Decentralized RLS With Data-Adaptive Censoring for Regressions Over Large-Scale Networks

Zifeng Wang, Zheng Yu, Qing Ling, *Senior Member, IEEE*, Dimitris Berberidis, *Student Member, IEEE*, and Georgios B. Giannakis, *Fellow, IEEE*

***Abstract*—The deluge of networked data motivates the development of algorithms for computation- and communication-efficient information processing. In this context, three data-adaptive censoring strategies are introduced to considerably reduce the computation and communication overhead of decentralized recursive least-squares solvers. The first relies on alternating minimization and the stochastic Newton iteration to minimize a network-wide cost, which discards observations with small innovations. In the resultant algorithm, each node performs local data-adaptive censoring to reduce computations while exchanging its local estimate with neighbors so as to consent on a network-wide solution. The communication cost is further reduced by the second strategy, which prevents a node from transmitting its local estimate to neighbors when the innovation it induces to incoming data is minimal. In the third strategy, not only transmitting, but also receiving estimates from neighbors is prohibited when data-adaptive censoring is in effect. For all strategies, a simple criterion is provided for selecting the threshold of innovation to reach a prescribed average data reduction. The novel censoring-based (C)D-RLS algorithms are proved convergent to the optimal argument in the mean-root deviation sense. Numerical experiments validate the effectiveness of the proposed algorithms in reducing computation and communication overhead.**

***Index Terms*—Decentralized estimation, networks, recursive least-squares (RLS), data-adaptive censoring.**

I. Introduction

In OUR big data era, various networks generate massive amounts of streaming data. Examples include wireless sensor networks, where a large number of inexpensive sensors cooperate to monitor, e.g., the environment [21], [22], or data centers, where a group of servers collaboratively handles dynamic user requests [24]. Since a single node has limited computational resources, decentralized information processing is preferable as the network size scales up [6], [8]. In this paper, we focus on a decentralized linear regression setup, and develop computation- and communication-efficient decentralized recursive least-squares (D-RLS) algorithms.

The main tool we adopt to reduce computation and communication costs is data-adaptive censoring, which leverages the redundancy present especially in big data. Upon receiving an observation, nodes determine whether it is informative or not. Less informative observations are discarded, while messages among neighboring nodes are exchanged only when necessary. We propose three censoring-based (C)D-RLS algorithms that can achieve estimation accuracy comparable to D-RLS without censoring, while significantly reducing the computation and communication overhead.

A. Related Works

The merits of RLS algorithms in solving centralized linear regression problems are well recognized [11], [25]. When streaming observations that depend linearly on a set of unknown parameters become available, RLS yields the least-squares parameter estimates online. RLS reduces the computational burden of finding a batch estimate per iteration, and can even allow for tracking time-varying parameters. The computational cost can be further reduced by data-adaptive censoring [4], where less informative data are discarded. On the other hand, decentralized versions of RLS without censoring have been advocated to solve linear regression tasks over networks [15]. In D-RLS, a node updates its estimate that is common to the entire network by fusing its local observations with the local estimates of its neighbors. As time evolves, all local estimates consent on the centralized RLS solution. This paper builds on both [4] and [15] by developing censoring-based decentralized RLS algorithms, thus catering to efficient online linear regression over large-scale networks.

Different from our in-network setting where operation is fully decentralized and nodes are only able to communicate with their neighbors, most of the existing distributed censoring algorithms apply to star topology networks that rely on a fusion center [2], [9], [10], [19], [23]. Their basic idea is that each node transmits data to the fusion center for further processing only when its local likelihood ratio exceeds a threshold [23]; see also [9] where communication constraints are also taken into

Manuscript received December 27, 2016; revised June 11, 2017, October 12, 2017, and December 11, 2017; accepted January 9, 2018. Date of publication January 23, 2018; date of current version February 7, 2018. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Joao Xavier. This work was supported in part by the NSF China under Grant 61573331, in part by the NSF Anhui under Grant 1608085QF130, and in part by the NSF under Grants 1442686, 1500713, and 1711471. This paper was presented in part at the 42nd IEEE International Conference on Acoustics, Speech, and Signal Processing, New Orleans, LA, USA, March 2017. *(Corresponding author: Qing Ling.)*

Z. Wang and Z. Yu are with the Special Class for the Gifted Young, University of Science and Technology of China, Hefei 230026, China (e-mail: wzfeng@mail.ustc.edu.cn; yz123@mail.ustc.edu.cn).

Q. Ling is with the School of Data and Computer Science, Sun Yat-Sen University, Guangzhou 510275, China (e-mail: lingqing556@mail.sysu.edu.cn).

D. Berberidis and G. B. Giannakis are with the Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis MN 55455 USA (e-mail: bermp001@umn.edu; georgios@umn.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSP.2018.2795594

account. Information fusion over fading channels is considered in [10]. Practical issues such as joint dependence of sensor decision rules, randomization of decision strategies as well as partially known distributions are reported in [2], while [19] also explores quantization jointly with censoring.

Other than the star topology studied in the aforementioned works, [20] investigates censoring for a tree structure. If a node's local likelihood ratio exceeds a threshold, its local data is sent to its parent node for fusion. A fully decentralized setting is considered in [3], where each node determines whether to transmit its local estimate to its neighbors by comparing the local estimate with the weighted average of its neighbors. Nevertheless, [3] aims at mitigating only the communication cost, while the present work also considers reduction of the computational cost across the network. Furthermore, the censoring-based decentralized linear regression algorithm in [13] deals with optimal full-complexity estimation when observations are partially known or corrupted. This is different from our context, where censoring is deliberately introduced to reduce computational and communication costs for decentralized linear regression.

B. Our Contributions and Organization

The present paper introduces three data-adaptive online censoring strategies for decentralized linear regression. The resultant CD-RLS algorithms incur low computational and communication costs, and are thus attractive for large-scale network applications requiring decentralized solvers of linear regressions. Unlike most related works that specifically target wireless sensor networks (WSNs), the proposed algorithms may be used in a broader context of decentralized linear regression using multiple computing platforms. Of particular interest are cases where a regression dataset is not available at a single machine, but it is distributed over a network of computing agents that are interested in accurately estimating the regression coefficients in an efficient manner.

In Section II, we formulate the decentralized online linear regression problem (Section II-A), and recast the D-RLS in [15] into a new form (Section II-B) that prompts the development of three censoring strategies (Section II-C). Section III develops the first censoring strategy (Section III-A), analyzes all three censoring strategies (Section III-B), and discusses how to set the censoring thresholds (Section III-C). Numerical experiments in Section IV demonstrate the effectiveness of the novel CD-RLS algorithms.

Notation: Lower (upper) case boldface letters denote column vectors (matrices). $(\cdot)^T$, $||\cdot||$, $||\cdot||_2$ and $E[\cdot]$ stand for transpose, 2-norm, induced matrix 2-norm and expectation, respectively. Symbols $\mathrm{tr}(\mathbf{X})$, $\lambda_{\min}(\mathbf{X})$ and $\lambda_{\max}(\mathbf{X})$ are used for the trace, minimum eigenvalue and maximal eigenvalue of matrix $\mathbf{X}$, respectively. Kronecker product is denoted by $\otimes$ and the uniform distribution over $[a,b]$ by $\mathcal{U}(a,b)$, and the Gaussian probability distribution function (pdf) with mean μ and variance σ^2 by $\mathcal{N}(\mu,\sigma^2)$. The standardized Gaussian pdf is $\phi(t)=(1/\sqrt{2\pi})\exp(-t^2/2)$, and its the associated complementary cumulative distribution function is represented by $Q(z):=\int_z^{+\infty}\phi(t)dt$.

II. Context and Algorithms

This section outlines the online linear regression setup over networks, and takes a fresh look at the D-RLS algorithm. Three strategies are then developed using data-adaptive censoring to reduce the computational and communication costs of D-RLS.

A. Problem Statement

Consider a bidirectionally connected network with J nodes, described by a graph $\mathcal{G}:=\{\mathcal{V},\mathcal{E}\}$, where $\mathcal{V}$ is the set of nodes with cardinality $|\mathcal{V}|=J$, and $\mathcal{E}$ denotes the set of edges. Each node j only communicates with its one-hop neighbors, collected in the set $\mathcal{N}_j\subset\mathcal{V}$. The decentralized network is deployed to estimate a real vector $\mathbf{s}_0\in\mathbb{R}^p$. Per time slot $t=1,2,\ldots$, node j receives a real scalar observation $x_j(t)$ involving the wanted $\mathbf{s}_0$ with a regression row $\mathbf{h}_j^T(t)$, so that $x_j(t)=\mathbf{h}_j^T(t)\mathbf{s}_0+\epsilon_j(t)$, with $\epsilon_j(t)\sim\mathcal{N}(0,\sigma_j^2)$.

Our goal is to devise efficient decentralized online algorithms to solve the following exponentially-weighted least-squares (EWLS) problem

$$\hat{\mathbf{s}}_{ewls}(t):=\arg\min_{\mathbf{s}}\frac{1}{2}\sum_{r=1}^{t}\sum_{j=1}^{J}\lambda^{t-r}[x_j(r)-\mathbf{h}_j^T(r)\mathbf{s}]^2 \quad (1)$$

where $\hat{\mathbf{s}}_{ewls}(t)$ is the EWLS estimate at slot t, and $\lambda\in(0,1]$ is a forgetting factor that de-emphasizes the importance of past measurements, and thus enables tracking of a non-stationary process. When $\lambda=1$, (1) boils down to a standard decentralized online least-squares estimate.

