(D/4)}\\&=\; \log(D/4+1)+1.\end{aligned}\tag{30}$$

Combining (29) and (30) yields (27).

To prove (28) from (26), we can choose the integer k in (26) as $k^*=1$ if $\gamma_i^h\le m/\rho$ and as $k^*=\lceil\sqrt{\rho\gamma_i^h/m}-1\rceil$ if $\gamma_i^h>m/\rho$, and by the similar arguments in (30), it is straightforward to see that (28) follows at once from (26). This completes the proof of Corollary 1. □

Now we are ready to present our results under the asymptotic regime when Type I and II error probabilities constraints $\alpha = \beta = \epsilon$ go to 0. The following theorem and its corollaries present the asymptotic properties of the symmetric CISPRT with $\gamma_i^l = -\gamma_i^h$, and indicate that our non-asymptotic upper bounds in (27) and (28) are actually first-order sharp under the asymptotic regime.

Theorem 2. *Under the symmetric local false alarm constraints $\alpha = \beta = \epsilon$ in (2), consider the CISPRT algorithm with the thresholds $\gamma_i^h = -\gamma_i^h$ defined as*

$$\gamma_i^h = -\gamma_i^l = \frac{Nr^2+1}{N}\Delta_\epsilon, \tag{31}$$

where Δ_ϵ is the solution Δ of

$$\Delta - \log(1+\frac{1}{4}\Delta) - 1 = \log(\rho^{-1}) + \log(\epsilon^{-1}). \tag{32}$$

Then this CISPRT algorithm satisfies local false alarm constraints $\alpha = \beta = \epsilon$ in (2), and

$$\lim_{\epsilon\to 0}\frac{\boldsymbol{E}_1(T_i)}{\mathcal{M}(\epsilon)} = Nr^2+1 \tag{33}$$

for all $i = 1, 2, \cdots, N$, as $\alpha = \beta = \epsilon \to 0$.

Corollary 2. *Among the family of the CISPRTs with the weight matrix $\mathbf{W}$ of the form in (17), the optimal one with the smallest asymptotically expected sample size in (33) is attained by the weight matrix $\mathbf{W}_{opt}$ in (17) with*

$$\delta_{opt} = 2/(\lambda_N(\mathbf{L}) + \lambda_2(\mathbf{L})), \tag{34}$$

which yields the minimum r value

$$r_{opt} = \frac{\lambda_N(\mathbf{L}) - \lambda_2(\mathbf{L})}{\lambda_N(\mathbf{L}) + \lambda_2(\mathbf{L})}, \tag{35}$$

where $\lambda_N(\mathbf{L})$ and $\lambda_2(\mathbf{L})$ are the largest and second smallest eigenvalues in (16) for the Laplacian matrix $\mathbf{L}$. In particular, for the complete graph when each sensor is connected to all other sensors, we have $\delta_{opt} = 1/N$ and $r_{opt} = 0$. In such a case, the CISPRT with the optimal weight matrix $\mathbf{W}_{opt}$ in (17) and (34) becomes the centralized SPRT T_c in (5), and attains the universal lower bound $\mathcal{M}(\epsilon)$ as $\alpha = \beta = \epsilon \to 0$.

The proofs of Theorem 2 and its corollary follow from Theorem 1 and other well-known facts, and below we present a high-level sketch of the proofs.

Proof of Theorem 2: Equating the upper bound (27) to ϵ and solving it to yield the desired threshold γ_i^h in (31) and (32). It remains to prove (33) as $\epsilon \to 0$.

Now as $\epsilon \to 0$, it is not difficult to show that $\Delta_\epsilon \sim \log(\epsilon^{-1})$ by (32). Here and below we denote by $x(\epsilon) \sim y(\epsilon)$ if and only if $\lim_{\epsilon\to 0}(x(\epsilon)/y(\epsilon)) = 1$. Hence, by (23), as $\epsilon \to 0$,

$$\gamma_i^h = -\gamma_i^l = (1+o(1))\frac{Nr^2+1}{N}\log(\epsilon^{-1}) \to \infty. \tag{36}$$

Thus it suffices to investigate the asymptotic expression of $\boldsymbol{E}_1(T_i)$ of the symmetric CISPRT with $\gamma_i^l = -\gamma_i^h$ in (36).

On the one hand, our improved upper bound (26) in Theorem 1 implies that

$$\boldsymbol{E}_1(T_i) \le (1+o(1))\frac{1}{m}\gamma_i^h, \tag{37}$$

since $\gamma_i^h \to \infty$. On the other hand, by the false alarm constraints, $\mathbb{P}_1(\delta_i = 0) \le \beta = \epsilon \to 0$ and thus $\mathbb{P}_1(\delta_i = 0)(\gamma_i^h - \gamma_i^l) = o(\gamma_i^h)$ as $\gamma_i^h \to \infty$ when $\gamma_i^l = -\gamma_i^h$. By the lower bound in (22) and relation (36), we have

$$\boldsymbol{E}_1(T_i) \ge (1 + o(1))\frac{1}{m}\gamma_i^h. \tag{38}$$

Relation (33) then follows at once from (36), (37), (38), and Relation (9) which implies $\mathcal{M}(\epsilon) \sim \frac{1}{Nm}\log(\epsilon^{-1})$ as $\epsilon \to 0$. Thus the theorem holds. □

Proof of Corollary 2: By (33), it suffices to minimize the second largest eigenvalue r of $\mathbf{W}$ among all weight matrices $\mathbf{W}$ of the form in (17). It is shown in reference [32] that the corresponding optimal solution and value are given by (34) and (35).

