Distributed Change Detection Based on Average Consensus

Qinghua Liu∗ and Yao Xie†

November, 2017

Abstract

Distributed change-point detection has been a fundamental problem when performing real-time monitoring using sensor-networks. Most existing algorithms require local sensors to exchange information with a global fusion center. We present a distributed detection algorithm, where each sensor only exchanges CUSUM statistic with their neighbors based on average consensus, and an alarm is fired when local statistic exceeds a pre-specified global threshold. We provide theoretical performance bounds showing that the performance of the fully distributed scheme can match the centralized algorithms under some mild conditions. Numerical experiments demonstrate the good performance of the algorithm.

Keywords: Change-point sequential detection, local communication, distributed network, CUSUM decision

1 Introduction

Detecting an abrupt change from data collected by sensor-networks has been a fundamental problem in various applications, such as cybersecurity [1, 2] and environmental monitoring [3]. The goal is to detect the change as quickly as possible while the false alarm is under control.

Various distributed change-point detection methods have been developed based on the classic CUSUM and Shiryaev-Roberts procedures [4, 5]. Most of the existing approaches [2, 6–8] require a fusion center that gathers information (raw data or statistics) from all sensors to perform decision centrally. For instance, [8] requires a global center to collect all the local CUSUM statistics from all the sensors and utilizes their sum to make decision. This may be difficult to implement with limited communication bandwidth: for instance, when some nodes are too far from the central hub. On the other hand, there are approaches where each sensor makes decision using its own data and only transmits a one-bit signal to central hub once a local alarm has been trigged (e.g., [7, 9]). However, this approach can be improved if neighboring sensors can exchange information, which has become practical in sensor networks with the ever developing technology.

In this paper, we study a new distributed multi-sensor change-point detection procedure based on the so-called average consensus scheme [10], by allowing sensors to exchange their local CUSUM statistics. This scheme does not involve explicit point-to-point message passing or routing; instead, it diffuses information across the network by updating their

∗Department of Electronic Engineering, Tsinghua University, China; Email: liu-qh14@mails.tsinghua.edu.cn
†ISYE, Georgia Institute of Technology, USA; Email: yao.xie@isye.gatech.edu.
Figure 1: Distributed change-point detection (a) with a fusion center in [7], where all sensors send data to central hub, and (b) without fusion center, where sensors perform local consensus average.

own statistics by performing a weighted average of neighbors’ statistics [11]. Therefore, each sensor performs detection locally using only the consensus result available to itself. An alarm is fired whenever any sensor detects a change. Hence, our scheme requires no global fusion center and allows each sensor to utilize neighboring data. Figure 1 demonstrates the different graph topologies of our approach and the conventional scheme with a fusion center. We demonstrate the good performance of our proposed method via numerical examples.

The main theoretical contributions of the paper are the analysis of our detection procedure in terms of the two fundamental performance metrics: the average run length (the reciprocal of false alarm rate) and the expected detection delay. We show that for a system consisting of $N$ sensors, using the average consensus scheme, the expected detecting delay can nearly be reduced by a factor of $N$ compared to a system without communication, under the same false alarm rate.

Some recent works [12–14] study a related but different problem, the distributed sequential hypothesis test based on average consensus, which aims to resolve two hypotheses. It can be viewed as a case when the change-point happens at the first moment. There exists a major difference in our analysis from the analysis for this problem. In sequential hypothesis test, the local log-likelihood statistic accumulates linearly, while in sequential change-point detection, the local detection statistic is recursively computed using CUSUM, and the local likelihood statistic accumulates nonlinearly due to a non-linear transformation in the CUSUM procedure. This results in a more challenging case for analysis and requires significantly different techniques.

2 Preliminary

We first introduce some necessary notations and background.

**Definition 1. Log-likelihood ratio (LLR)** Suppose we have two distinct distributions $\mathcal{P}_1$ and $\mathcal{P}_2$. Define the probability density function of $\mathcal{P}_1$ and $\mathcal{P}_2$ as $f_1(x)$ and $f_2(x)$, respectively. Then the log-likelihood ratio function (LLR) between distribution $\mathcal{P}_2$ and $\mathcal{P}_1$ is defined as

$$L(x) = \log \frac{f_2(x)}{f_1(x)}.$$  

Assume a sequence of observations $\{x^t\}_{t=1}^{+\infty}$. There exists a change-point $\tau$, and two different distributions $\mathcal{P}_1$, $\mathcal{P}_2$ such that for $t < \tau$, $x^t \overset{i.i.d.}{\sim} \mathcal{P}_1$ and for $t \geq \tau$, $x^t \overset{i.i.d.}{\sim} \mathcal{P}_2$. The classical CUSUM procedure is based on the likelihood ratio to detect the change of the data distribution.
Definition 2. (CUSUM Procedure) The CUSUM procedure is a stopping time that stops the first time the likelihood based statistic exceeds a threshold $b$:

$$T_s = \inf \{ t > 0 : \max_{1 \leq i \leq t} \sum_{k=i}^{t} L(x^k) \geq b \}. \tag{2}$$

Under normal circumstances, the larger $b$ is, the larger the detection delay and the smaller the false alarm rate. The stopping time $T_s$ has a recursive implementation as below

$$y^{t+1} = \max\{y^t + L(x^{t+1}), 0\}, \quad T_s = \inf \{ t > 0 : y^t \geq b \}.$$ 

3 Consensus based detection procedure

We represent an $N$-sensor network using a graph $G = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ and $\mathcal{E}$ are the sensor set and edge set, respectively. There exists an edge between sensor $i$ and sensor $j$ if and only if they can communicate with each other. Without loss of generality, we assume that the graph is connected; since if there is more than one connected component, we can apply our algorithm to each of them separately. Assume the topology of the sensor network is known (e.g., by system design).