B. D-RLS Revisited

The D-RLS algorithm of [15] solves (1) as follows. Per time slot t, node j receives $x_j(t)$ and $\mathbf{h}_j^T(t)$ and uses them to update the per-node inverse $p\times p$ covariance matrix as

$$\mathbf{\Phi}_j^{-1}(t)=\lambda^{-1}\mathbf{\Phi}_j^{-1}(t-1) - \frac{\lambda^{-1}\mathbf{\Phi}_j^{-1}(t-1)\mathbf{h}_j(t)\mathbf{h}_j^T(t)\mathbf{\Phi}_j^{-1}(t-1)}{\lambda+\mathbf{h}_j^T(t)\mathbf{\Phi}_j^{-1}(t-1)\mathbf{h}_j(t)} \quad (2)$$

along with the per-node $p\times 1$ cross-covariance vector as

$$\boldsymbol{\psi}_j(t)=\lambda\boldsymbol{\psi}_j(t-1)+\mathbf{h}_j(t)x_j(t). \quad (3)$$

Using $\mathbf{\Phi}_j^{-1}(t)$ and $\boldsymbol{\psi}_j(t)$, node j then updates its local parameter estimate using

$$\mathbf{s}_j(t)=\mathbf{\Phi}_j^{-1}(t)\left[\boldsymbol{\psi}_j(t)-\frac{1}{2}\sum_{j'\in\mathcal{N}_j}\left(\mathbf{v}_j^{j'}(t-1)-\mathbf{v}_{j'}^{j}(t-1)\right)\right] \quad (4)$$

where $\mathbf{v}_j^{j'}(t-1)$ denotes the Lagrange multiplier of node j corresponding to its neighbor j' at slot $t-1$, that captures the accumulated differences of neighboring estimates, recursively obtained as ($\rho>0$ is a step-size)

$$\mathbf{v}_j^{j'}(t-1)=\mathbf{v}_j^{j'}(t-2)+\rho\big[\mathbf{s}_j(t-1)-\mathbf{s}_{j'}(t-1)\big]. \quad (5)$$

Next, we develop an equivalent novel form of D-RLS recursions (2)–(5) that is convenient for our incorporation of data-adaptive censoring. Detailed derivation of the equivalence can be found in Appendix B. The inverse covariance matrix is updated as in (2). However, the update of $\mathbf{s}_j(t)$ in (4) is replaced by

$$\mathbf{s}_j(t) = \mathbf{s}_j(t-1) + \mathbf{\Phi}_j^{-1}(t)\mathbf{h}_j(t)\big[x_j(t) - \mathbf{h}_j^T(t)\mathbf{s}_j(t-1)\big] - \rho\mathbf{\Phi}_j^{-1}(t)\boldsymbol{\delta}_j(t-1) \tag{6}$$

where $\boldsymbol{\delta}_j(t)$ stands for a Lagrange multiplier conveying network-wide information that is updated as

$$\boldsymbol{\delta}_j(t) = \boldsymbol{\delta}_j(t-1) + \sum_{j'\in\mathcal{N}_j}[\mathbf{s}_j(t) - \mathbf{s}_{j'}(t)] - \lambda\sum_{j'\in\mathcal{N}_j}[\mathbf{s}_j(t-1) - \mathbf{s}_{j'}(t-1)]. \tag{7}$$

Observe that $\boldsymbol{\delta}_j(t)$ stores the weighted sum of differences between the local estimate of node j, and all estimates of its neighbors. Interestingly, if the network is disconnected and the nodes are isolated, then $\boldsymbol{\delta}_j(t) = \mathbf{0}$ so long as $\boldsymbol{\delta}_j(0) = \mathbf{0}$, and the update of $\mathbf{s}_j(t)$ in (6) basically boils down to the centralized RLS one [11], [25]. That is, the current estimate is modified from its previous value using the prediction error $x_j(t) - \mathbf{h}_j^T(t)\mathbf{s}_j(t-1)$, which is known as the incoming data *innovation*. If on the other hand the network is connected, nodes can leverage estimates of their neighbors (captured by $\boldsymbol{\delta}_j(t)$), which provide new information from the network other than its own observations $\{x_j(t)\}$. The term $\rho\mathbf{\Phi}_j^{-1}(t)\boldsymbol{\delta}_j(t-1)$ can be viewed as a Laplacian smoothing regularizer, which encourages all nodes of the graph to reach consensus on their estimates.

Remark 1: In D-RLS, (2) incurs computational complexity $O(p^2)$, since calculating the products $\mathbf{\Phi}_j^{-1}(t-1)\mathbf{h}_j(t)$ and $\mathbf{\Phi}_j^{-1}(t-1)\boldsymbol{\psi}_j(t)$ requires $O(p^2)$ multiplications. Similarly, (6) incurs *computational* complexity $O(p^2)$, that is dominated by the matrix-vector multiplications $\mathbf{\Phi}_j^{-1}(t)\mathbf{h}_j(t)$ and $\mathbf{\Phi}_j^{-1}(t)\boldsymbol{\delta}_j(t-1)$. The cost of carrying out (7) is relatively minor. Regarding *communication* cost per slot t, node j needs to transmit its local estimate $\mathbf{s}_j(t)$ to its neighbors and receive estimates $\mathbf{s}_{j'}(t)$ from all neighbors $j' \in \mathcal{N}_j$. The computational burden of D-RLS recursions (2)–(5) is comparable to that of (2), (6) and (7), with the cost of (4) being the same as what (6) requires. Meanwhile, the original form requires neighboring nodes j and j' to exchange $\mathbf{v}_j(t)$ and $\mathbf{v}_{j'}(t)$ in addition to $\mathbf{s}_j(t)$ and $\mathbf{s}_{j'}(t)$, which doubles the communication cost relative to (6) and (7).

C. Censoring-Based D-RLS Strategies

The D-RLS algorithm has well documented merits for decentralized online linear regression [15]. However, its computational and communication costs per iteration are fixed, regardless of whether observations and/or the estimates from neighboring nodes are informative or not. This fact motivates our idea of permeating benefits of data-adaptive censoring to *decentralized* RLS, through three novel censoring-based (C)D-RLS strategies. They are different from the RLS algorithms in [4], where the focus is on *centralized* online linear regression.

Algorithm 1: CD-RLS-1.

Initialize $\boldsymbol{\delta}_j(0)$, $\{\mathbf{s}_j(0)\}_{j=1}^J$ and $\{\mathbf{\Phi}_j^{-1}(0)\}_{j=1}^J$
for $t = 1, 2, \ldots$ **do**
  All $j \in \mathcal{V}$:
  if $|x_j(t) - \mathbf{h}_j^T(t)\mathbf{s}_j(t-1)| \leq \tau\sigma_j(t)$ **then**
    update $\mathbf{\Phi}_j^{-1}(t)$ using (9)
    update $\mathbf{s}_j(t)$ using (10)
  else
    update $\mathbf{\Phi}_j^{-1}(t)$ using (2)
    update $\mathbf{s}_j(t)$ using (6)
  end if
  transmit $\mathbf{s}_j(t)$ to and receive $\mathbf{s}_{j'}(t)$ from all $j' \in \mathcal{N}_j$
  compute $\boldsymbol{\delta}_j(t)$ using (7)
end for

Our first censoring strategy (CD-RLS-1) can be intuitively motivated as follows. If a given datum $(x_j(t), \mathbf{h}_j(t))$ is not informative enough, we do not have to use it since its contribution to the local estimate of node j, as well as to those of all network nodes, is limited. With $\{\tau\sigma_j(t)\}$ specifying proper thresholds to be discussed later, this intuition can be realized using a censoring indicator variable

$$c_j(t) := \begin{cases} 0, & \text{if } |x_j(t) - \mathbf{h}_j^T(t)\mathbf{s}_j(t-1)| \leq \tau\sigma_j(t) \\ 1, & \text{if } |x_j(t) - \mathbf{h}_j^T(t)\mathbf{s}_j(t-1)| > \tau\sigma_j(t). \end{cases} \tag{8}$$

If the absolute value of the innovation is less than $\tau\sigma_j(t)$, then $(x_j(t), \mathbf{h}_j(t))$ is censored; otherwise $(x_j(t), \mathbf{h}_j(t))$ is used. Section III-C will provide rules for selecting the threshold τ along with the local noise variance $\sigma_j^2(t)$, whose computations are lightweight. If data censoring is in effect, we simply throw away the current datum by letting $\mathbf{h}_j(t) = \mathbf{0}$ in (2), to obtain

$$\mathbf{\Phi}_j^{-1}(t) = \lambda^{-1}\mathbf{\Phi}_j^{-1}(t-1). \tag{9}$$

Likewise, letting $x_j(t) = 0$ and $\mathbf{h}_j(t) = \mathbf{0}$ in (6), yields

$$\mathbf{s}_j(t) = \mathbf{s}_j(t-1) - \rho\mathbf{\Phi}_j^{-1}(t)\boldsymbol{\delta}_j(t-1). \tag{10}$$

CD-RLS-1 is summarized in Algorithm 1. If censoring is in effect, computation cost per node and per slot is a fraction $2/7$ of the D-RLS in (4) and (7) without censoring. To recognize why, observe that the scalar-matrix multiplication $\lambda^{-1}\mathbf{\Phi}_j^{-1}(t-1)$ in (9) is not necessary as the update of $\mathbf{\Phi}_j^{-1}(t)$ can be merged to wherever it is needed, e.g., in (10) and the next slot. In addition, carrying out the $O(p^2)$ multiplications to obtain $\mathbf{\Phi}_j^{-1}(t)\mathbf{h}_j(t)$ is no longer necessary, while the $O(p^2)$ multiplications required to obtain $\mathbf{\Phi}_j^{-1}(t)\boldsymbol{\delta}_j(t-1)$ remain the same.

The first censoring strategy still requires nodes to communicate with neighbors per time slot; hence, the communication cost remains the same. Reducing this communication cost, motivates our second censoring strategy (CD-RLS-2), where each node does not perform extra computations relative to CD-RLS-1, but only receives neighboring estimates if its current datum is censored. The intuition behind this strategy is that if a datum

Algorithm 2: CD-RLS-2.

Initialize $\boldsymbol{\delta}_j(0)$, $\{\mathbf{s}_j(0)\}_{j=1}^J$ and $\{\boldsymbol{\Phi}_j^{-1}(0)\}_{j=1}^J$
for $t = 1, 2, \ldots$ **do**
  All $j \in \mathcal{V}$:
  if $|x_j(t) - \mathbf{h}_j^T(t)\mathbf{s}_j(t-1)| \leq \tau\sigma_j(t)$ **then**
    receives $\mathbf{s}_{j'}(t)$ from all $j' \in \mathcal{N}_j$
  else
    set $\mathbf{s}_{j'}(t-1)$ as recently received ones from all $j' \in \mathcal{N}_j$
    update $\boldsymbol{\Phi}_j^{-1}(t)$ using (2)
    update $\mathbf{s}_j(t)$ using (6)
    transmit $\mathbf{s}_j(t)$ to and receive $\mathbf{s}_{j'}(t)$ from all $j' \in \mathcal{N}_j$
    compute $\boldsymbol{\delta}_j(t)$ using (7)
  end if
end for

Algorithm 3: CD-RLS-3.