As for the complete graph, its Laplacian matrix is $\mathbf{L} = N\mathbf{I}_{N\times N} - \mathbf{1}\mathbf{1}^T$, where the vector $\mathbf{1}$ is an N-dimensional all one vector. An elementary algebra shows that the eigenvalues of $\mathbf{L}$ are $\lambda_1 = 0$ and $\lambda_2 = \lambda_3 = \cdots = \lambda_N = N$. By (34) and (35), we have $\delta_{opt} = 1/N$ and $r_{opt} = 0$. Hence, for the CISPRT with the optimal weight matrix under the complete graph scenario, each local sensor is to put equal weights to all raw sensor observations, and each and every local sensor essentially runs the optimal centralized SPRT in (5). □

Remark: Note that Corollary 2 implies that the threshold in (31) is first-order asymptotically tight in the sense that if the thresholds $\gamma_i^h = -\gamma_i^l = M_\epsilon\frac{Nr^2+1}{N}\Delta_\epsilon$ with $\liminf_{\epsilon\to 0} M_\epsilon = M < 1$ being a constant that does not depend on N and r, then the corresponding CISPRT algorithm cannot satisfy the local false alarm constraints $\alpha = \beta = \epsilon$ in (2). To see this, (33) would become $\liminf_{\epsilon\to 0}\frac{\boldsymbol{E}_1(T_i)}{\mathcal{M}(\epsilon)} = M(Nr^2+1)$ for any network structures. By Corollary 2, a special case is the complete graph with $r = 0$, which would imply that $\liminf_{\epsilon\to 0}\frac{\boldsymbol{E}_1(T_i)}{\mathcal{M}(\epsilon)} = M$. If $M < 1$, then we would have $\boldsymbol{E}_1(T_i) < \mathcal{M}(\epsilon)$, and this is a contradiction that $\mathcal{M}(\epsilon)$ is the universal lower bounds for any stopping times satisfying the local false alarm constraints $\alpha = \beta = \epsilon$ in (2). Thus, the upper bound (25) on Type I error probability is asymptotically accurate up to first-order in the logarithm scale to derive the thresholds.

IV. Simulation Studies

In this section, we report our simulation study results to illustrate the usefulness of our improved performance properties of CISPRT algorithm.

We use random graph to generate the neighborhood structure of sensor as follow. Assume the N sensors correspond to N random points in a unit square $[0,1]\times[0,1]$. Two sensors are connected if and only if the distance of the corresponding two points is less than the connectivity parameter g. In our simulation studies, we consider two choices of $N = 30$ and 300, and focus on two choices of the connectivity parameter $g = 0.3$ and 0.9. In other words, we consider a total of $2\times 2 = 4$ different distributed network structures. For the comparison, we will also consider the case of $g = 1.5 > \sqrt{2}$, in which the network becomes the centralized system. For each network structure, the raw sensor observations $y_i(t)$'s are assumed to be $N(\mu_1, \sigma^2)$ under H_1 and $N(-\mu_1, \sigma^2)$ under H_0. Here we set $\mu_1 = 1/\sqrt{2}$ and $\sigma = 1$, so that the local Kullback-Leibler divergence in (8) becomes $m = 2\mu^2/\sigma^2 = 1$ at each local sensor. In each of these networks, for simplicity, we consider the symmetric scenario when $\alpha = \beta = \epsilon$ and when the lower and upper bounds γ_i^l and γ_i^h of the CISPRT in (14) are given by $\gamma_i^l \equiv -\gamma$ and $\gamma_i^h \equiv \gamma$ for all $i = 1, \cdots, N$ for some $\gamma > 0$.

In order to validate the non-asymptotic upper bound in (27) or (25) on the Type I error probability of the CISPRT, Table I reports the Monte Carlo (MC) estimates of the Type I error probabilities of the CISPRT when the thresholds $\gamma_i^h = -\gamma_i^l$ are given by (23) when $\alpha = \beta = \epsilon = 5\%$. Two kinds of Monte Carlo estimates (MCE) will be reported based on 10^6 runs: one is

Network structures	$N=30$			$N=300$		
	$g=1.5$	$g=0.9$	$g=0.3$	$g=1.5$	$g=0.9$	$g=0.3$
$r=r_{opt}$ in (35)	0	0.1406	0.9384	0	0.2027	0.8588
$\gamma=\gamma_i^h=-\gamma_i^l$ in (23)	0.1405	0.1823	5.3471	0.0141	0.1874	4.1826
Classical Central Bound in (6)	0.0148	–	–	0.0146	–	–
Original bound in (20)	0.0500	0.0500	0.0500	0.0500	0.0500	0.0500
Our improved bound (25)	0.0398	0.0370	0.0397	0.0398	0.0398	0.0373
Our loose bound (27)	0.0824	0.0805	0.0824	0.0824	0.0824	0.0809
MCE1 of $\mathbb{P}_0(\delta_i=1)$	5×10^{-6}	3×10^{-6}	0	0	0	0
MCE2 of $\mathbb{P}_0(\delta_i=1)$	5.14×10^{-6}	1.686×10^{-6}	5.354×10^{-43}	1.264×10^{-84}	1.264×10^{-84}	0

TABLE I: This table reports different upper bounds and two Monte Carlo estimates (MCE) of Type I error probability of the CISPRT with thresholds in (23): one is the naive method (MCE1) and the other is the importance sampling method (MCE2). It confirms that both the original bound in (20) and our upper bounds in (25) or (27) do provide the upper bounds on the Type I error probability of the CISPRT. Unfortunately, all upper bounds are very loose, and it is still an open problem to find accurate upper bound on error probabilities in the finite-sample regime when the number N of sensors is large.

the naive method based on how many of 10^6 runs makes a false decision, and the other is the importance sampling technique: for any (centralized or distributed) stopping time T,

$$\begin{aligned}\mathbb{P}_0(\delta_i=1) &= \mathbb{E}_1(\prod_{t=1}^{T}\prod_{i=1}^{N}\log\frac{f_0(y_i(t))}{f_1(y_i(t))};\delta_i=1)\\ &= \mathbb{E}_1\left(\exp(-S_c(T))\mathbf{1}\{\delta_i=1\}\right),\end{aligned} \tag{39}$$

where $S_c(T)$ is the centralized likelihood ratio statistic in (4) at the stopping time T. The importance sampling technique in (39) allows us to provide an accurate estimate of Type I error probabilities when they are very small, see [15].