Denote data observed by the sensor $v$ at time $t$ as $x^t_v$. There exists a change time $\tau$ such that for $t < \tau$, $x^t_v \sim \mathcal{P}_{v1}$ and for $t \geq \tau$, $x^t_v \sim \mathcal{P}_{v2}$. The data observed at different sensors are mutually independent of each other. We also assume the change occurs to all sensors at the same time.

Assume the nodes exchange information via a weighted consensus matrix $W \in \mathbb{R}^{N \times N}$. We assume consensus matrix $W \in \mathbb{R}^{N \times N}$ satisfying the following conditions (by design):

- $W_{ij} > 0$ if sensor $i$ and sensor $j$ are connected and $W_{ij} = 0$ if sensor $i$ and sensor $j$ are not connected.
- $W1 = 1$ and $W^T = W$.
- The second largest eigenvalue modulus of $W$, defined as $\lambda_2(W)$, is smaller than 1.

As long as the graph is connected, the consensus matrix $W$ satisfying the above conditions always exists. We will present a method to design $W$ satisfying the above conditions in Section 5.3.

Our algorithm contains three steps at each sensor, as summarized in Algorithm 1, which we present below:

**Step 1**: Form a local CUSUM statistic at each sensor:

$$y^t_{v+1} = \max\{y^t_v + L_v(x^{t+1}_v), 0\}, \quad v \in \mathcal{V},$$

where $L_v(\cdot)$ is the LLR between $\mathcal{P}_{v2}$ and $\mathcal{P}_{v1}$.

**Step 2**: Sensors exchange information with their neighbors according to the pre-determined topology and weights, as follows

$$z^{t+1}_v = \sum_{u \in \mathcal{N}(v)} W_{vu} (z^t_u + y^{t+1}_u - y^t_u), \quad v \in \mathcal{V},$$

where $\mathcal{N}(v)$ contains sensor $v$ and its neighbor sensors. To update the auxiliary variable $z^t_v$ at sensor $v$ in this process, we only need the information from sensor $v$’s neighbors due to the first property of the consensus matrix. Therefore, the communication process is completely local and the detection is fully distributed.
Algorithm 1: Distributed Change-point Detection.

Parameter: $\{o_v\}_{v \in \mathcal{V}}, \{y_v\}_{v \in \mathcal{V}}, \{z_v\}_{v \in \mathcal{V}}$ and $b$.

Initialization: $o_v = y_v = z_v = 0$ for all $v \in \mathcal{V}$.

For $t = 1, 2 \ldots$
1) Sample $x^t_v, v \in \mathcal{V}$.
2) Compute CUSUM statistic at each sensor:
   \[
o_v = y_v, \quad v \in \mathcal{V}.
   \]
   \[
y_v = \max\{y_v + L_v(x^t_v), 0\}, \quad v \in \mathcal{V}.
   \]
3) Communicate with neighbors:
   \[
z_v = \sum_{u \in \mathcal{N}(v)} W_{vu} (z_u + y_u - o_u), \quad v \in \mathcal{V}.
   \]
4) Detect change if there exists a $z_v \geq b, v \in \mathcal{V}$.

End

Remark 1. Since $W_1 = 1$ and $W^T = W$, it is not difficult to verify that
\[
\sum_{v \in \mathcal{V}} z^t_v = \sum_{v \in \mathcal{V}} y^t_v \text{ holds for all } t.
\]

Step 3: Perform detection by comparing $z^t_v$ with a predetermined threshold $b$ at each sensor $v \in \mathcal{V}$. As long as there exists one $v \in \mathcal{V}$ such that $z^t_v \geq b$, a global alarm is fired.

In summary, our detection procedure corresponds to the following stopping time
\[
T_s = \inf \{t > 0 : \max_{v \in \mathcal{V}} z^t_v \geq b\}.
\]

4 Theoretical analysis of ARL and EDD

In this section, we present the main theoretical results. First, we define necessary performance metrics: the average run length (ARL) and the expected detection delay (EDD) [5]:

Definition 3. (ARL and EDD)

\[
\text{EDD} = \mathbb{E}[T_s | \tau = 1],
\]
\[
\text{ARL} = \mathbb{E}[T_s | \tau = \infty].
\]

In the definition above, $\tau = \infty$ means that the change-point never occurs. Intuitively, EDD can be interpreted as the delay time before detecting the change and ARL can be interpreted as the expected duration between two false alarms, which is the reciprocal of false alarm rate.

In addition, we make the following assumptions

Assumption 1. 1. All the sensors share the same distribution $\mathcal{P}_1$ and $\mathcal{P}_2$, and thus the same log-likelihood radio function $L(\cdot)$.

2. For $x \sim \mathcal{P}_1$ and $x \sim \mathcal{P}_2$, random variable $L(x)$ obeys sub-Gaussian distribution.

The sub-Gaussian distribution includes many important cases such as Student’s t-distribution, Gaussian distribution and Beta Distribution, and it is a commonly seen property for the LLR. For instance, [13] studied distributed sequential hypothesis test such that all the sensors share the same $\mathcal{P}_1 = \mathcal{N}(0, I), \mathcal{P}_2 = \mathcal{N}(\mathbf{u}, I)$ and $L(x) = \mathbf{u}^T x - \|\mathbf{u}\|^2 / 2$. Hence, when $x \sim \mathcal{P}_1$, $L(x) \sim \mathcal{N}(-\|\mathbf{u}\|^2 / 2, \|\mathbf{u}\|^2)$ and when $x \sim \mathcal{P}_2$, $L(x) \sim \mathcal{N}(\|\mathbf{u}\|^2 / 2, \|\mathbf{u}\|^2)$. 

4
Remark 2. The above assumption is made purely to simplify theoretical analysis. The detection procedure can still be implemented without this assumption.