Initialize $\boldsymbol{\delta}_j(0)$, $\{\mathbf{s}_j(0)\}_{j=1}^J$ and $\{\boldsymbol{\Phi}_j^{-1}(0)\}_{j=1}^J$
for $t = 1, 2, \ldots$ **do**
  All $j \in \mathcal{V}$:
  if $|x_j(t) - \mathbf{h}_j^T(t)\mathbf{s}_j(t-1)| \leq \tau\sigma_j(t)$ **then**
    stay idle
  else
    set $\mathbf{s}_{j'}(t-1)$ as recently received ones from all $j' \in \mathcal{N}_j$
    update $\boldsymbol{\Phi}_j^{-1}(t)$ using (2)
    update $\mathbf{s}_j(t)$ using (6)
    transmit $\mathbf{s}_j(t)$ to and receive $\mathbf{s}_{j'}(t)$ from all $j' \in \mathcal{N}_j$
    compute $\boldsymbol{\delta}_j(t)$ using (7)
  end if
  if do not receive from any $j' \in \mathcal{N}_j$ for $d_{\max}$ time **then**
    receive $\mathbf{s}_{j'}(t)$
  end if
end for

is censored, then very likely the current local estimate is sufficiently accurate, and the node does not need to account for estimates from its neighbors. Estimates from neighbors, are only stored for future usage. Likewise, neighbors in $\mathcal{N}_j$ do not need node j's current estimate either, because they have already received a very similar estimate. CD-RLS-2 is summarized in Algorithm 2.

The third censoring strategy (CD-RLS-3) given by Algorithm 3 is more aggressive than the second one. If a node has its datum censored at a certain slot, then it neither transmits to nor receives from its neighbors, and in that sense it remains "isolated" from the rest of the network in this slot. Apparently, we should not allow any node to be forever isolated. To this end, we can force each node to receive the local estimate from any of its neighbors at least once every $d_{\max}$ slots, which upper bounds the delay of information exchange to $d_{\max}$. Interestingly, the ensuing section will prove convergence of all three strategies to the optimal argument in the mean-square deviation sense under mild conditions.

III. Development and Performance Analysis

This section starts with a criterion-based development of CD-RLS-1. Convergence analysis of all three censoring strategies will follow, before developing practical means of setting the censoring threshold $\tau\sigma_j(t)$.

A. Derivation of Censoring-Based D-RLS-1

Consider the following truncated quadratic cost that is similar to the one used in the censoring-based but centralized RLS [4]

$$f_{j,t}(\mathbf{s}) := \begin{cases} 0, & |x_j(t) - \mathbf{h}_j^T(t)\mathbf{s}| \leq \tau\sigma_j(t) \\ \frac{1}{2}[x_j(t) - \mathbf{h}_j^T(t)\mathbf{s}]^2 - \frac{1}{2}\tau^2\sigma_j(t)^2, & |x_j(t) - \mathbf{h}_j^T(t)\mathbf{s}| > \tau\sigma_j(t) \end{cases} \tag{11}$$

which is convex, but non-differentiable on $\{\mathbf{s} : |x_j(t) - \mathbf{h}_j^T(t)\mathbf{s}| = \tau\sigma_j(t)\}$. Using (11) to replace the quadratic loss $[x_j(\tau) - \mathbf{h}_j^T(\tau)\mathbf{s}]^2$ in (1), our CD-RLS-1 criterion is

$$\min_{\mathbf{s}} \sum_{r=1}^{t}\sum_{j=1}^{J} \lambda^{t-r} f_{j,r}(\mathbf{s}). \tag{12}$$

To solve (12) in a decentralized manner, we introduce a local estimate $\mathbf{s}_j$ per node j, along with auxiliary vectors $\bar{\mathbf{z}}_j^{j'}$ and $\tilde{\mathbf{z}}_j^{j'}$ per edge (j, j'). By constraining all local estimates of neighbors to consent, we arrive at the following equivalent separable convex program per slot t

$$\min_{\{\mathbf{s}_j\}_{j\in\mathcal{V}}} \quad \sum_{r=1}^{t}\sum_{j=1}^{J} \lambda^{t-r} f_{j,r}(\mathbf{s}_j) \tag{13}$$

$$\text{s.t.} \quad \mathbf{s}_j = \bar{\mathbf{z}}_j^{j'}, \mathbf{s}_{j'} = \tilde{\mathbf{z}}_j^{j'}, \bar{\mathbf{z}}_j^{j'} = \tilde{\mathbf{z}}_j^{j'}, j \in \mathcal{V}, j' \in \mathcal{N}_j.$$

Next, we employ alternating minimization and the stochastic Newton iteration to derive our first censoring-based solver of (13). To this end, consider the Lagrangian of (13) that is given by

$$\mathcal{L}(\mathbf{s}, \mathbf{z}, \mathbf{v}, \mathbf{u}) = \sum_{j\in\mathcal{V}}\sum_{r=1}^{t} \lambda^{t-r} f_{j,r}(\mathbf{s}_j) + \sum_{j=1}^{J}\sum_{j'\in\mathcal{N}_j} \left[(\mathbf{v}_j^{j'})^T(\mathbf{s}_j - \bar{\mathbf{z}}_j^{j'}) + (\mathbf{u}_j^{j'})^T(\mathbf{s}_{j'} - \tilde{\mathbf{z}}_j^{j'})\right] \tag{14}$$

where $\mathbf{s} := \{\mathbf{s}_j\}_{j\in\mathcal{V}}$ and $\mathbf{z} := \{\bar{\mathbf{z}}_j^{j'}, \tilde{\mathbf{z}}_j^{j'}\}_{j\in\mathcal{V}}^{j'\in\mathcal{N}_j}$ are primal variables, while $\mathbf{v} := \{\mathbf{v}_j^{j'} \in \mathbb{R}^p\}_{j\in\mathcal{V}}^{j'\in\mathcal{N}_j}$ and $\mathbf{u} := \{\mathbf{u}_j^{j'} \in \mathbb{R}^p\}_{j\in\mathcal{V}}^{j'\in\mathcal{N}_j}$ are dual variables. Consider also the augmented Lagrangian of (13), namely

$$\mathcal{L}_\rho(\mathbf{s}, \mathbf{z}, \mathbf{v}, \mathbf{u}) = \mathcal{L}(\mathbf{s}, \mathbf{z}, \mathbf{u}, \mathbf{v}) + \frac{\rho}{2}\sum_{j=1}^{J}\sum_{j'\in\mathcal{N}_j} \left[||\mathbf{s}_j - \bar{\mathbf{z}}_j^{j'}||^2 + ||\mathbf{s}_{j'} - \tilde{\mathbf{z}}_j^{j'}||^2\right] \tag{15}$$

where ρ is a positive regularization scale. Note that the constraints on $\mathbf{z}$ are not dualized, but they are collected in the set $\mathcal{C}_{\mathbf{z}} := \{\mathbf{z}|\bar{\mathbf{z}}_j^{j'} = \tilde{\mathbf{z}}_j^{j'}, j \in \mathcal{V}, j' \in \mathcal{N}_j, j \neq j'\}$.

To minimize (13) per slot $t > 0$, we rely on alternating minimization [27] in an online manner, which entails an iterative procedure consisting of three steps.

[S1] Local estimate updates:

$$\mathbf{s}(t) = \arg\min_{\mathbf{s}} \mathcal{L}(\mathbf{s}, \mathbf{z}(t-1), \mathbf{v}(t-1), \mathbf{u}(t-1))$$

[S2] Auxiliary variable updates:

$$\mathbf{z}(t) = \arg\min_{\mathbf{z}\in\mathcal{C}_{\mathbf{z}}} \mathcal{L}_\rho(\mathbf{s}(t), \mathbf{z}, \mathbf{v}(t-1), \mathbf{u}(t-1))$$

[S3] Multiplier updates:

$$\mathbf{v}_j^{j'}(t) = \mathbf{v}_j^{j'}(t-1) + \rho\big[\mathbf{s}_j(t) - \bar{\mathbf{z}}_j^{j'}(t)\big]$$
$$\mathbf{u}_j^{j'}(t) = \mathbf{u}_j^{j'}(t-1) + \rho\big[\mathbf{s}_{j'}(t) - \tilde{\mathbf{z}}_j^{j'}(t)\big].$$

Observe that [S2] is a linearly constrained quadratic program, for which if $\mathbf{v}_j^{j'}(t-1) + \mathbf{u}_j^{j'}(t-1) = \mathbf{0}$, we always have

$$\mathbf{s}_{j'}(t) + \mathbf{s}_j(t) = \tilde{\mathbf{z}}_j^{j'}(t) + \bar{\mathbf{z}}_j^{j'}(t) \quad \text{and} \quad \tilde{\mathbf{z}}_j^{j'}(t) = \bar{\mathbf{z}}_j^{j'}(t).$$

Therefore, the initial values of $\mathbf{v}_j^{j'}$ and $\mathbf{u}_j^{j'}$ in [S3] are selected to satisfy $\mathbf{v}_j^{j'}(0) + \mathbf{u}_j^{j'}(0) = \mathbf{0}$ (the simplest choice is $\mathbf{v}_j^{j'}(0) = \mathbf{u}_j^{j'}(0) = \mathbf{0}$). It then holds for $t \geq 0$ that

$$\mathbf{v}_j^{j'}(t) + \mathbf{u}_j^{j'}(t) = \mathbf{0}.$$

Using the latter to eliminate $\mathbf{u}_j^{j'}$ in [S3], we obtain

$$\begin{aligned}\mathbf{v}_j^{j'}(t) &= \mathbf{v}_j^{j'}(t-1) + \frac{\rho}{2}\big[\mathbf{s}_j(t) - \bar{\mathbf{z}}_j^{j'}(t) - \mathbf{s}_{j'}(t) + \tilde{\mathbf{z}}_j^{j'}(t)\big] \\ &= \mathbf{v}_j^{j'}(t-1) + \frac{\rho}{2}\big[\mathbf{s}_j(t) - \mathbf{s}_{j'}(t)\big] \qquad (16)\end{aligned}$$

where the first equality comes from subtracting the two lines in [S3], and the second equality is due to $\tilde{\mathbf{z}}_j^{j'}(t) = \bar{\mathbf{z}}_j^{j'}(t)$. The auxiliary variables $\tilde{\mathbf{z}}_j^{j'}$ and $\bar{\mathbf{z}}_j^{j'}$ can be also eliminated. When $\mathbf{v}_j^{j'}$ is initialized by $\mathbf{v}_j^{j'}(0) = \mathbf{0}$, summing up both sides of (16) from $r = 1$ to $r = t$, we arrive, after telescopic cancellation, at

$$\mathbf{v}_j^{j'}(t) = \frac{\rho}{2}\sum_{r=1}^{t}\big[\mathbf{s}_j(r) - \mathbf{s}_{j'}(r)\big]. \qquad (17)$$