From Table I, our improved non-asymptotic upper bound in (25) is less than the original bound in (20) for the CISPRT, and both are indeed the upper bounds of Type I error probability of the CISPRT. Unfortunately, while our loose upper bound in (27) is first-order tight as in Corollary 2, the convergence rate is extremely slow, especially when the number N of sensors is large. Note that this slow convergence rate is intrinsic for high-dimension N under the centralized setting too. For instance, the classical bounds in (6) for the centralized SPRT test are known to be first-order asymptotically sharp when $\log\epsilon^{-1}$ is much larger than N, but Table I shows that they were poor estimates in the finite-sample setting for the centralized system with $g=1.5$, which are much larger than the MCE. Thus it remains an open problem to find accurate estimates of Type I and Type II error probabilities in centralized or distributed system under the modern asymptotic setting when the number N of sensors goes to ∞.

In order to validate the non-asymptotic upper bound in (26), we simulate the expected sample size of the CISPRT in (14) with threshold γ^h varying from 0 to 100 with step size 1 under H_1 based on $M=2000$ MC runs. Figure 1 plots three different estimating curves of $\boldsymbol{E}_1(T_i)$ of the CISPRT as the function of γ^h: (i) the Monte Carlo simulated expected sample size $\boldsymbol{E}_1(T_i)$ (purple dotted line); (ii) Sahu and Kar's upper bound in (22) (blue dashed line); and (iii) our improved upper bound in (26) (red solid line). The lower bound in (22) contains some constants that cannot be computed, and thus we focus on our improved upper bounds. From Figure 1, our upper bounds on $\boldsymbol{E}_1(T_i)$ are indeed larger than the Monte Carlo simulated expected sample size $\boldsymbol{E}_1(T_i)$, and are also better than Sahu and Kar's original upper bound, no matter the number of sensors and the connectivity parameter.