4.1 Theoretical Analysis of ARL

First we present an asymptotic lower bound for the ARL defined in (4). Assume that the mean and variance of $L(x)$ when $x \sim \mathcal{P}_1$ are given by $\mu_1$ and $\sigma_1$, respectively. Note that since $-\mu_1$ corresponds to the Kullback-Leibler (KL)-divergence from $\mathcal{P}_2$ to $\mathcal{P}_1$, $\mu_1$ is always nonpositive (can be shown using Jensen’s inequality).

The following theorem establishes a lower bound for ARL and shows that it increases at least exponentially as $N$ increases.

**Theorem 1 (Lower-bound for ARL).** When the detection threshold $b$ tends to infinity, we have

$$\text{ARL} \geq \exp \left\{ -\frac{\mu_1 b}{\sigma_1^2} \left[ 2N - 2 \left( \frac{N}{N + 1} \right)^2 \right] + \frac{\sqrt{-\mu_1 \mu_1 \lambda_2(W)}}{\sigma_1^2 \left[ 1 - \lambda_2(W) \right]} \left( 2N - 2 \left( \frac{N}{N + 1} \right)^2 \right) + o(1) \right\} \sqrt{b}.$$ 

The main ingredient of the proof is concentration inequality, which is used to find $p \in \mathbb{R}$ as large as possible such that

$$\lim_{b \to +\infty} \mathbb{P}(T_s \leq p) = 0.$$ 

Then we can obtain $\text{ARL} \geq p$. The detailed proof is delegated to appendix.

Remark 3. Intuitively, as the number of sensors increases, the false alarm rate increases because we take the one-shot scheme similar to [9]. The situation in our algorithm, however, is different due to local communications. Moreover, Theorem 1 demonstrates that the lower bound of ARL is related with the second largest eigenvalue modulus of the consensus matrix. The smaller the second largest eigenvalue modulus, the larger the lower bound of ARL. This is similar to observation made for the original average consensus paper [10] and mixing Markov chain problem [15], where smaller second largest eigenvalue modulus can lead to faster convergence speed. Previous works [10, 15, 16] have studied how to design the consensus matrix to get the smallest second largest eigenvalue modulus. Their results can be adapted to design the consensus matrix in this paper, as we shown in numerical examples in Section 5.3.

4.2 Theoretical Analysis of EDD

In this section, we analyze the EDD, i.e. the expected detection delay, assuming that the distribution of samples changes at time 0 (which provides an upper bound to EDD [17]). Denote the mean and the variance of $L(x)$ for $x \sim \mathcal{P}_2$ as $\mu_2$ and $\sigma_2$, respectively. Note that $\mu_2$ corresponds to the KL-divergence from $\mathcal{P}_1$ to $\mathcal{P}_2$, so $\mu_2$ is nonnegative.

**Lemma 1 (Upper-bound for EDD).** When the detection threshold $b$ tends to infinity, we have

$$\text{EDD} \leq \frac{b}{\mu_2} \left( 1 + o(1) \right).$$
The main ingredient of the proof is again concentration inequality, which shows

\[
\lim_{b \to +\infty} \sum_{t = \left\lfloor \frac{\delta (1 + \varepsilon)}{\mu_2} \right\rfloor + 1}^{+\infty} \mathbb{P} (T_s = t) t = 0, \quad \forall \varepsilon > 0.
\]

The proof can be found in the appendix.

By combining Theorem 1 with Lemma 1, we explicitly characterizes the relationship to ARL.

**Theorem 2.** When the detection threshold \( b \) tends to infinity, under the constraint \( ARL \geq \gamma \), we have

\[
EDD \leq \frac{\log \gamma (1 + o(1))}{N \mu_2} \times \frac{\sigma_1^2}{-2 \mu_1 (1 - N^2 / (N + 1)^2)}.\]

Theorem 2 shows that EDD can be reduced by roughly a factor of \( N \) by using local communications compared to the scheme with no communication \([9]\). Moreover, by comparing Theorem 2 with the results in \([6, 8]\), one can find that the theoretical performance of our decentralized algorithm is nearly as optimal as the existing centralized algorithms.

## 5 Numerical Experiments

In this section, we present several numerical experiments to verify the performance of our algorithm. Assume \( \mathcal{P}_1 \) is \( \mathcal{N}(0, 1) \) and \( \mathcal{P}_2 \) is \( \mathcal{N}(1, 1) \). Thus, \( \mu_1 = -0.5, \mu_2 = 0.5 \) and \( \sigma_1 = \sigma_2 = 1 \).

### 5.1 Comparison with one-shot scheme

First consider a simple four-node network. We compare a line topology with a network of isolated nodes (i.e., no communication) to show the effect of communication. Here the isolated case corresponds to the one-shot detection scheme in \([9]\), where each sensor performs a CUSUM procedure separately and does not communicate with their neighbors; as long as there is one CUSUM statistic exceeding the threshold, a change is detected. In our case, sensors cooperate with each other. The consensus matrix used for communication is as below (here the diagonal entries are non-zero meaning that a sensor incorporates its own data when computing the local statistic):

\[
W = \begin{pmatrix}
\frac{1}{3} & \frac{2}{3} & 0 & 0 \\
\frac{2}{3} & \frac{1}{6} & \frac{1}{6} & 0 \\
0 & \frac{1}{6} & \frac{1}{2} & \frac{1}{3} \\
0 & 0 & \frac{1}{3} & \frac{2}{3}
\end{pmatrix}.
\]

The detection performance shown in Figure 2 demonstrates that the change can be detected approximately four times faster under the same ARL with our scheme, which is consistent with the theoretical prediction in Theorem 2.

### 5.2 The impact of network topology

In the second example, we compare the performance of our algorithm on two different topologies, a line network consisting of four nodes and K4, which is a complete graph consisting of four mutually connected nodes \([18]\). The line is the worst topology for communication
Figure 2: Communication versus No communication [9] in fully distributed network.