Moving on to [S1], observe that it can be split into J per-node subproblems

$$\begin{aligned}\mathbf{s}_j(t) = \arg\min_{\mathbf{s}_j} &\sum_{r=1}^{t}\lambda^{t-r} f_{j,r}(\mathbf{s}_j) \\ &+ \sum_{j'\in\mathcal{N}_j}[\mathbf{v}_j^{j'}(t-1) - \mathbf{v}_{j'}^{j}(t-1)]^T\mathbf{s}_j.\end{aligned}$$

Before solving (11) with the stochastic Newton iteration [1], eliminate $\mathbf{v}_j^{j'}$ using (17) to obtain

$$\begin{aligned}\mathbf{s}_j(t) = \arg\min_{\mathbf{s}_j} &\sum_{r=1}^{t}\lambda^{t-r} f_{j,r}(\mathbf{s}_j) \\ &+ \rho\sum_{r=1}^{t-1}\sum_{j'\in\mathcal{N}_j}\big[\mathbf{s}_j(r) - \mathbf{s}_{j'}(r)\big]^T\mathbf{s}_j\end{aligned}$$

which after manipulating the double sum yields

$$\begin{aligned}\mathbf{s}_j(t) = \arg\min_{\mathbf{s}_j} &\sum_{r=1}^{t}\lambda^{t-r} f_{j,r}(\mathbf{s}_j) \\ &+ \sum_{r=1}^{t}\lambda^{t-r}\rho\sum_{j'\in\mathcal{N}_j}\Bigg[\mathbf{s}_j(r-1) - \mathbf{s}_{j'}(r-1) \\ &+ (1-\lambda)\sum_{\xi=1}^{r-1}\big(\mathbf{s}_j(\xi-1) - \mathbf{s}_{j'}(\xi-1)\big)\Bigg]^T\mathbf{s}_j.\end{aligned}$$

If the update in (7) is initialized with $\boldsymbol{\delta}_j(0) = \mathbf{0}$, summing up both sides from $\xi = 1$ to $\xi = r-1$, we find after telescopic cancellation

$$\begin{aligned}\boldsymbol{\delta}_j(r-1) = \sum_{j'\in\mathcal{N}_j}\Bigg[&\mathbf{s}_j(r-1) - \mathbf{s}_{j'}(r-1) \\ &+ (1-\lambda)\sum_{\xi=1}^{r-1}\big(\mathbf{s}_j(\xi-1) - \mathbf{s}_{j'}(\xi-1)\big)\Bigg]. \qquad (18)\end{aligned}$$

Thus, optimization of $\mathbf{s}_j(t)$ reduces to

$$\mathbf{s}_j(t) = \arg\min_{\mathbf{s}_j}\sum_{r=1}^{t}\lambda^{t-r} g_{j,r}(\mathbf{s}_j) \qquad (19)$$

where the instantaneous cost per slot t is

$$g_{j,t}(\mathbf{s}_j) := f_{j,t}(\mathbf{s}_j) + \rho\boldsymbol{\delta}_j^T(t-1)\mathbf{s}_j. \qquad (20)$$

The stochastic gradient of the latter is given by

$$\begin{aligned}&\nabla g_{j,t}(\mathbf{s}_j(t-1)) \\ &= -c_j(t)\Big[\big(x_j(t) - \mathbf{h}_j(t)\mathbf{s}_j(t-1)\big)\mathbf{h}_j(t)\Big] + \rho\boldsymbol{\delta}_j(t-1).\end{aligned}$$

In the stochastic Newton method, the Hessian matrix is given by

$$\mathbf{M}_j(t) = E[\nabla^2 g_{j,t}(\mathbf{s}_j(t-1))] = E[c_j(t)\mathbf{h}_j(t)\mathbf{h}_j^T(t)]$$

where the second equality comes from (11) and (8). A reasonable approximation of the expectation is provided by sample averaging. However, presence of $\lambda \neq 1$ affects attenuation of regressors, which leads to

$$\begin{aligned}\mathbf{M}_j(t) &= \frac{1}{t}\sum_{r=1}^{t}\lambda^{t-r} c_j(r)\mathbf{h}_j(r)\mathbf{h}_j^T(r) \\ &= \lambda\frac{t-1}{t}\mathbf{M}_j(t-1) + \frac{1}{t}c_j(t)\mathbf{h}_j(t)\mathbf{h}_j^T(t).\end{aligned}$$

Applying the matrix inversion lemma, we obtain

$$\mathbf{M}_j^{-1}(t) = \frac{t}{t-1}\Bigg[\lambda^{-1}\mathbf{M}_j^{-1}(t-1) - c_j(t)\frac{\lambda^{-1}\mathbf{M}_j^{-1}(t-1)\mathbf{h}_j(t)\mathbf{h}_j^T(t)\mathbf{M}_j^{-1}(t-1)}{(t-1)\lambda - \mathbf{h}^T(t)\mathbf{M}_j^{-1}(t-1)\mathbf{h}_j(t)}\Bigg] \tag{21}$$

and after adopting a diminishing step size $1/t$, the stochastic Newton update becomes

$$\mathbf{s}_j(t) = \mathbf{s}_j(t-1) - \frac{1}{t}\mathbf{M}_j^{-1}(t)\nabla g_{j,t}(\mathbf{s}_j(t-1)).$$

For rational convenience, let $\mathbf{\Phi}_j^{-1}(t) := \mathbf{M}_j^{-1}(t)/t$, and rewrite (21) as (cf. (2))

$$\mathbf{\Phi}_j^{-1}(t) = \lambda^{-1}\mathbf{\Phi}_j^{-1}(t-1) - c_j(t)\frac{\lambda^{-1}\mathbf{\Phi}_j^{-1}(t-1)\mathbf{h}_j(t)\mathbf{h}_j^T(t)\mathbf{\Phi}_j^{-1}(t-1)}{\lambda + \mathbf{h}_j^T(t)\mathbf{\Phi}_j^{-1}(t-1)\mathbf{h}_j(t)}. \tag{22}$$

Substituting $\nabla g_{j,t}(\mathbf{s}_j(t-1))$ and $\mathbf{\Phi}_j^{-1}(t)$ into the stochastic Netwon iteration yields (cf. (6))

$$\mathbf{s}_j(t) = \mathbf{s}_j(t-1) + c_j(t)\mathbf{\Phi}_j^{-1}(t)\mathbf{h}_j(t)\big[x_j(t) - \mathbf{h}_j^T(t)\mathbf{s}_j(t-1)\big] - \rho\mathbf{\Phi}_j^{-1}(t)\boldsymbol{\delta}_j(t-1)$$

which completes the development of CD-RLS-1.

B. Convergence analysis

Here we establish convergence of all three novel strategies for $\lambda = 1$. With $\lambda < 1$, the EWLS estimator can even adapt to time-varying parameter vectors, but analyzing its tracking performance goes beyond the scope of this paper. For the time-invariant case ($\lambda = 1$), we will rely on the following assumption.

(**as1**) Observations obey the linear model $x_j(t) = \mathbf{h}_j(t)\mathbf{s}_0 + \epsilon_j(t)$, where $\epsilon_j(t) \sim \mathcal{N}(0, \sigma_j^2)$ is correlated across j and t. Rows $\mathbf{h}_j^T(t)$ are uniformly bounded and independent of $\epsilon_j(t)$. Covariance matrices $\mathbf{R}_{h_j} := E[\mathbf{h}_j(t)\mathbf{h}_j^T(t)] \succ \mathbf{0}_{p\times p}$ are time-invariant and positive definite. Process $\{c_j(t)\mathbf{h}_j(t)\mathbf{h}_j^T(t)\}$ is mean ergodic, while $\{\epsilon_j(t)\}$ and $\{c_j(t)\}$ are uncorrelated. Eigenvalues of $\mathbf{\Phi}_j(t)/t$, which approximate the true positive definite Hessian matrices $E[c_j(t)\mathbf{h}_j(t)\mathbf{h}_j^T(t)]$, are bounded below by a positive constant when t is large enough.

We will assess convergence of our iterative algorithms using the squared mean-root deviation (SMRD) metric, defined as

$$\text{SMRD}(t) := \left\{E\left[\left(\sum_{j=1}^{J}||\mathbf{s}_j(t) - \mathbf{s}_0||^2\right)^{\frac{1}{2}}\right]\right\}^2. \tag{23}$$

Letting $\mathbf{e}_j(t) := \mathbf{s}_j(t) - \mathbf{s}_0 \in \mathbb{R}^p$ denote the estimation error of node j and $\mathbf{e}(t) := [\mathbf{e}_1^T(t), \ldots, \mathbf{e}_J^T(t)]^T \in \mathbb{R}^{Jp}$ the estimation error across all nodes, one can see that $\text{SMRD}(t) = \{E[\|\mathbf{e}(t)\|]\}^2$. Observe that $\text{SMRD}(t)$ is a lower-bound approximation of the mean-square deviation (MSD) metric $\text{MSD}(t) := E[\|\mathbf{e}(t)\|^2]$ [14], [26], since by Jensen's inequality $\{E[\|\mathbf{e}(t)\|]\}^2 \leq E[\|\mathbf{e}(t)\|^2]$.

Under (as1), convergence of CD-RLS-1 and CD-RLS-2 is asserted as follows; see Appendix B for the proof.