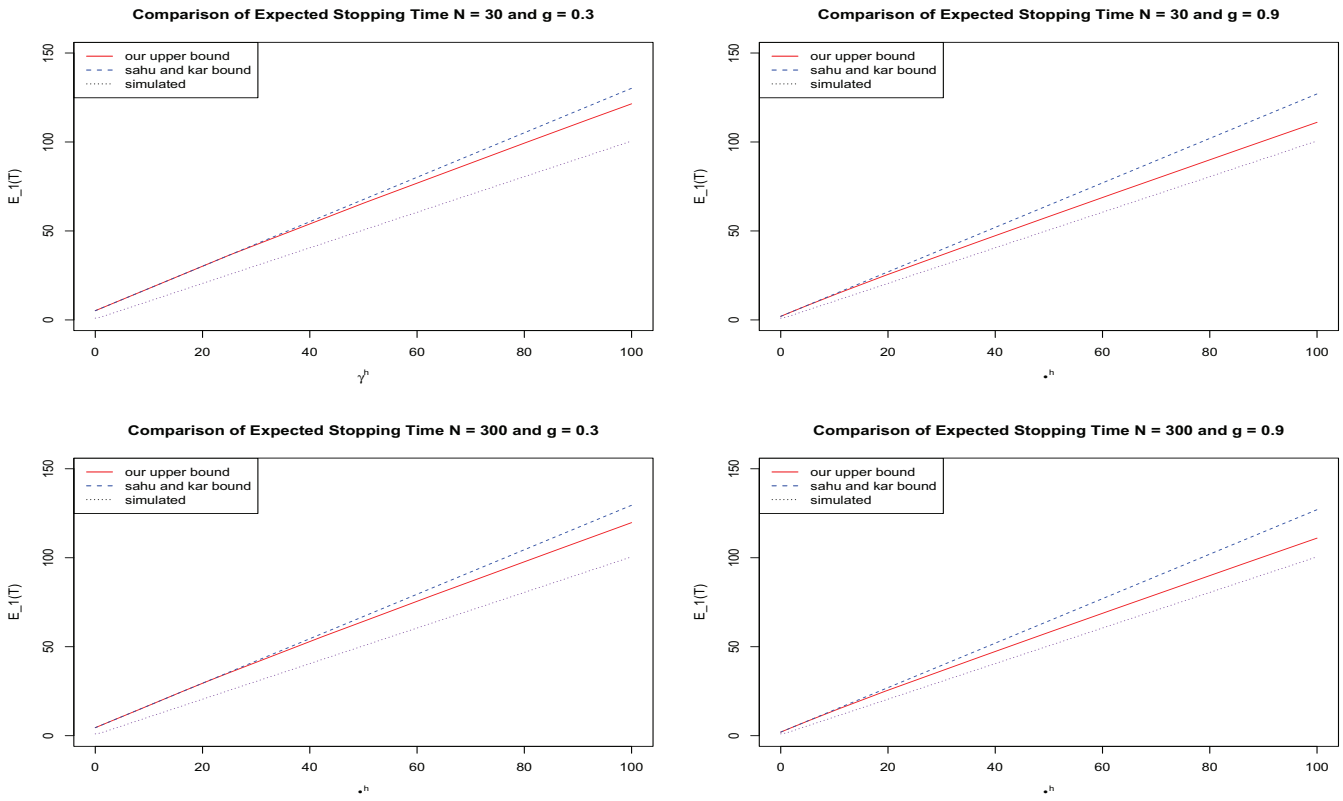

Fig. 1: A comparison of three non-asymptotic bounds or estimates of $\boldsymbol{E}_1(T_i)$ of the CISPRT under four different setting of random graph depending on the number N of sensors and the connectivity parameter g. In each plot, three curves represent three different bounds/estimates as threshold γ^h varies from 0 to 100 with step size 1, and these three methods ranking from largest to smallest are as follows: (i) The blue dashed line is Sahu and Kar's upper bound in (22); (ii) The red solid line is our improved bound in (26); and (iii) The purple dotted line is the Monte Carlo simulated estimate of $\boldsymbol{E}_1(T_i)$. The plots validate our improved non-asymptotic upper bound in (26), though the convergence rate seems to be slow.

V. Proof of Theorem 1

This section is devoted to prove Theorem 1 under the non-asymptotic setting. Let us fix the i-th local sensor, and investigate the properties of the stopping time T_i in (13) of the CISPRT at this specific local sensor that are related to the upper bound γ_i^h. Since the detailed proof is technical and involves many subscripts, we decide to abuse the notation and denote the stopping time T_i and the upper bound γ_i^h in (13) simply by T and γ.

Let us first provide the high-level idea to prove Theorem 1. Note that the Type I error probability can be written as

$$\mathbb{P}_0(\delta_i = 1) = \mathbb{P}_0(S_i(T) \geq \gamma), \tag{40}$$

where $S_i(T)$ is the value of the local test statistic in (10) at the stopping time T. Note that in the classical sequential analysis for the centralized setting, it is standard to use the change of measures arguments, and rewrite the Type I error probability as its equivalent form of $\boldsymbol{E}_1(e^{-S_c(T)}I(S(T) \geq \gamma))$, where $S_c(T)$ is the centralized log-likelihood ratio in (4). The analysis on this error probability analysis and the expected sample size $\mathbf{E}_1(T)$ is then based on the renewal theory and overshoot analysis for random walks over time t. Unfortunately, such approach breaks down for distributed setting when the centralized test statistic $S_c(t)$ and the local test statistic $S(t)$ can be completely different. Moreover, with large number N sensors, the overshoot

analysis often involves constants that are exponentially increasing as N increases and thus the corresponding analysis can be meaningless under the practical setting.

Sahu and Kar [22] proposed an alternative method to bound the Type I error probability and expected sample size directly. Specifically, note that

$$\begin{aligned}\mathbb{P}_0(\delta_i = 1) &= \mathbb{P}_0(S_i(T) \geq \gamma) \\ &= \sum_{t=1}^{\infty} \mathbb{P}_0(T = t, S_i(t) \geq \gamma) \\ &\leq \sum_{t=1}^{\infty} \mathbb{P}_0(S_i(t) \geq \gamma) \\ &= \sum_{t=1}^{\infty} \mathbb{Q}\left(\frac{\gamma - \mu_0^*(t)}{\sqrt{V_0^*(t)}}\right). \end{aligned} \tag{41}$$

Here $\mathbb{Q}(u) = \boldsymbol{P}(N(0,1) > u)$, and the local test statistics $S_i(t)$ are Gaussian distributed under H_0, say, $N(\mu_0^*(t), V_0^*(t))$, at any fixed t, since the raw sensor observations are Gaussian. Meanwhile, the expected sample size, $\boldsymbol{E}_1(T)$, is bounded by

$$\begin{aligned}\boldsymbol{E}_1(T) &= \sum_{t=0}^{\infty} \boldsymbol{P}_1(T > t) \\ &\leq \sum_{t=0}^{\infty} \boldsymbol{P}_1(S_i(t) < \gamma) \\ &\leq \sum_{t=0}^{\infty} \mathbb{Q}\left(\frac{\gamma - \mu_1^*(t)}{\sqrt{V_1^*(t)}}\right), \end{aligned} \tag{42}$$

where the local test statistics $S_i(t)$ are Gaussian $N(\mu_1^*(t), V_1^*(t))$ under H_1.

Due to the similarity between (41) and (42), below we will focus on the Type I error probability analysis in (41). By (8) and (10), it was showed in [22] that

$$\mu_0^*(t) = -mt \qquad \text{and} \qquad V_0^*(t) \leq \frac{2m(Nr^2+1)t}{N}. \tag{43}$$

In [22], the authors then combined these above results together to bound the infinite sum in (41) by splitting the interval $t \in [1, \infty)$ into four subintervals:

$$[1, \frac{\gamma}{2m}], \quad (\frac{\gamma}{2m}, \frac{\gamma}{m}], \quad (\frac{\gamma}{m}, \frac{2\gamma}{m}], \quad (\frac{2\gamma}{m}, \infty). \tag{44}$$

Bounding the sum in each of these four subintervals leads to the result in (20) in the original CISPRT paper.

The direct approach in (41) - (44) is non-asymptotic, but unfortunately it is too crude in general. Indeed, if we apply them directly to investigate the Type I error probability or expected sample sizes of the centralized SPRT, the results will be much looser as compared with those from the classical renewal theory: while the first-order terms have the same order, the coefficients from the direct approach in (41) - (44) are much larger.