Figure 3: Two different topology consisting of four nodes.
because the information at one end needs to travel all edges before it reaches another end while $K4$ is the best because the information at one sensor can be directly passed to other sensors via only one edge. We use the following consensus matrices for the two networks, respectively

$$
\begin{align*}
\text{Line:} & \begin{pmatrix}
1/3 & 2/3 & 0 & 0 \\
2/3 & 1/6 & 1/6 & 0 \\
0 & 1/6 & 1/2 & 1/3 \\
0 & 0 & 1/3 & 2/3 \\
\end{pmatrix} \\
K4: & \begin{pmatrix}
1/4 & 1/4 & 1/4 & 1/4 \\
1/4 & 1/4 & 1/4 & 1/4 \\
1/4 & 1/4 & 1/4 & 1/4 \\
1/4 & 1/4 & 1/4 & 1/4 \\
\end{pmatrix}
\end{align*}
$$

Their second largest eigenvalue modulus are 0.86 and 0, respectively. According to our previous analysis, the graph with smaller second largest eigenvalue modulus should have smaller EDD under the same ARL, which is consistent with the experiment results shown in Figure 3(c).

5.3 Optimized weights versus maximum degree chain

In the last example, we compare two networks with the same topology: the first network is the maximum degree chain network [15], which uses unity weights on all edges, and the second network uses optimized weights using the algorithm in [15]. The optimized weights are obtained for a fixed topology by minimizing the second largest eigenvalue modulus to achieve faster convergence. The optimal consensus matrix obtained using the algorithm in [15], which is solved numerically from an optimization problem and the resulted weight matrix given as below

$$
W = \begin{pmatrix}
0 & 0 & 0.34 & 0 & 0.34 & 0.08 & 0 & 0.09 & 0.15 & 0 \\
0 & 0 & 0.13 & 0 & 0.13 & 0.13 & 0 & 0.26 & 0 & 0.35 \\
0.34 & 0.13 & 0.27 & 0 & 0.13 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0.13 & 0.13 & 0 & 0 & 0 & 0 & 0 \\
0.34 & 0.13 & 0.13 & 0 & 0 & 0.16 & 0.11 & 0 & 0 & 0 \\
0.08 & 0.13 & 0 & 0.13 & 0 & 0.26 & 0 & 0.14 & 0.26 & 0 \\
0 & 0 & 0 & 0.26 & 0.16 & 0.13 & 0.1 & 0.34 & 0 \\
0.09 & 0.26 & 0 & 0 & 0.11 & 0.14 & 0.1 & 0.04 & 0.25 & 0 \\
0.15 & 0 & 0 & 0 & 0.26 & 0.34 & 0.25 & 0 & 0 & 0 \\
0 & 0.35 & 0.35 & 0 & 0 & 0 & 0 & 0 & 0 & 0.3 \\
\end{pmatrix}.
$$

The maximum degree chain network has the following weights (with the same topology as the network above):

$$
W_{ij} = \begin{cases} 
\frac{1}{\max_{i \in V} d_i}, & i \neq j \text{ and } (i, j) \in \mathcal{E} \\
1 - \sum_{j \in \mathcal{N}(i)} W_{ij}, & i = j \\
0, & \text{otherwise},
\end{cases}
$$

Figure 4: Optimized consensus matrix [15] versus Maximum degree chain matrix
where $d_i$ is the number of neighbors of sensor $i$. Their second largest eigenvalue modulus are 0.5722 (optimized weights) and 0.7332 (maximum degree network), respectively. Figure 4 shows that the optimized consensus matrix achieves certain performance gain by optimizing weights for the same network topology, which is consistent with Theorem 1. This example shows that when fixing the network topology (which corresponds to fixing the support of $W$, i.e., the location of the non-zeros), there are still certain benefits of optimizing the weights on the consensus schemes to achieve better performance for the change-point detection procedures.

6 Conclusion

In this paper, we present a new distributed change-point detection algorithm based on average consensus, where sensors can exchange CUSUM statistic with their neighbors and perform local detection, and thus no fusion center is needed. We show by allowing sensors to communicate and share information with their neighbors, the expected detection delay can be reduced by a factor of $N$ for a network consisting of $N$ sensors, while the ARL is kept constant.

References

[1] A. Lakhina, M. Crovella, and C. Diot, “Diagnosing network-wide traffic anomalies,” in ACM SIGCOMM Computer Communication Review, vol. 34, no. 4. ACM, 2004, pp. 219–230.

[2] A. G. Tartakovsky and V. V. Veeravalli, “An efficient sequential procedure for detecting changes in multichannel and distributed systems,” in Information Fusion, 2002. Proceedings of the Fifth International Conference on, vol. 1. IEEE, 2002, pp. 41–48.

[3] J. Chen, S.-H. Kim, and Y. Xie, “$S^3T$: An efficient score-statistic for spatio-temporal surveillance,” arXiv preprint arXiv:1706.05331, 2017.

[4] E. S. Page, “Continuous inspection schemes,” Biometrika, vol. 41, no. 1/2, pp. 100–115, 1954.

[5] G. Lorden, “Procedures for reacting to a change in distribution,” The Annals of Mathematical Statistics, pp. 1897–1908, 1971.

[6] A. G. Tartakovsky and V. V. Veeravalli, “Quickest change detection in distributed sensor systems,” in Proceedings of the 6th International Conference on Information Fusion, 2003, pp. 756–763.

[7] ———, “Asymptotically optimal quickest change detection in distributed sensor systems,” Sequential Analysis, vol. 27, no. 4, pp. 441–475, 2008.

[8] Y. Mei, “Efficient scalable schemes for monitoring a large number of data streams,” Biometrika, vol. 97, no. 2, pp. 419–433, 2010.

[9] O. Hadjiliadis, H. Zhang, and H. V. Poor, “One shot schemes for decentralized quickest change detection,” IEEE Transactions on Information Theory, vol. 55, no. 7, pp. 3346–3359, 2009.
[10] L. Xiao and S. Boyd, “Fast linear iterations for distributed averaging,” *Systems & Control Letters*, vol. 53, no. 1, pp. 65–78, 2004.