Theorem 1: For CD-RLS-1 and CD-RLS-2 Algorithms 1 and 2, set $\sigma_j(t) = \sigma_j$ and $\mathbf{\Phi}_j^{-1}(0) = \gamma\mathbf{I}_p$ per node j. Let $\mu := \min\{\lambda_{\min}(\mathbf{R}_{h_j}), j \in \mathcal{V}\}$, and suppose $0 < \rho < 1/(\gamma\lambda_{\max}(\mathbf{L}))$ for CD-RLS-1 and correspondingly $0 < \rho < \rho_0$ for CD-RLS-2, while $\mathbf{L}$ is the network Laplacian and the constant ρ_0 depends on $\lambda_{\max}(\mathbf{L}), \gamma, \tau, \mu$, and the upper bound of $\mathbf{h}_j(t)$. Under (as1), there exists $t_0 > 0$ for which it holds for $t > t_0$ that

$$\left\{E\left[\left(\sum_{j=1}^{J}||\mathbf{s}_j(t) - \mathbf{s}_0||^2\right)^{\frac{1}{2}}\right]\right\}^2 \leq \sum_{j=1}^{J}\frac{\gamma^{-1}||\mathbf{s}_j(0) - \mathbf{s}_0||^2 + \gamma t_0\sigma_j^2\text{tr}(\mathbf{R}_{h_j})}{2Q(\tau)\mu t} + \frac{\gamma\sigma_j^2\lambda_{\max}(\mathbf{R}_{h_j}^{-1})\text{tr}(\mathbf{R}_{h_j})\ln(t)}{4Q^2(\tau)\mu t}. \tag{24}$$

Theorem 1 establishes that the SMRD in (23) converges to zero at a rate $O(\ln(t)/t)$. The constant of the convergence rate is related to $\mathbf{R}_{h_j}$ through $\lambda_{\max}(\mathbf{R}_{h_j}^{-1})$, $\text{tr}(\mathbf{R}_{h_j})$ and μ; the noise covariance σ_j^2, and the threshold τ through $Q(\tau)$. Theorem 1 also indicates the impact of the initial states (determined by γ and $\mathbf{s}_j(0)$), which disappears at a faster rate of $O(1/t)$. To guarantee convergence, the step size ρ must be small enough.

The proof for CD-RLS-3 is more challenging. Because a node does not receive any information from its neighbors when censoring is in effect, it has to rely on outdated neighboring estimates when the incoming datum is not censored. This delay in percolating information may cause computational instability. For this reason, we will impose an additional constraint to guarantee that all local estimates do not grow unbounded. In practice, this can be realized by truncating local estimates when they exceed a certain threshold.

(**as2**) Local estimates $\{\mathbf{s}_j(t)\}_{j=1}^J$ are uniformly bounded $\forall t \geq 0$.

Convergence of CD-RLS-3 is then asserted as follows. Similar to CD-RLS-1 and CD-RLS-2, the SMRD of CD-RLS-3 converges to zero with rate $O(\ln(t)/t)$, as stated in the following theorem. The proof is omitted due to space limitations, but can be found in the arxiv version [28].

Theorem 2: For CD-RLS-3 given by Algorithms 3, set $\sigma_j(t) = \sigma_j$ and $\mathbf{\Phi}_j^{-1}(-1) = \gamma\mathbf{I}_p$ per node j. Under (as1) and (as2) with $0 < \rho < \rho_0$ as in Theorem 1, there exists $t_0 > 0$ for which it holds $\forall t > t_0$, that

$$\left\{E\left[\left(\sum_{j=1}^{J}||\mathbf{s}_j(t) - \mathbf{s}_0||^2\right)^{\frac{1}{2}}\right]\right\}^2 \leq \frac{a + b\ln(t)}{t} \tag{25}$$

where a and b are positive constants that depend on the upper bounds of $\mathbf{h}_j(t)$ and $\mathbf{s}_j(t)$, parameters ρ and τ, the covariance $\mathbf{R}_{h_j}(t)$, the Laplacian matrix $\mathbf{L}$, and t_0.

Although the bounds asserted by Theorems 1 and 2 could be loose, they demonstrate that $\limsup_{t\to\infty}\text{SMRD}(t) = 0$, which

TABLE I
AVERAGE PER STEP PER NODE COMMUNICATION AND COMPUTATIONAL COSTS, GIVEN THE AVERAGE CENSORING RATIO π^*

Algorithm	Communication	Computation
D-RLS	$2p\|\mathcal{E}\|/J$	$7p^2/2+O(p)$
CD-RLS-1	$2p\|\mathcal{E}\|/J$	$7p^2(1-\pi^*)/2+p^2\pi^*+O(p)$
CD-RLS-2	$2p\|\mathcal{E}\|(1-\pi^*)/J$	$7p^2(1-\pi^*)/2+O(p)$
CD-RLS-3	$2p\|\mathcal{E}\|(1-\pi^*)^2/J$	$7p^2(1-\pi^*)/2+O(p)$

establishes that the decentralized estimates converge to the ground truth asymptotically.

C. Threshold Setting and Variance Estimation

The threshold τ influences considerably the performance of all CD-RLS algorithms. Its value trades off estimation accuracy for computation and communication overhead. We provide a simple criterion for setting τ using the average censoring ratio π^*, which is defined as the number of censored data over the total number of data [19]. The goal is to choose τ so that the actual censoring ratio approaches π^* as t goes to infinity – since we are dealing with streaming big data, such an asymptotic property is certainly relevant. When t is large enough, $\mathbf{s}$ is very close to $\mathbf{s}_0$; thus, the innovation $x_j(t)-\mathbf{h}_j^T(t)\mathbf{s}_j(t-1)\approx x_j(t)-\mathbf{h}_j^T(t)\mathbf{s}_0=\epsilon_j(t)\sim\mathcal{N}(0,\sigma_j^2)$. As a consequence, $\Pr(c_j(t)=0)=\Pr(|x_j(t)-\mathbf{h}_j^T(t)\mathbf{s}_j(t-1)|\leq\tau\sigma_j)\approx\Pr(|\epsilon_j(t)|\leq\tau\sigma_j)=\Pr(|\epsilon_j(t)/\sigma_j|\leq\tau)=1-2Q(\tau)$, where the last equality holds because $\epsilon_j(t)/\sigma_j\sim\mathcal{N}(0,1)$. Therefore, $\pi^*=\lim_{t\to\infty}\frac{1}{t}\sum_{\tau=0}^{t}E[c_j(\tau)]\approx 1-2Q(\tau)$, which implies that

$$\tau=Q^{-1}((1-\pi^*)/2)\ .$$

Given the average censoring ratio π^*, Table I compares the average per step per node communication and computational costs of D-RLS and the proposed CD-RLS algorithms. We assume that transmitting or receiving a p-dimensional local estimate vector to or from a neighboring node incurs a cost of p. Thus, for D-RLS and CD-RLS-1, the average communication costs are both $2p|\mathcal{E}|/J$. In CD-RLS-2, a node does not transmit to its neighbors when it censors a datum, which leads to an average communication cost of $2p|\mathcal{E}|(1-\pi^*)/J$. CD-RLS-3 avoids communication over a link as long as one of the two end nodes censors a datum, and hence reduces the cost to $2p|\mathcal{E}|(1-\pi^*)^2/J$. As discussed in Section II-C, the computational costs of CD-RLS-1 for the non-censoring and censoring cases are $O(7p^2/2)$ and $O(p^2)$, respectively. For the censoring case, CD-RLS-2 and CD-RLS-3 reduce their computational costs to $O(p)$, and are more computationally efficient.

If the variances $\{\sigma_j^2\}$ were known, one could simply choose $\sigma_j(t)=\sigma_j$. However, σ_j in practice is often unknown. In this case, we consider the running average $\sigma_j^2(t+1)\approx t^{-1}\sum_{\tau=1}^{t+1}[x_j(\tau)-\mathbf{h}_j^T(\tau)\mathbf{s}_0]^2=(t-1)\sigma_j^2(t)/t+[x_j(t+1)-\mathbf{h}_j^T(t+1)\mathbf{s}_0]^2/t$, which suggests the recursive variance estimate

$$\sigma_j^2(t+1)=(t-1)\sigma_j^2(t)/t+[x_j(t+1)-\mathbf{h}_j^T(t+1)\mathbf{s}_j(t)]^2/t.$$

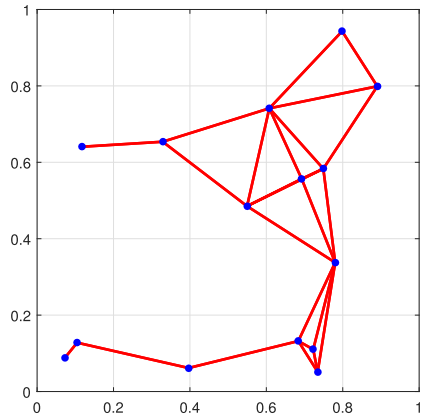


Fig. 1. The network topology used in the numerical experiments.

IV. NUMERICAL EXPERIMENTS

This section provides numerical results to validate the effectiveness of our novel censoring strategies. We simulate a network of $J=15$ nodes, which are uniformly randomly deployed over a 1×1 square. Two nodes within communication range 0.3 are deemed as being neighbors. The resultant network topology is depicted in Fig. 1. We compare six algorithms: the centralized adaptive censoring (AC)-RLS that runs in every node independently, the distributed diffusion least mean-square (Diffusion-LMS) algorithm [5], [16], D-RLS without censoring [17], and the three censoring-based D-RLS algorithms, namely CD-RLS-1, CD-RLS-2 and CD-RLS-3. All algorithms are evaluated on two data sets, one synthetic and one real. The empirical SMRD is used as performance metric.