After a careful analysis, we find out that the main reason is caused by the middle two subintervals in (44), and the direct approach in (41) - (44) can be refined to provide better bounds if we further split each of the middle two subintervals into k sub-subintervals, for some suitable choice of k that will be optimally determined later. In fact, when we applied this new refined approach to investigate the Type I error probability or expected sample sizes of the centralized SPRT, then the corresponding results are first-order asymptotically equivalent to those from the classical renewal theory. This suggests that the refined direct approach yields an accurate upper bound for complete graph regardless of the number N of sensors, and thus may also lead to good bounds for other network structures. We acknowledge that the proof techniques are essentially that of [22], except for the new found techniques as far as splitting the intervals are concerned in (45) below.

Now we are ready to provide the detailed, rigorous proof of (25). First, we further split each of the middle two subintervals into k sub-subintervals as follows. Let $\ell = \frac{\gamma}{2m}$ or $\frac{\gamma}{m}$, and we propose to further split the subinterval $[\ell, 2\ell]$ as k sub-subintervals:

$$(\frac{k+j-1}{k}\ell,\ \frac{k+j}{k}\ell] \qquad \text{for } j = 1, \cdots, k. \tag{45}$$

Relation (25) in Theorem 1 can be proved by bounding the infinite sum in (41) through these subintervals.

Second, we will use heavily the following well-known fact for $N(0,1)$ distribution:

$$\mathbb{Q}(x) = \boldsymbol{P}(N(0,1) > x) \le \frac{1}{2}\exp\Big(-\frac{x^2}{2}\Big) \text{ for all } x > 0. \tag{46}$$

Also $Q(x)$ is decreasing as a function of x, and thus replacing $V_0^*(t)$ by its upper bound in (43) yields an upper bound of (41).

Next, by (41), (43) and (46), we have

$$\begin{aligned}
\mathbb{P}_0(\delta_i = 1) &\le \sum_{t=1}^{\infty} \mathbb{Q}\left(\frac{\gamma + mt}{\sqrt{\frac{2mt(Nr^2+1)}{N}}}\right) \\
&\le \frac{1}{2}\sum_{t=1}^{\infty} \exp\left\{\frac{-N(\gamma+mt)^2}{4mt(Nr^2+1)}\right\} \\
&= \frac{1}{2}(A_1 + A_2 + A_3 + A_4) \\
&= \frac{1}{2}(A_1 + \sum_{j=1}^{k} A_{2j} + \sum_{j=1}^{k} A_{3j} + A_4),
\end{aligned} \tag{47}$$

where A_1, A_2, A_3 and A_4 denote the corresponding sum when the integer index t ranges over the subintervals in (44). Here A_{2j} and A_{3j} are defined as the summation over the sub-subintervals in (45) for $\ell = \frac{\gamma}{2m}$ or $\frac{\gamma}{m}$. In other words,

$$A_{2j} = \sum_{t=\lfloor \frac{k+j-1}{k}\ell \rfloor + 1}^{\lfloor \frac{k+j}{k}\ell \rfloor} \exp\left\{\frac{-N(\gamma+mt)^2}{4mt(Nr^2+1)}\right\}$$

with $\ell = \frac{\gamma}{2m}$, and A_{3j} is defined similarly with $\ell = \frac{\gamma}{m}$. Here and below $\lfloor x \rfloor$ denotes the largest integer $\le x$.

Sahu and Kar [22] proved their results by bounding A_1, A_2, A_3 and A_4, and here we refine their results by bounding A_{2j}'s and A_{3j}'s. The main mathematical tool is the simple fact that when $a \le t \le b$, for $c = N/(4(Nr^2+1))$,

$$\begin{aligned}
&\sum_{t=a}^{b} \exp\left\{-c\frac{(\gamma+mt)^2}{mt}\right\} \\
\le\ &\sum_{t=a}^{b} \exp\left\{-c\big(\frac{\gamma^2}{mb} + 2\gamma + mt\big)\right\} \\
=\ &\exp\left\{-c\big(\frac{\gamma^2}{mb} + 2\gamma\big)\right\} \frac{\exp(-cma) - \exp(-cm(b+1))}{1-\exp(-cm)} \\
\le\ &\rho^{-1}\exp\left\{-c\big(\frac{\gamma^2}{mb} + 2\gamma\big)\right\} \exp(-cma),
\end{aligned} \tag{48}$$

where the constant $\rho = 1 - \exp(-cm)$ is defined in (19).

In order to help casual readers better understand our main ideas, let us first provide the bounds of the original CISPRT paper [22] on A_1 and A_2. Applying (48) to the case of $a = 1$ and $b = \gamma/(2m)$, we have

$$\begin{aligned}
A_1 &\le \rho^{-1}\exp(-4c\gamma)\exp(-cm) \\
&\le \rho^{-1}\exp\Big(-\frac{N\gamma}{Nr^2+1}\Big) = \rho^{-1}\exp(-D),
\end{aligned} \tag{49}$$

where the second to last relation follows from the fact that $\exp(-cm) < 1$ and the definition of $c = N/(4(Nr^2+1))$, and the term D is defined in (21) with $\gamma = \gamma_i^h$.

Similarly, applying (48) to the case of $a = \frac{\gamma}{2m}$ and $b = \frac{\gamma}{m}$, we have

$$\begin{aligned} A_2 &\leq \rho^{-1}\exp(-3c\gamma)\exp(-\frac{1}{2}c\gamma) \\ &= \rho^{-1}\exp(-\frac{7}{2}c\gamma) \\ &= \rho^{-1}\exp\Big(-\frac{7}{8}\frac{N\gamma}{Nr^2+1}\Big) = \rho^{-1}\exp(-\frac{7}{8}D). \end{aligned} \tag{50}$$

It is easy to see that A_3 satisfies (50), whereas A_4 satisfies (49). A combination of (47) with the bounds in (49) and (50) yields (20), which is the upper bound of $\mathbb{P}_0(\delta_i = 1)$ derived in [22].