[11] L. Xiao, S. Boyd, and S. Lall, “A scheme for robust distributed sensor fusion based on average consensus,” in *Proceedings of the 4th international symposium on Information processing in sensor networks*. IEEE Press, 2005, p. 9.

[12] S. Li and X. Wang, “Order-2 asymptotic optimality of the fully distributed sequential hypothesis test,” *arXiv preprint arXiv:1606.04203*, 2016.

[13] A. K. Sahu and S. Kar, “Distributed sequential detection for Gaussian shift-in-mean hypothesis testing,” *IEEE Transactions on Signal Processing*, vol. 64, no. 1, pp. 89–103, 2016.

[14] K. Liu and Y. Mei, “Improved performance properties of the CISPRT algorithm for distributed sequential detection,” *Submitted*, 2017.

[15] S. Boyd, P. Diaconis, and L. Xiao, “Fastest mixing markov chain on a graph,” *SIAM review*, vol. 46, no. 4, pp. 667–689, 2004.

[16] L. Xiao, S. Boyd, and S.-J. Kim, “Distributed average consensus with least-mean-square deviation,” *Journal of Parallel and Distributed Computing*, vol. 67, no. 1, pp. 33–46, 2007.

[17] Y. Xie and D. Siegmund, “Sequential multi-sensor change-point detection,” *Annals of Statistics*, vol. 41, no. 2, pp. 670–692, 2013.

[18] J. A. Bondy, U. S. R. Murty *et al.*, *Graph theory with applications*. Citeseer, 1976, vol. 290.
7 Appendix

For the simplicity of proof, we first number the sensors from 1 to \(N\). Then, we use vector \(L^t\) to represent \((L(x_1^t), \ldots, L(x_N^t))^T\), vector \(y^t\) to represent \((y_1^t, \ldots, y_N^t)^T\) and vector \(z^t\) to represent \((z_1^t, \ldots, z_N^t)^T\). Now, our algorithm can be rewritten in the following form

\[
\begin{align*}
    y^{t+1} &= (y^t + L^{t+1})^+,
    z^{t+1} &= W(z^t + y^{t+1} - y^t),
    T_s &= \inf \{ t > 0 : \|z^t\|_\infty \geq b \}.
\end{align*}
\]  

Firstly, we prove some useful lemmas before reaching the main results.

**Lemma 2. (Hoeffding Inequality)** Let \(X_i\) be independent, mean-zero, \(\sigma_i^2\)-sub-Gaussian random variables. Then for \(K > 0\),

\[
P\left( \sum_{i=1}^{n} X_i \geq K \right) \leq \exp \left( - \frac{K^2}{2 \sum_{i=1}^{n} \sigma_i^2} \right).
\]

**Lemma 3.** Consider a sequence of random variables \(X_k\) \(i.i.d.\) \(\sim P\), for \(k = 1, 2, \ldots, t\). \(P\) is a sub-Gaussian distribution and its mean and variance are defined as \(\mu_1 < 0\) and \(\sigma_1\), respectively. Given \(K > 0\) large enough, we have

\[
\sum_{k=1}^{t} \mathbb{P} \left( \sum_{q=1}^{k} X_q > K \right) < \frac{2K}{\mu_1} \exp \left( \frac{2K\mu_1}{\sigma_1^2} \right).
\]

*Proof.*

**Case 1.** For \(0 < t \leq \left\lfloor \frac{-2K}{\mu_1} \right\rfloor\), by Hoeffding Inequality, we have

\[
\sum_{k=1}^{t} \mathbb{P} \left( \sum_{q=1}^{k} X_q > K \right) < \sum_{k=1}^{t} \exp \left( - \frac{1}{2} \left( \frac{K - k\mu_1}{\sqrt{k}\sigma_1} \right)^2 \right). \tag{8}
\]

Using \(\frac{K - k\mu_1}{\sqrt{k}} \geq 2\sqrt{-K\mu_1}\) and \(t \leq \frac{-2K}{\mu_1}\), we get

\[
\sum_{k=1}^{t} \mathbb{P} \left( \sum_{q=1}^{k} X_q > K \right) < \frac{2K}{\mu_1} \exp \left( \frac{2K\mu_1}{\sigma_1^2} \right). \tag{9}
\]

**Case 2.** For \(\left\lfloor \frac{-2K}{\mu_1} \right\rfloor + 1 \leq t\), by (9), we have

\[
\sum_{k=1}^{t} \mathbb{P} \left( \sum_{q=1}^{k} X_q > K \right) < \frac{2K}{\mu_1} \exp \left( \frac{2K\mu_1}{\sigma_1^2} \right) + \sum_{k=\left\lfloor \frac{-2K}{\mu_1} \right\rfloor + 1}^{t} \mathbb{P} \left( \sum_{q=1}^{k} X_q > K \right). \tag{10}
\]

Utilizing Hoeffding Inequality and \(k \geq \left\lfloor \frac{-2K}{\mu_1} \right\rfloor + 1\), we get