For the synthetic data set, the unknown $\mathbf{s}_0$ is p-dimensional with $p=4$. The setting is the one in [17], where WSN-based decentralized power spectrum estimation is sought for a signal modeled as an autoregressive process. In this context, consider an auxiliary sequence $r_j(t)$ that evolves according to $r_j(t)=(1-q)\beta_j r_j(t-1)+\sqrt{q}\omega_j(t)$. Starting from $r_j(t)$, the row $\mathbf{h}_j^T(t)$ is formed by taking the next p observations, namely $\mathbf{h}_j^T(t)=[r_j(t+p-1);\ldots;r_j(t)]$. Parameters are selected as $q=0.5$, $\beta_j\sim\mathcal{U}(0,1)$, and also uniformly distributed driving white noise $\omega_j(t)\sim\mathcal{U}(-\sqrt{3}\sigma_{\omega_j},\sqrt{3}\sigma_{w_j})$ with $\sigma_{\omega_j}^2\sim\mathcal{U}(0,2)$. Observation of node j is subject to additive white Gaussian noise, with covariance $\sigma_j^2=10^{-3}\alpha_j$, where $\alpha_j\sim\mathcal{U}(0,1)$. The true signal vector is $\mathbf{s}_0=\mathbf{1}_p$, for which $\lambda=1$ is set for all algorithms. For D-RLS, CD-RLS-1, CD-RLS-2 and CD-RLS-3, the step size $\rho=0.01$ and $\mathbf{\Phi}_j^{-1}(0)=\gamma\mathbf{I}_p$ where $\gamma=30$, leading to fastest convergence of D-RLS. Regarding the four censoring-based algorithms AC-RLS, CD-RLS-1, CD-RLS-2 and CD-RLS-3, we set the average censoring ratio to $\pi^*=0.6$, which is approached using $\tau=Q^{-1}((1-\pi^*)/2)\approx 0.84$. The variances σ_j^2 are estimated in an online manner as described in Section III-C. AC-RLS uses $\mathbf{\Phi}_j^{-1}(0)=\gamma\mathbf{I}_p$, where $\gamma=10^5$ leads to the fastest convergence. Diffusion-LMS uses the nearest-neighbor diffusion matrix and $1.5/\sqrt{t}$ step size, which is tuned to obtain fastest convergence. For all curves obtained by running the algorithms, the ensemble averages are approximated via sample averaging over 100 Monte Carlo runs.

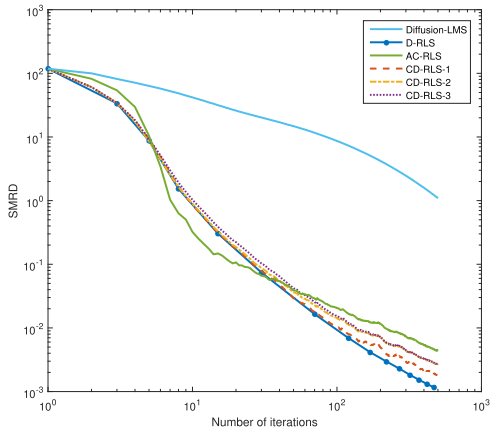


Fig. 2. SMRD of the six algorithms versus number of iterations.

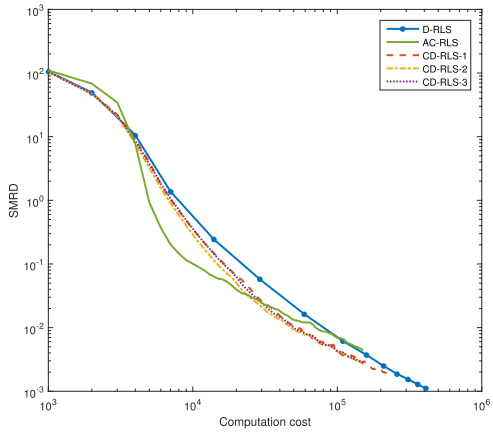


Fig. 3. SMRD of the five algorithms versus computational cost, defined as the number of multiplications.

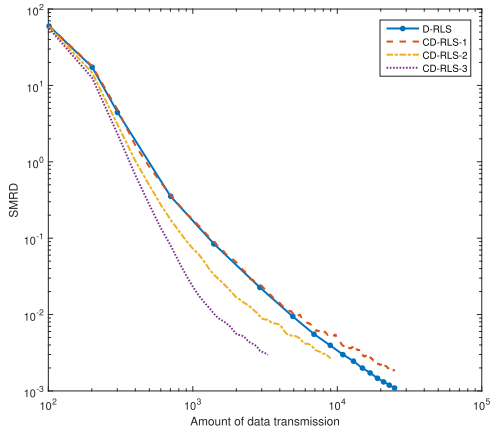


Fig. 4. SMRD of the four decentralized algorithms versus amount of data transmission in the unicast mode.

Fig. 2 depicts the SMRD versus the number of iterations. Not surprisingly, since D-RLS does not censor data, its convergence rate with respect to the number of iterations is the fastest. Among the three proposed CD-RLS algorithms, CD-RLS-2 and CD-RLS-3 are slower than CD-RLS-1, because the former two incur smaller communication cost than the latter. Though CD-RLS-3 adopts a more aggressive censoring strategy than CD-RLS-2, its convergence does not degrade as confirmed by Fig. 2. AC-RLS is the slowest among all except for Diffusion-LMS, because it is run at all nodes independently, without sharing information over the network. Even though the SMRD of Diffusion-LMS vanishes as $t \to \infty$ (with rate $1/t$), its finite-sample SMRD decays slower than our CD-RLS schemes for which SMRD also vanishes as $t \to \infty$ (with rate upper bounded by $\ln(t)/t$). This is analogous to *centralized* LMS that for finite samples exhibits SMRD decaying slower than that of *centralized* RLS. Note that due to the non-differentiable cost function (11), Diffusion-LMS is unable to achieve a linear convergence rate as in the differentiable case [5], [18]. We shall not compare with Diffusion-LMS in the rest of the numerical experiments.

The merits of censoring are further appreciated when one considers computational costs. Recall that the target average censoring ratio is $\pi^* = 0.6$, meaning that $3/5$ of the data are discarded (actual values are 0.6320 for AC-RLS, 0.6292 for CD-RLS-1, 0.6277 for CD-RLS-2, and 0.6237 for CD-RLS-3, averaged over 100 runs). As confirmed by Fig. 3, the three CD-RLS algorithms consume considerably less computational resources relative to D-RLS that does not censor data. Indeed, whenever a datum is censored, CD-RLS-1 only requires $2/7$ of the computations relative to D-RLS, while CD-RLS-2 and CD-RLS-3 incur minimal computational overhead. Although AC-RLS is the most computationally efficient algorithm at the beginning, absence of collaboration undermines its performance in steady state.

Regarding the amount of data exchanged to communicate local estimates in a unicast mode, CD-RLS-1 is the worst because nodes need to transmit their local estimate to neighbors, no matter whether local data are censored or not. Fig. 4 corroborates that CD-RLS-2 and CD-RLS-3 show significant improvement over D-RLS, demonstrating their potential for reducing both communication and computation costs in solving decentralized linear regression problems over large-scale networks.

We further numerically quantify the savings of computation and communication that the three censoring-based D-RLS algorithms enjoy over RLS without censoring. We set the target SMRD to 0.015 and plot the computational and communication costs required to reach it. According to Fig. 5, the computational costs of the three censoring-based algorithms decrease to about half of that of D-RLS as the censoring ratio grows to 0.7, while CD-RLS-2 outperforms the other two. Though CD-RLS-2 uses more iterations (hence more data) to achieve the target SMRD than CD-RLS-1 (see Fig. 2), it requires less computation when a datum is censored. On the other hand, CD-RLS-3 uses more iterations to achieve the target SMRD than CD-RLS-2, and hence it incurs more computational cost. The saving of CD-RLS-3 over CD-RLS-2 is mainly in the communication cost. In Fig. 6, the communication cost of CD-RLS-2 and CD-RLS-3 decreases as the censoring ratio grows, but that of CD-RLS-1 increases and is larger than that of D-RLS when the censoring ratio exceeds 0.5. CD-RLS-3 exhibits best performance in terms of communication cost.

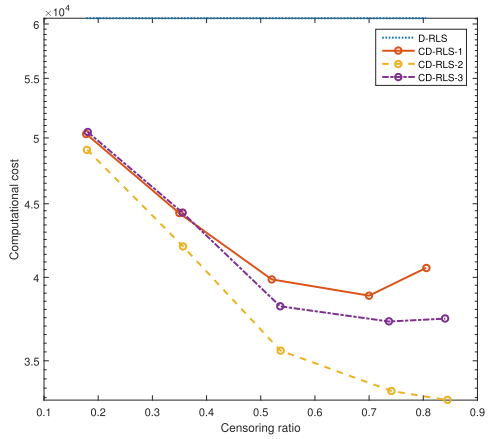


Fig. 5. Computational cost of the four decentralized algorithms for variable censoring ratios when target SMRD is 0.015.

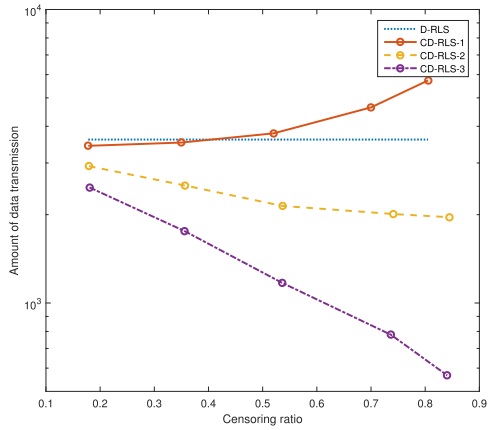


Fig. 6. Amount of data transmission of the four decentralized algorithms for variable censoring ratios when target SMRD is 0.015.

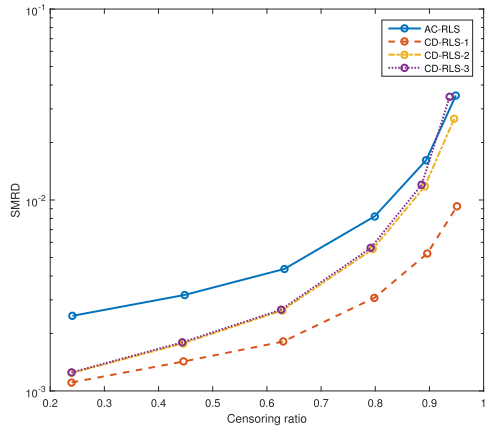


Fig. 7. SMRD after 500 iterations of the four censoring algorithms for variable censoring ratios.

Next, we vary π and evaluate its impact on SMRD, as shown in Fig. 7. The SMRD here is computed after 500 iterations. When π is close to 0.5, meaning about $1/2$ of the data is censored, the three proposed CD-RLS algorithms are still able to reach SMRD of 10^{-4}, which is the limit of D-RLS without censoring. Among

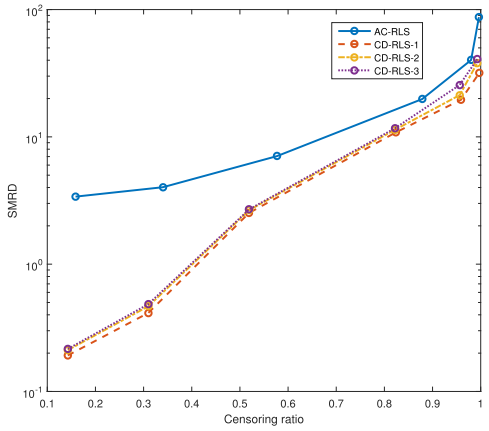


Fig. 8. SMRD of the four censoring algorithms versus the censoring ratio on a real data set of protein tertiary structures.