To improve the upper bound in (20) of [22], our key observation is that the bound in (50) for A_2 and A_3 can be further reduced. For that purpose, let us consider the A_{2j} over the j-th sub-subinterval in (45), and apply (48) to the case of $a = \frac{k+j-1}{k}\frac{\gamma}{2m}$ and $b = \frac{k+j}{k}\frac{\gamma}{2m}$. Then for $j = 1, 2, \cdots, k$, we have

$$\begin{aligned} A_{2j} &\leq \rho^{-1}\exp\Big(-2c\gamma\frac{2k+j}{k+j}\Big)\exp\Big(-c\gamma\frac{k+j-1}{2k}\Big) \\ &= \rho^{-1}\exp(-2c\gamma)\exp\Big(-c\gamma(\frac{2k}{k+j}+\frac{k+j-1}{2k})\Big) \\ &\leq \rho^{-1}\exp(-2c\gamma)\exp\Big(-c\gamma(2-\frac{1}{2k})\Big) \\ &= \rho^{-1}\exp\Big\{-\frac{8k-1}{8k}D\Big\}, \end{aligned} \tag{51}$$

where the second to last relation follows from the simple fact that $u + 1/u \geq 2$ for $u = 2k/(k+j)$, and the last equation is from the definition of $c = N/(4(Nr^2+1))$ and D in (21).

Hence, we have

$$\begin{aligned} A_2 &= \sum_{j=1}^{k} A_{2j} \\ &\leq \rho^{-1}k\exp\Big\{-\frac{8k-1}{8k}D\Big\} \end{aligned} \tag{52}$$

for any $k = 1, 2, \cdots$. Similarly, the same technique of (51) is applied to A_{3j} or A_3, and we have

$$A_3 \leq \rho^{-1}k\exp\Big\{-\frac{4k+3}{4k+4}D\Big\}. \tag{53}$$

for any $k = 1, 2, \cdots$. Relation (25) in Theorem 1 then follows at once from (49), (52) and (53) by taking the minimum values in the last two equations over all possible k's.

The proof of (26) for the expected sample size is similar, except with different subintervals. By (8) and (10), we can show that

$$\mu_1^*(t) = mt \qquad \text{and} \qquad V_1^*(t) \leq \frac{2m(Nr^2+1)t}{N}. \tag{54}$$

By (42), for the CISPRT,

$$\begin{aligned} \boldsymbol{E}_1(T) &\leq \sum_{t=0}^{\infty}\mathbb{Q}(\frac{tm-\gamma}{\sqrt{2tm(Nr^2+1)/N}}) \\ &= B_1 + B_2 + B_3 + B_4, \end{aligned} \tag{55}$$

where B_1, B_2, B_3, B_4 correspond to the summation over t in each of the following four subintervals:

$$[0, \frac{\gamma}{m}], \quad (\frac{\gamma}{m}, \frac{3\gamma}{2m}], \quad (\frac{3\gamma}{2m}, \frac{2\gamma}{m}], \quad (\frac{2\gamma}{m}, \infty). \tag{56}$$

It turns out that the bounds derived in [22] for B_1, B_3, B_4 are tight, and the bound on B_2 is too loose. To be more specific, in [22], the authors used the similar technique for Type I error to show that $B_3 \le \frac{1}{2}\rho^{-1}$ and $B_4 \le \frac{1}{2}\rho^{-1}$, and also bound B_1 and B_2 by

$$\begin{aligned} B_1 &= \sum_{t=0}^{\lfloor \gamma/m \rfloor} \mathbb{Q}\left(\frac{tm-\gamma}{\sqrt{2tm(Nr^2+1)/N}}\right) \\ &\le \sum_{t=0}^{\lfloor \gamma/m \rfloor} 1 \le \frac{\gamma}{m}+1, \\ B_2 &= \sum_{t=\lfloor \gamma/m \rfloor+1}^{\lfloor 3\gamma/(2m) \rfloor} \mathbb{Q}\left(\frac{tm-\gamma}{\sqrt{2tm(Nr^2+1)/N}}\right) \\ &\le \sum_{t=\lfloor \gamma/m \rfloor+1}^{\lfloor 3\gamma/(2m) \rfloor} \frac{1}{2} = \frac{\gamma}{4m}, \\ B_3 &\le \frac{1}{2}\rho^{-1}, \quad \text{and} \quad B_4 \le \frac{1}{2}\rho^{-1} \end{aligned} \tag{57}$$

Here B_1 and B_2 are based on the two simple facts: (1) $Q(u) = \boldsymbol{P}(N(0,1) > u) \le 1$ for all $-\infty < u < \infty$ and (2) $Q(u) \le 1/2$ when $u > 0$.