\[
\mathbb{P} \left( \sum_{q=1}^{k} X_q > K \right) < \exp \left( - \frac{1}{2} \left( \frac{K - k\mu_1}{\sqrt{k}\sigma_1} \right)^2 \right) \leq \exp \left( \frac{9K\mu_1}{4\sigma_1^2} \right). \tag{11}
\]
Besides, for $k \geq \lceil -\frac{2K}{\mu_1} \rceil + 1$, we have
\[
\exp\left( -\frac{1}{2} \left( \frac{K-(k+1)\mu_1}{\sqrt{k+1}\sigma_1} \right)^2 \right) = \exp\left( -\frac{\mu_1^2}{2\sigma_1^2} + \frac{K^2}{2k(k+1)\sigma_1^2} \right) < \exp\left( -\frac{3\mu_1^2}{8\sigma_1^2} \right). \tag{12}
\]
Then, utilizing Hoeffding Inequality, (11) and (12), we get
\[
\sum_{k=\lceil -\frac{2K}{\mu_1} \rceil + 1}^{t} \mathbb{P}\left( \sum_{q=1}^{k} X_q > K \right) < \sum_{k=\lceil -\frac{2K}{\mu_1} \rceil + 1}^{t} \exp\left( -\frac{1}{2} \left( \frac{K-k\mu_1}{\sqrt{k}\sigma_1} \right)^2 \right)
\]
\[
< \sum_{k=\lceil -\frac{2K}{\mu_1} \rceil + 1}^{t} \exp\left( \frac{9K\mu_1}{4\sigma_1^2} \right) \times \exp\left( -\frac{3\mu_1^2}{8\sigma_1^2} \left( k - \lfloor -\frac{2K}{\mu_1} \rfloor - 1 \right) \right)
\]
\[
< \exp\left( \frac{9K\mu_1}{4\sigma_1^2} \right) \frac{1}{1 - \exp\left( -\frac{3\mu_1^2}{8\sigma_1^2} \right)}. \tag{13}
\]
From (13), we know that the second term on the RHS of (10) is a small quantity compared with the first term provided $K$ large enough, so we just neglect it and get
\[
\sum_{k=1}^{t} \mathbb{P}\left( \sum_{q=1}^{k} X_q > K \right) < \frac{-2K}{\mu_1} \exp\left( \frac{2K\mu_1}{\sigma_1^2} \right). \tag{14}
\]
have

\[ |z_j^* - \frac{\sum_{i=1}^{N} y_i^t}{N}| = |z_j^* - \frac{\sum_{i=1}^{N} z_i^t}{N}| \]
\[ \leq \|z^t - \frac{\sum_{i=1}^{N} z_i^t}{N}1\|_2 \]
\[ = \|\sum_{k=1}^{t} (W^{t-k+1} - \frac{1}{N}11^T)(y^k - y^{k-1})\|_2 \]
\[ \leq \sum_{k=1}^{t} \lambda_2^{t-k+1}\|y^k - y^{k-1}\|_2 \]
\[ \leq \sum_{k=1}^{t} \lambda_2^{t-k+1}\|L^k\|_2 \]
\[ \leq \sum_{k=1}^{t} \lambda_2^{t-k+1}\sqrt{N}\varepsilon b \]
\[ \leq \frac{\sqrt{N}\varepsilon \lambda_2 b}{1 - \lambda_2}, \]

where \(\lambda_2\) is the second largest eigenvalue modulus of \(W\). If \(\{T_s = t\}\) happens, then \(z_j^* > b\) holds for some \(j\), which, together with the inequality above, leads to

\[ \sum_{j=1}^{N} y_j^t \geq (1 - \frac{\sqrt{N}\varepsilon \lambda_2}{1 - \lambda_2})b. \]

\[ \square \]

**Lemma 5.** Suppose we have a sequence of independent random variables \(Y_1, \cdots, Y_N\). Take any integer \(M > N\) and let \(C(M, N)\) to be the set below,

\[ \{(i_1, \cdots, i_N) : i_j \in \mathbb{N} \text{ and } M - N \leq \sum_{j=1}^{N} i_j \leq M\}. \]

Then we have

\[ \mathbb{P}\left( \sum_{j=1}^{N} Y_j > K \right) \leq \sum_{C(M, N)} \prod_{j=1}^{N} \mathbb{P}\left( Y_j > \frac{i_j K}{M} \right). \]

**Proof.** \(\forall(y_1, \cdots, y_N) \in \{(Y_1, \cdots, Y_N) : \sum_{j=1}^{N} Y_j > K\}\), take

\[ i_j = \left\lfloor \frac{y_j M}{\sum_{j=1}^{N} y_j} \right\rfloor \text{ for } j = 1, \ldots, N. \]

We can easily verify that \(M - N \leq \sum_{j=1}^{N} i_j \leq M\) and \(y_j > i_j K/M\). Therefore, we have

\[ \{(Y_1, \cdots, Y_N) : \sum_{j=1}^{N} Y_j > K\} \subset \bigcup_{C(M, N)} \left\{(Y_1, \cdots, Y_N) : Y_j > \frac{i_j K}{M} \text{ for } j = 1, \ldots, N\right\}. \]
Because $Y_j$ is independent with each other, together with the property of probability measure, we get
\[
\mathbb{P}\left(\sum_{j=1}^{N} Y_j > K\right) \leq \sum_{C(M,N)} \prod_{j=1}^{N} \mathbb{P}\left(Y_j > \frac{i_j K}{M}\right).
\]

7.1 Proof of Theorem 1

First, we calculate the probability that our algorithm stops within time $p$. The value of $p$ is to be specified later.

\[
\sum_{t=1}^{p} \mathbb{P}(T_s = t) = \sum_{t=1}^{p} \left(\mathbb{P}\left(\{T_s = t\} \land B(\varepsilon, p)\right) + \mathbb{P}\left(\{T_s = t\} \land \overline{B}(\varepsilon, p)\right)\right) \\
\leq \sum_{t=1}^{p} \mathbb{P}\left(\{T_s = t\} \land B(\varepsilon, p)\right) + \mathbb{P}\left(\overline{B}(\varepsilon, p)\right) \\
\leq \sum_{t=1}^{p} \mathbb{P}\left(\{T_s = t\} \land B(\varepsilon, p)\right) + 2Np \times \exp\left(-\frac{(\varepsilon b + \mu_1)^2}{2\sigma_1^2}\right),
\]

where the last inequality is from Hoeffding Inequality and Assumption 1. The value of $\varepsilon$ is to be specified later.