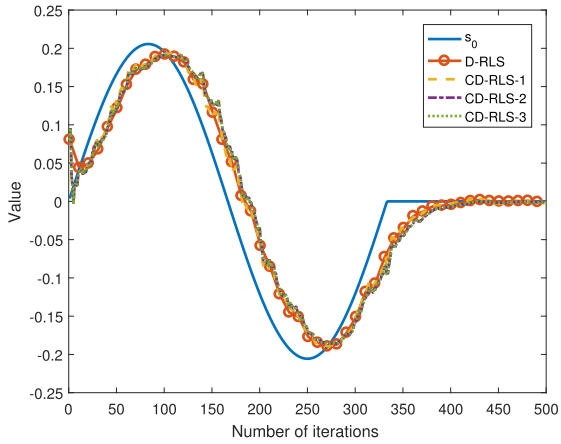


Fig. 9. First entries in the vector estimates of the four algorithms versus number of iterations when $\lambda = 0.95$.

the three algorithms, CD-RLS-1 exhibits the best SMRD curve, but its computation and communication costs are the highest. AC-RLS does not perform well especially for low censoring ratios due to the lack of network-wide collaboration. CD-RLS-2 and CD-RLS-3 perform comparably in this experiment.

The effectiveness of the novel censoring-based strategies is further assessed on a real data set of protein tertiary structures [12]. The premise here is that a given dataset is not available at a single location, but it is distributed over a network whose nodes are interested in obtaining accurate regression coefficients while suppressing the communication and computational overhead. Again, the graph in Fig. 1 is used to model the network of regression-performing agents. The number of control variables is $p = 9$. The first 45,720 (out of 45,730) observations are normalized and divided evenly into $J = 15$ parts, one per node. For CD-RLS-1, CD-RLS-2 and CD-RLS-3, we set $\rho = 0.05$ and $\mathbf{\Phi}_j^{-1}(0) = 5\mathbf{I}_p$, while for AC-RLS we choose $\gamma = 10$. The ground truth vector $\mathbf{s}_0$ is estimated by solving a batch least-squares problem on the entire data set. Similar to what we deduced from Fig. 7 in the synthetic data set, the novel CD-RLS algorithms outperform AC-RLS in terms of SMRD, as one

varies the average censoring ratio from 15% to nearly 100% in Fig. 8.

When $\lambda < 1$, the three censoring-based strategies are also able to track time-varying signals well. Note that to track the signal dynamics in this case, the censoring ratio cannot be too large. We use the same setting of the synthetic data but change the true $\mathbf{s}_0$ such that its ith element is $\tilde{\beta}_i \sin(3\pi t/500)$ when $t \leq 1000/3$, and remains constant after $t = 1000/3$. The magnitudes $\tilde{\beta}_i$ are i.i.d. and follow $\mathcal{U}(0,1)$. The parameters of the four decentralized algorithms are the same as those in the previous synthetic experiments, except that the censoring ratio is 0.3 when the censoring strategies are applied. Fig. 9 depicts the evolution of the first entries in the vector estimates of the four algorithms. They show similar tracking performance, but the censoring-based algorithms incur lower communication and computation costs over D-RLS.

V. Concluding Remarks

This paper introduced three data-adaptive censoring strategies that significantly reduce the computation and communication costs of the RLS algorithm over large-scale networks. The basic idea behind these strategies is to avoid inefficient computation and communication when the local observations and/or the neighboring messages are not informative. We proved convergence of the resulting algorithms in the mean-square deviation sense. Numerical experiments validated the merits of the novel schemes.

The notion of identifying and discarding less informative observations can be widely used in various large-scale online machine learning tasks including nonlinear regression, matrix completion, clustering and classification, to name a few. These constitute our future research directions.

Appendix A
Equivalent Form of D-RLS

Here we prove that D-RLS recursions (2)–(5) are equivalent to (2), (6) and (7). It follows from (4) that

$$\begin{aligned}
&\boldsymbol{\Phi}_j(t)\mathbf{s}_j(t) - \lambda\boldsymbol{\Phi}_j(t-1)\mathbf{s}_j(t-1)\\
&= \left[\boldsymbol{\psi}_j(t) - \frac{1}{2}\sum_{j'\in\mathcal{N}_j}\left(\mathbf{v}_j^{j'}(t-1) - \mathbf{v}_{j'}^{j}(t-1)\right)\right]\\
&\quad - \lambda\left[\boldsymbol{\psi}_j(t-1) - \frac{1}{2}\sum_{j'\in\mathcal{N}_j}\left(\mathbf{v}_j^{j'}(t-2) - \mathbf{v}_{j'}^{j}(t-2)\right)\right].
\end{aligned} \tag{26}$$

Applying the matrix inversion lemma to (2) yields

$$\boldsymbol{\Phi}_j(t) = \lambda\boldsymbol{\Phi}_j(t-1) + \mathbf{h}_j(t)\mathbf{h}_j^T(t). \tag{27}$$

Substituting $\boldsymbol{\psi}_j(t) - \lambda\boldsymbol{\psi}_j(t-1) = \mathbf{h}_j(t)x_j(t)$ from (3) and $\lambda\boldsymbol{\Phi}_j(t-1) = \boldsymbol{\Phi}_j(t) - \mathbf{h}_j(t)\mathbf{h}_j^T(t)$ from (27) into (26), leads to

$$\begin{aligned}
\boldsymbol{\Phi}_j(t)\big[\mathbf{s}_j(t) - \mathbf{s}_j(t-1)\big] &= \mathbf{h}_j(t)\big[x_j(t) - \mathbf{h}_j^T(t)\mathbf{s}_j(t-1)\big]\\
&\quad - \frac{1}{2}\sum_{j'\in\mathcal{N}_j}\left(\mathbf{v}_j^{j'}(t-1) - \lambda\mathbf{v}_j^{j'}(t-2)\right)\\
&\quad + \frac{1}{2}\sum_{j'\in\mathcal{N}_j}\left(\mathbf{v}_{j'}^{j}(t-1) - \lambda\mathbf{v}_{j'}^{j}(t-2)\right).
\end{aligned} \tag{28}$$

Next, we will show that if $\boldsymbol{\delta}(t)$ is defined as

$$\begin{aligned}
\boldsymbol{\delta}(t) &:= \frac{1}{2\rho}\sum_{j'\in\mathcal{N}_j}\left(\mathbf{v}_j^{j'}(t) - \lambda\mathbf{v}_j^{j'}(t-1)\right)\\
&\quad - \frac{1}{2\rho}\sum_{j'\in\mathcal{N}_j}\left(\mathbf{v}_{j'}^{j}(t) - \lambda\mathbf{v}_{j'}^{j}(t-1)\right)
\end{aligned} \tag{29}$$

then its update is exactly (7). This can be done by taking the difference between slots t and $t-1$ for (29), and substituting the update of $\mathbf{v}_j^{j'}$ in (5). Due to (29), it follows that (28) is equivalent to

$$\begin{aligned}
\boldsymbol{\Phi}_j(t)\big[\mathbf{s}_j(t) - \mathbf{s}_j(t-1)\big] &= \mathbf{h}_j(t)\big[x_j(t) - \mathbf{h}_j^T(t)\mathbf{s}_j(t-1)\big]\\
&\quad - \rho\boldsymbol{\delta}(t-1).
\end{aligned} \tag{30}$$

Left multiplying (30) with $\boldsymbol{\Phi}_j^{-1}(t)$, yields the update of $\mathbf{s}_j$ in (6), and completes the proof.

Appendix B
Proof of Theorem 1

Proof: We need the following lemma in [7, Chapter 7, Theorem 4].

Lemma 1: Let $X, X_1, X_2, ...$ be random variables on some probability space. If $X_n \to X$ in probability and $Pr(|X_n| \leq k) = 1$ for all n and some k, then $X_n \to X$ in rth mean for all $r \geq 1$.

Starting with CD-RLS-1, the proof proceeds in five stages.

Stage 1: We first investigate the spectral properties of $\boldsymbol{\Phi}_j(t)$ when t is sufficiently large. Letting $\lambda = 1$ and applying the matrix inversion lemma to the censoring form (2), we have

$$\boldsymbol{\Phi}_j(t) = \boldsymbol{\Phi}_j(t-1) + c_j(t)\mathbf{h}_j(t)\mathbf{h}_j^T(t). \tag{31}$$

Summing up from $r = 1$ to $r = t$ and using the telescopic cancellation, (31) yields

$$\boldsymbol{\Phi}_j(t) = \sum_{r=1}^{t} c_j(r)\mathbf{h}_j(r)\mathbf{h}_j^T(r) + \gamma^{-1}\mathbf{I}_p. \tag{32}$$

Thanks to the strong law of large numbers, $\boldsymbol{\Phi}_j(t)/t$ converges to $E[c_j(t)\mathbf{h}_j(t)\mathbf{h}_j^T(t)]$ almost surely as $t \to \infty$. Observe that

$$\begin{aligned}
&E[c_j(t)\mathbf{h}_j(t)\mathbf{h}_j^T(t)]\\
&= E\big[\mathbf{h}_j(t)\mathbf{h}_j^T(t)E[c_j(t)|\mathbf{h}_j(t), \mathbf{s}_j(t-1)]\big]\\
&= E[\mathbf{h}_j(t)\mathbf{h}_j^T(t)\Pr(c_j(t) = 1|\mathbf{h}_j(t), \mathbf{s}_j(t-1))].\\
&= E\left[\mathbf{h}_j(t)\mathbf{h}_j^T(t)\left(1 - \int_{-\tau+\sigma_j^{-1}[\mathbf{h}_j^T(t)(\mathbf{s}_j(t-1)-\mathbf{s}_0)]}^{\tau+\sigma_j^{-1}[\mathbf{h}_j^T(t)(\mathbf{s}_j(t-1)-\mathbf{s}_0)]}\phi(x)dx\right)\right].
\end{aligned} \tag{33}$$