To improve the upper bound of B_2, we propose to further split the subinterval $(\frac{\gamma}{m}, \frac{3\gamma}{2m}]$ into k sub-subintervals:

$$\left(\frac{\gamma}{m}, \frac{k+2}{k+1}\frac{\gamma}{m}\right] \text{ and } \left(\frac{j+2}{j+1}\frac{\gamma}{m}, \frac{j+1}{j}\frac{\gamma}{m}\right], \text{ for } j = 2, \cdots, k. \tag{58}$$

Denote by $B_2^{(1)}$ and $B_2^{(j)}$ the summation as in B_2 in (57) except when t is over the first and j-th sub-interval in (58), respectively, for $j = 2, \cdots, k$. For the first subinterval in (58), by the simple fact that $\mathbb{Q}(u) \le \frac{1}{2}$ when $u \ge 0$, we have

$$B_2^{(1)} = \frac{1}{2}\left(\frac{1}{k+1}\frac{\gamma}{m}\right). \tag{59}$$

For the j-th subinterval in (58) with $j = 2, \cdots, k$, we propose to explore relation (46), which provides a much improved bound for $Q(u)$ than the constant $1/2$ when $u > 0$. That is, by (46), for $j = 2, \cdots, k$, we have

$$B_2^{(j)} \le \frac{1}{2} \sum_{\lfloor \frac{j+2}{j+1}\frac{\gamma}{m} \rfloor + 1}^{\lfloor \frac{j+1}{j}\frac{\gamma}{m} \rfloor} \exp\left\{\frac{-N(\gamma - mt)^2}{4mt(Nr^2+1)}\right\} \tag{60}$$

Our remaining arguments are similar to those in (48), with a minor twist to reflect the change of mean from $\mu_0^*(t)$ to $\mu_1^*(t)$. To be more specific, as in (48), it is not difficult to see that

$$\begin{aligned} &\sum_{t=a}^{b} \exp\left\{-c\frac{(\gamma - mt)^2}{mt}\right\} \\ \le\ & \rho^{-1} \exp\left\{-c\left(\frac{\gamma^2}{mb} - 2\gamma\right)\right\} \exp(-cma) \\ =\ & \rho^{-1}, \end{aligned} \tag{61}$$

when $a = \frac{j+2}{j+1}\frac{\gamma}{m}$ and $b = \frac{j+1}{j}\frac{\gamma}{m}$. Combining this with (60) yields that

$$B_2^{(j)} \le \frac{1}{2}\rho^{-1}, \tag{62}$$

where the right-hand side upper bound does not depend on γ. By (59) and (62), we have

$$B_2 \le \frac{1}{2}\frac{1}{k+1}\frac{\gamma}{m} + \sum_{j=2}^{k}\frac{1}{2}\rho^{-1}$$

$$= \frac{1}{k+1}\frac{\gamma}{2m} + (k-1)\frac{1}{2}\rho^{-1}. \tag{63}$$

Hence, by (55), (57) and (63), the expected sample size of T under H_1 satisfies

$$\begin{aligned}
\boldsymbol{E}_1(T) &\leq B_1 + B_2 + B_3 + B_4 \\
&\leq \frac{\gamma}{m} + 1 + \left[\frac{1}{k+1}\frac{\gamma}{2m} + \frac{k-1}{2}\rho^{-1}\right] + \frac{1}{2}\rho^{-1} + \frac{1}{2}\rho^{-1} \\
&= \frac{\gamma}{m} + \frac{\gamma}{2(k+1)m} + \frac{k+1}{2}\rho^{-1} + 1, \qquad (64)
\end{aligned}$$

for any integer $k \geq 1$. When $k = 1$, this is just the upper bound of $\boldsymbol{E}_1(T)$ in (22) derived by [22]. Relation (26) follows directly from (64) which holds for any $k = 1, 2, \cdots$. This completes the proof of the theorem.

VI. Conclusions

In this article, we investigate the performance properties of the CISPRT algorithm proposed by [22]. Our focus is on improving the non-asymptotic upper bounds on the error probabilities and expected sample sizes of the CISPRT algorithm through a novel approach to bound the infinite sum of tail probabilities of Gaussian distributions, and it turns out that our improved upper bound are first-order asymptotic sharp. Our results show that the more the number of sensors or the sparser the network neighborhood connectivity is, the larger the information loss is, i.e., the larger expected sample size is needed to achieve the desired Type I and II error probabilities.

Several future directions can be pursued for distributed sequential detection. First, it will be interesting to provide more accurate approximations on the performance analysis of the CISPRT algorithm for any network structures, especially under the modern asymptotic setting when the number of sensors goes to ∞. New techniques and new ideas will need to be developed. Second, instead of binary simple hypothesis testing, it will be interesting to develop efficient algorithms when there are nuisance parameters in the alternative hypothesis. Third, here we assume that all local sensors having different distributions simultaneously under the alternative hypothesis H_1, and in some applications, one may want to develop efficient algorithms where only a few unknown subset of local sensors have distributions from H_1. This might be closely related to sparsity detection or false discovery rate in the modern statistics literature. Fourth, in our paper the neighborhood or network structure is pre-specified, and it will be useful to investigate the time-varying network structure. Finally, it will also be interesting to adapt the analysis and the technical tools in this paper to non-linear observations models or non-Gaussian noises.

References

[1] Y. Liang, L. Lai, and J. Halloran, "Distributed algorithm for collaborative detection in cognitive radio networks," in *2009 47th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, Sept 2009, pp. 394–399.

[2] ——, "Distributed cognitive radio network management via algorithms in probabilistic graphical models," *IEEE Journal on Selected Areas in Commun.*, vol. 29, no. 2, pp. 338–348, February 2011.

[3] C. Tekin, S. Zhang, and M. van der Schaar, "Distributed online learning in social recommender systems," *IEEE Journal of Selected Topics in Signal Process.*, vol. 8, no. 4, pp. 638–652, Aug 2014.

[4] F. Yan, S. Sundaram, S. V. N. Vishwanathan, and Y. Qi, "Distributed autonomous online learning: Regrets and intrinsic privacy-preserving properties," *IEEE Trans. on Knowl. and Data Eng.*, vol. 25, no. 11, pp. 2483–2493, Nov 2013.