Denote $\bar{b} = N(1 - \sqrt{\frac{\varepsilon b}{1 - \sqrt{\lambda_2}}})b$, then $\bar{b}$ will also tend to infinity as $b$ tends to infinity provided $\varepsilon$ small enough. By Lemma 4, we have

\[
\sum_{t=1}^{p} \mathbb{P}(T_s = t) \leq \sum_{t=1}^{p} \mathbb{P}\left(\left\{\sum_{j=1}^{N} y_j > \bar{b}\right\} \land B(\varepsilon, p)\right) + 2Np \times \exp\left(-\frac{(\varepsilon b + \mu_1)^2}{2\sigma_1^2}\right) . \quad (15)
\]

By Lemma 5, we have

\[
\mathbb{P}\left(\left\{\sum_{j=1}^{N} y_j^t > \bar{b}\right\} \land B(\varepsilon, p)\right) \leq \sum_{C(M,N)} \prod_{j=1}^{N} \mathbb{P}\left(\left\{y_j^t > \frac{i_j \bar{b}}{M}\right\} \land B(\varepsilon, p)\right) , \quad (16)
\]

where the value of $M$ is to be specified later. If $y_j^t > i_j \bar{b}/M$, then there must exist $1 \leq k \leq t$ such that $y_j^t = \sum_{q=k}^{t} L(x_j^q) \geq i_j \bar{b}/M$. So, we have

\[
\mathbb{P}\left(\left\{y_j^t > \frac{i_j \bar{b}}{M}\right\} \land B(\varepsilon, p)\right) \leq \sum_{k=1}^{t} \mathbb{P}\left(\left\{\sum_{q=k}^{t} L(x_j^q) > \frac{i_j \bar{b}}{M}\right\} \land B(\varepsilon, p)\right). \quad (17)
\]

The influence of $B(\varepsilon, p)$ in (17) can be interpreted as truncating the original distribution of $L(\cdot)$. It’s obvious that the new distribution is still sub-Gaussian. Besides, the mean and variance almost keep unchanged provided $\varepsilon b$ large enough.
If \( i_j = 0 \), we just set the upper bound of the probability in (17) to be 1. If \( i_j \neq 0 \), by Lemma 3, we have

\[
\sum_{k=1}^{t} \mathbb{P} \left( \left\{ \sum_{q=k}^{t} L(x_j^q) > \frac{i_j \bar{b}}{M} \right\} \land B(\varepsilon, p) \right) < -\frac{2i_j \bar{b} \mu_1}{M \mu_1} \exp \left( \frac{2i_j \bar{b} \mu_1}{M \sigma_1^2} \right). \tag{18}
\]

Plugging (18) into (16), we get

\[
\sum_{C(M,N)} \prod_{j=1}^{N} \mathbb{P} \left( \left\{ y_j > \frac{i_j \bar{b}}{M} \right\} \land B(\varepsilon, p) \right) < \sum_{C(M,N), i_j \neq 0} \prod_{j \neq \not 0} \frac{2i_j \bar{b}}{-M \mu_1} \exp \left( \frac{2i_j \bar{b} \mu_1}{M \sigma_1^2} \right) \\
\leq \sum_{C(M,N)} \left( \frac{2\bar{b}}{-\mu_1} \right)^N \exp \left( \frac{2b \mu_1}{\sigma_1^2} (1 - \frac{N}{M}) \right) \\
= |C(M, N)| \left( \frac{2\bar{b}}{-\mu_1} \right)^N \exp \left( \frac{2b \mu_1}{\sigma_1^2} (1 - \frac{N}{M}) \right). \tag{19}
\]

Plugging (19) and (16) into (15), we get

\[
\sum_{t=1}^{p} \mathbb{P}(T_s = t) < \sum_{t=1}^{p} |C(M, N)| \left( \frac{2\bar{b}}{-\mu_1} \right)^N \exp \left( \frac{2b \mu_1}{\sigma_1^2} (1 - \frac{N}{M}) \right) + 2Np \times \exp \left( -\frac{(\varepsilon b + \mu_1)^2}{2\sigma_1^2} \right) \\
= p |C(M, N)| \left( \frac{2\bar{b}}{-\mu_1} \right)^N \exp \left( \frac{2b \mu_1}{\sigma_1^2} (1 - \frac{N}{M}) \right) + 2Np \times \exp \left( -\frac{(\varepsilon b + \mu_1)^2}{2\sigma_1^2} \right). \tag{20}
\]

Next, we will show that as \( b \) tends to infinity, the second term on the RHS of (20) is a small quantity in comparison with the first term if we choose the value of \( M \) and \( \varepsilon \) properly. Note that \( 2Np \) is a small quantity in comparison with \( p |C(M, N)| \left( \frac{2\bar{b}}{-\mu_1} \right)^N \), so we only require

\[
\frac{2b \mu_1}{\sigma_1^2} (1 - \frac{N}{M}) \geq -\frac{(\varepsilon b + \mu_1)^2}{2\sigma_1^2}. \tag{21}
\]

Choosing \( M = (N + 1)^2 \) and recalling that \( \bar{b} = N(1 - \frac{\sqrt{N}\varepsilon \lambda_2}{1 - \lambda_2})b \), (21) can be rewritten as

\[
\frac{(\varepsilon b + \mu_1)^2}{2\sigma_1^2} \geq -\frac{2(N^3 + N^2 + N) \mu_1 b}{(N + 1)^2 \sigma_1^2} \left( 1 - \frac{\sqrt{N}\varepsilon \lambda_2}{1 - \lambda_2} \right). \tag{22}
\]