[5] K. Gimpel, D. Das, and N. A. Smith, "Distributed asynchronous online learning for natural language processing," in *Proc. of the 14th Conference on Computational Natural Language Learning*. Association for Computational Linguistics, 2010, pp. 213–222.

[6] A. Galstyan, B. Krishnamachari, K. Lerman, and S. Pattem, "Distributed online localization in sensor networks using a moving target," in *3rd International Symposium on Inf. Proc. in Sensor Networks, 2004. IPSN 2004*, April 2004, pp. 61–70.

[7] M. Rabbat and R. Nowak, "Distributed optimization in sensor networks," in *3rd International Symposium on Inf. Proc. in Sensor Networks, 2004. IPSN 2004*, April 2004, pp. 20–27.

[8] Y. Zhang, D. Sow, D. Turaga, and M. van der Schaar, "A fast online learning algorithm for distributed mining of bigdata," *ACM SIGMETRICS Performance Evaluation Review*, vol. 41, no. 4, pp. 90–93, 2014.

[9] P. D. Lorenzo and S. Barbarossa, "Distributed estimation and control of algebraic connectivity over random graphs," *IEEE Trans. on Signal Process.*, vol. 62, no. 21, pp. 5615–5628, Nov 2014.

[10] M. G. Rabbat and R. D. Nowak, "Decentralized source localization and tracking [wireless sensor networks]," in *2004 IEEE International Conference on Acoustics, Speech, and Signal Process.*, vol. 3, May 2004, pp. iii–921–4 vol.3.

[11] R. S. Blum, S. A. Kassam, and H. V. Poor, "Distributed detection with multiple sensors i. advanced topics," *in Proc. of the IEEE*, vol. 85, no. 1, pp. 64–79, Jan 1997.

[12] R. Viswanathan and P. K. Varshney, "Distributed detection with multiple sensors i. fundamentals," *in Proc. of the IEEE*, vol. 85, no. 1, pp. 54–63, Jan 1997.

[13] Q. Zhou, D. Li, S. Kar, L. Huie, H. V. Poor, and S. Cui, "Learning-based distributed detection-estimation in sensor networks with unknown sensor defects," *arXiv preprint arXiv:1510.02371*, 2015.

[14] A. Wald, "Sequential tests of statistical hypotheses," *The Annals of Mathematical Statistics*, vol. 16, no. 2, pp. 117–186, 1945.

[15] D. Siegmund, *Sequential Analysis: Tests and Confidence Intervals*. Springer Science & Business Media, 2013.

[16] M. Basseville, I. V. Nikiforov *et al.*, *Detection of Abrupt Changes: Theory and Application*. Prentice Hall Englewood Cliffs, 1993, vol. 104.

[17] A. Tartakovsky, I. Nikiforov, and M. Basseville, *Sequential analysis: Hypothesis testing and Changepoint Detection*. CRC Press, 2014.

[18] G. Fellouris, "Asymptotically optimal parameter estimation under communication constraints," *The Annals of Statistics*, vol. 40, no. 4, pp. 2239–2265, 2012.

[19] H. R. Hashemi and I. B. Rhodes, "Decentralized sequential detection," *IEEE Trans. on Inf. Theory*, vol. 35, no. 3, pp. 509–520, May 1989.

[20] Y. Mei, "Asymptotic optimality theory for decentralized sequential hypothesis testing in sensor networks," *IEEE Trans. on Inf. Theory*, vol. 54, no. 5, pp. 2072–2089, May 2008.

[21] V. V. Veeravalli, T. Basar, and H. V. Poor, "Decentralized sequential detection with a fusion center performing the sequential test," in *1992 American Control Conference*, June 1992, pp. 1177–1181.

[22] A. K. Sahu and S. Kar, "Distributed sequential detection for gaussian shift-in-mean hypothesis testing," *IEEE Trans. on Signal Process.*, vol. 64, no. 1, pp. 89–103, Jan 2016.

[23] S. Kar, S. Aldosari, and J. M. F. Moura, "Topology for distributed inference on graphs," *IEEE Trans. on Signal Process.*, vol. 56, no. 6, pp. 2609–2613, June 2008.

[24] S. Kar and J. M. Moura, "Consensus based detection in sensor networks: Topology optimization under practical constraints," in *Proc. 1st Intl. Wrkshp. Inform. Theory Sensor Networks*, 2007.

[25] S. Kar and J. M. F. Moura, "Sensor networks with random links: Topology design for distributed consensus," *IEEE Trans. on Signal Process.*, vol. 56, no. 7, pp. 3315–3326, July 2008.

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

[27] V. Matta, P. Braca, S. Marano, and A. H. Sayed, "Diffusion-based adaptive distributed detection: Steady-state performance in the slow adaptation regime," *IEEE Trans. on Inf. Theory*, vol. 62, no. 8, pp. 4710–4732, Aug 2016.

[28] S. Li and X. Wang, "Fully distributed sequential hypothesis testing: Algorithms and asymptotic analyses," *IEEE Trans. on Inf. Theory*, vol. 64, no. 4, pp. 2742–2758, February 2018.

[29] A. K. Sahu and S. Kar, "Recursive distributed detection for composite hypothesis testing: Nonlinear observation models in additive gaussian noise," *IEEE Trans. on Inf. Theory*, vol. 63, no. 8, pp. 4797–4828, March 2017.

[30] M. Mesbahi and M. Egerstedt, *Graph Theoretic Methods in Multiagent Networks*. Princeton University Press, 2010.

[31] F. R. Chung, *Spectral Graph Theory*. American Mathematical Soc., 1997, no. 92.

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