To ensure that (22) holds as \( b \) tends to infinity, \( \varepsilon = 2\sqrt{-N\mu_1}/b \) is sufficient. Plugging the value of \( M \) and \( \varepsilon \) into (20) and neglecting the second term, we get

\[
\mathbb{P}(T_s \leq p) \leq |C((N + 1)^2, N)| \left( \frac{2\bar{b}}{-\mu_1} \right)^N \\
\times \exp \left( \frac{2(N^3 + N^2 + N) \mu_1 b}{(N + 1)^2 \sigma_1^2} \right) - \frac{4(N^3 + N^2 + N) \sqrt{-\mu_1 \mu_1 \lambda_2 \sqrt{5}}}{(N + 1)^2(1 - \lambda_2) \sigma_1^2} + \ln(p)).
\]
So \( \forall l > -\frac{4(N^3 + N^2 + N)\sqrt{-\mu_1 \lambda_2}}{(N+1)^2(1-\lambda_2)\sigma_1^2} \), if we take

\[
p = \exp \left( -\frac{2(N^3 + N^2 + N)\mu_1 b}{(N+1)^2\sigma_1^2} - l\sqrt{b} \right).
\]

Then we have

\[
\lim_{b \to +\infty} \mathbb{P} (T_s \leq p) \leq \lim_{b \to +\infty} |C((N+1)^2, N)| \left( \frac{-2\sqrt{b}}{-\mu_1} \right)^N \exp \left( -l + \frac{4(N^3 + N^2 + N)\sqrt{-\mu_1 \lambda_2}}{(N+1)^2(1-\lambda_2)\sigma_1^2} \right) \sqrt{b} = 0,
\]

which together with the definition of ARL leads to

\[
\text{ARL} \geq p = \exp \left( -\frac{2(N^3 + N^2 + N)\mu_1 b}{(N+1)^2\sigma_1^2} - l\sqrt{b} \right), \quad \forall l > -\frac{4(N^3 + N^2 + N)\sqrt{-\mu_1 \lambda_2}}{(N+1)^2(1-\lambda_2)\sigma_1^2},
\]

which is identical to

\[
\text{ARL} \geq \exp \left( -\frac{2(N^3 + N^2 + N)\mu_1 b}{(N+1)^2\sigma_1^2} \right) + \left( o(1) + \frac{4(N^3 + N^2 + N)\sqrt{-\mu_1 \lambda_2}}{(N+1)^2(1-\lambda_2)\sigma_1^2} \right) \sqrt{b}
\]

when \( b \) tends to infinity.

### 7.2 Proof of Lemma 1

First of all, note that \( \mathbb{P} (T_s = +\infty) = 0 \), so given \( \varepsilon > 0 \), we have

### (23)

\[
\text{EDD} \leq \frac{b(1+\varepsilon)}{\mu_2} + \sum_{t=\left(\frac{b(1+\varepsilon)}{n_2}\right)+1}^{+\infty} \mathbb{P} (T_s = t) t.
\]

If \( T_s = t \), then we have that \( z_j^{t-1} < b \) holds for all \( j \). Since \( \sum_{j=1}^{N} z_j^{t-1} = \sum_{j=1}^{N} y_j^{t-1} \), there must exist some \( y_j^{t-1} < b \). Therefore, we have

### (24)

\[
\sum_{t=\left(\frac{b(1+\varepsilon)}{n_2}\right)+1}^{+\infty} \mathbb{P} (T_s = t) t \leq \sum_{t=\left(\frac{b(1+\varepsilon)}{n_2}\right)+1}^{+\infty} \sum_{j=1}^{N} \mathbb{P} (y_j^{t-1} < b) t.
\]

Note that \( y_j^{t-1} \geq \sum_{q=1}^{t-1} L(x_q^j) \), together with Hoeffding Inequality, we get

\[
\mathbb{P} (y_j^{t-1} < b) t \leq \mathbb{P} \left( \sum_{q=1}^{t-1} L(x_q^j) < b \right) = \exp \left( -\frac{1}{2} \left( \frac{b - (t-1)\mu_2}{\sqrt{t-1}\sigma_2} \right)^2 \right) t.
\]

Note that \( y_j^{t-1} \geq \sum_{q=1}^{t-1} L(x_q^j) \), together with Hoeffding Inequality, we get

\[
\mathbb{P} (y_j^{t-1} < b) t \leq \mathbb{P} \left( \sum_{q=1}^{t-1} L(x_q^j) < b \right) = \exp \left( -\frac{1}{2} \left( \frac{b - (t-1)\mu_2}{\sqrt{t-1}\sigma_2} \right)^2 \right) t.
\]
When $b$ is large enough, for any $t > \frac{b(1+\varepsilon)}{\mu_2}$, utilizing the similar technique in (12), we get

$$\exp\left(-\frac{1}{2} \left( \frac{b - \mu_2}{\sqrt{t} \sigma_2} \right)^2 \right) \times \frac{t + 1}{t} \leq \exp\left(-\frac{b}{2} \left( 1 - \frac{1}{(1+\varepsilon)^2} \right) \right).$$

(26)

Plugging (25) and (26) into (24), utilizing the similar technique in (13), we get

$$\sum_{t=\left\lfloor \frac{b(1+\varepsilon)}{\mu_2} \right\rfloor + 1}^{+\infty} \mathbb{P} (T_s = t) t \leq \sum_{t=\left\lfloor \frac{b(1+\varepsilon)}{\mu_2} \right\rfloor + 1}^{+\infty} N \times \exp\left(-\frac{1}{2} \left( \frac{b - (t - 1)\mu_2}{\sqrt{t - 1} \sigma_2} \right)^2 \right) t$$

$$\leq N \left( \frac{b(1+\varepsilon)}{\mu_2} + 1 \right) \exp\left(-\frac{\varepsilon^2 b}{2(1+\varepsilon)\sigma_2^2} \right) \frac{1}{1 - \exp\left(-\frac{b}{2} \left( 1 - \frac{1}{(1+\varepsilon)^2} \right) \right)}. \quad (27)$$

Note that $\forall \varepsilon > 0$, as $b$ tends to infinity, the RHS of (27) would converge to zero. Therefore, by (23), we get

$$\text{EDD} \leq \frac{b(1+o(1))}{\mu_2}.$$