Observing the integral in (33), we know that

$$1 > 1 - \int_{-\tau+\sigma_j^{-1}[\mathbf{h}_j^T(t)(\mathbf{s}_j(t-1)-\mathbf{s}_0)]}^{\tau+\sigma_j^{-1}[\mathbf{h}_j^T(t)(\mathbf{s}_j(t-1)-\mathbf{s}_0)]} \phi(x)dx$$
$$\geq 1 - \int_{-\tau}^{\tau} \phi(x)dx = 2Q(\tau) \quad (34)$$

where the event set that the second inequality strictly holds (namely, "$\geq$" becomes "$>$") is with nonzero measure. Thus, substituting (34) into (33) yields

$$E[c_j(t)\mathbf{h}_j(t)\mathbf{h}_j^T(t)] \prec E[\mathbf{h}_j(t)\mathbf{h}_j^T(t)] = \mathbf{R}_{h_j}$$

and

$$E[c_j(t)\mathbf{h}_j(t)\mathbf{h}_j^T(t)] \succ 2Q(\tau)E[\mathbf{h}_j(t)\mathbf{h}_j^T(t)] = 2Q(\tau)\mathbf{R}_{h_j}.$$

Since $\mathbf{\Phi}_j(t)/t$ converges to $E[c_j(t)\mathbf{h}_j(t)\mathbf{h}_j^T(t)]$ almost surely as $t \to \infty$ and $\mathbf{h}_j(t)$ is uniformly bounded such that $\mathbf{\Phi}_j(t)/t$ is also bounded (cf. (32)), we have $E[\|\mathbf{\Phi}_j(t)/t\|_2]$ converges to $E[\|c_j(t)\mathbf{h}_j(t)\mathbf{h}_j^T(t)\|_2]$ as $t \to \infty$ by Lemma 1. Therefore, $2Q(\tau)\mathbf{R}_{h_j} \prec E[c_j(t)\mathbf{h}_j(t)\mathbf{h}_j^T(t)] \prec \mathbf{R}_{h_j}$ implies that there exists $t_1 > 0$, for which it holds $\forall t \geq t_1$ that

$$2Q(\tau)\|\mathbf{R}_{h_j}\|_2 < E[\|\mathbf{\Phi}_j(t)/t\|_2] < \|\mathbf{R}_{h_j}\|_2$$

and consequently the expected maximum eigenvalue of $\mathbf{\Phi}_j(t)$ satisfies

$$2Q(\tau)\lambda_{\max}(\mathbf{R}_{h_j})t < E[\lambda_{\max}(\mathbf{\Phi}_j(t))] < \lambda_{\max}(\mathbf{R}_{h_j})t. \quad (35)$$

Observe that $t\mathbf{\Phi}_j^{-1}(t)$ converges to $\left\{E[c_j(t)\mathbf{h}_j(t)\mathbf{h}_j^T(t)]\right\}^{-1}$ almost surely as $t \to \infty$ due to the convergence of $\mathbf{\Phi}_j(t)/t$ to $E[c_j(t)\mathbf{h}_j(t)\mathbf{h}_j^T(t)]$. Since eigenvalues of $\mathbf{\Phi}_j(t)/t$ are bounded below by a positive constant when t is large enough, there exists $t_2 > 0$ such that $t\mathbf{\Phi}_j^{-1}(t)$ is bounded $\forall t \geq t_2$. Following the same analysis to obtain (35), it holds $\forall t \geq t_2$ that

$$\lambda_{\max}(\mathbf{R}_{h_j}^{-1})/t < E[\lambda_{\max}(\mathbf{\Phi}_j^{-1}(t))] < \lambda_{\max}(\mathbf{R}_{h_j}^{-1})/(2Q(\tau)t). \quad (36)$$

Letting $t_0 := \max(t_1, t_2)$, (35) and (36) hold $\forall t \geq t_0$.

Stage 2: Rewrite the update of $\mathbf{s}_j$ as

$$\mathbf{s}_j(t) = \mathbf{s}_j(t-1) + c_j(t)\mathbf{\Phi}_j^{-1}(t)\mathbf{h}_j(t)\big[x_j(t) - \mathbf{h}_j^T(t)\mathbf{s}_j(t-1)\big] - \rho\mathbf{\Phi}_j^{-1}(t)\boldsymbol{\delta}_j(t-1).$$

Note also that for $\lambda = 1$, the update of $\boldsymbol{\delta}_j$ is equivalent to (cf. (18))

$$\boldsymbol{\delta}_j(t-1) = \sum_{j' \in \mathcal{N}_j} \big[\mathbf{s}_j(t-1) - \mathbf{s}_{j'}(t-1)\big].$$

Letting $\mathbf{e}_j(t) := \mathbf{s}_j(t) - \mathbf{s}_0$, the estimation error obeys the recursion

$$\mathbf{e}_j(t) = \mathbf{e}_j(t-1) + c_j(t)\mathbf{\Phi}_j^{-1}(t)\mathbf{h}_j(t)[x_j(t) - \mathbf{h}_j^T(t)\mathbf{s}_j(t-1)] - \rho\mathbf{\Phi}_j^{-1}(t)\sum_{j' \in \mathcal{N}_j}\big[\mathbf{e}_j(t-1) - \mathbf{e}_{j'}(t-1)\big].$$

Substituting $x_j(t) = \mathbf{h}_j(t)\mathbf{s}_0 + \epsilon_j(t)$ to eliminate $\mathbf{s}_j(t-1)$, we obtain

$$\mathbf{e}_j(t) = \mathbf{e}_j(t-1) - c_j(t)\mathbf{\Phi}_j^{-1}(t)\mathbf{h}_j(t)\mathbf{h}_j^T(t)\mathbf{e}_j(t-1) + c_j(t)\mathbf{\Phi}_j^{-1}(t)\mathbf{h}_j(t)\epsilon_j(t) - \rho\mathbf{\Phi}_j^{-1}(t)\sum_{j' \in \mathcal{N}_j}\big[\mathbf{e}_j(t-1) - \mathbf{e}_{j'}(t-1)\big]. \quad (37)$$

Left multiplying (37) with $\mathbf{\Phi}_j(t)$ yields

$$\begin{aligned}
&\mathbf{\Phi}_j(t)\mathbf{e}_j(t)\\
&= \mathbf{\Phi}_j(t)\mathbf{e}_j(t-1) - c_j(t)\mathbf{h}_j(t)\mathbf{h}_j^T(t)\mathbf{e}_j(t-1)\\
&\quad + c_j(t)\mathbf{h}_j(t)\epsilon_j(t) - \rho\sum_{j' \in \mathcal{N}_j}\big[\mathbf{e}_j(t-1) - \mathbf{e}_{j'}(t-1)\big]\\
&= \mathbf{\Phi}_j(t-1)\mathbf{e}_j(t-1)\\
&\quad + c_j(t)\mathbf{h}_j(t)\epsilon_j(t) - \rho\sum_{j' \in \mathcal{N}_j}\big[\mathbf{e}_j(t-1) - \mathbf{e}_{j'}(t-1)\big].
\end{aligned} \quad (38)$$

Our convergence analysis result will rely on a matrix form of (38) that accounts for all nodes j. Define vectors $\mathbf{e}(t) := [\mathbf{e}_1^T(t), \ldots, \mathbf{e}_J^T(t)]^T \in \mathbb{R}^{Jp}$, $\boldsymbol{\epsilon}(t) := [\epsilon_1^T(t), \ldots, \epsilon_J^T(t)]^T \in \mathbb{R}^J$, as well as block-diagonal matrices $\mathbf{\Phi}(t) := \text{diag}(\{\mathbf{\Phi}_j(t)\}) \in \mathbb{R}^{Jp \times Jp}$, $\mathbf{C}(t) := \text{diag}(\{c_j(t)\}) \in \mathbb{R}^{J \times J}$, and $\mathbf{H}(t) := \text{diag}(\{\mathbf{h}_j(t)\}) \in \mathbb{R}^{Jp \times J}$. Then (38) can be written in matrix form as

$$\begin{aligned}
&\mathbf{\Phi}(t)\mathbf{e}(t)\\
&= \big[\mathbf{\Phi}(t-1) - \rho\mathbf{L} \otimes \mathbf{I}_p\big]\mathbf{e}(t-1) + \mathbf{H}(t)\mathbf{C}(t)\boldsymbol{\epsilon}(t)
\end{aligned} \quad (39)$$

which after left multiplication with $\mathbf{\Phi}^{-\frac{1}{2}}(t)$ yields

$$\mathbf{\Phi}^{\frac{1}{2}}(t)\mathbf{e}(t) = \mathbf{\Phi}^{-\frac{1}{2}}(t)\big[\mathbf{\Phi}(t-1) - \rho\mathbf{L} \otimes \mathbf{I}_p\big]\mathbf{e}(t-1) + \mathbf{\Phi}^{-\frac{1}{2}}(t)\mathbf{H}(t)\mathbf{C}(t)\boldsymbol{\epsilon}(t). \quad (40)$$

From (40), we have ($\otimes$ denotes Kronecker product)

$$\begin{aligned}
&E[\mathbf{e}^T(t)\mathbf{\Phi}(t)\mathbf{e}(t)]\\
&= E[\mathbf{e}^T(t-1)(\mathbf{\Phi}(t-1) - \rho\mathbf{L} \otimes \mathbf{I}_p)^T\mathbf{\Phi}^{-1}(t)\\
&\times (\mathbf{\Phi}(t-1) - \rho\mathbf{L} \otimes \mathbf{I}_p)\mathbf{e}(t-1)]\\
&+ 2E[\mathbf{e}^T(t-1)(\mathbf{\Phi}(t-1) - \rho\mathbf{L} \otimes \mathbf{I}_p)^T\mathbf{\Phi}^{-1}(t)\mathbf{H}(t)\mathbf{C}(t)\boldsymbol{\epsilon}(t)]\\
&+ E[\boldsymbol{\epsilon}^T(t)\mathbf{C}^T(t)\mathbf{H}^T(t)\mathbf{\Phi}^{-1}(t)\mathbf{H}(t)\mathbf{C}(t)\boldsymbol{\epsilon}(t)].
\end{aligned}$$

Since $\mathbf{C}(t)$ and $\boldsymbol{\epsilon}(t)$ are irrelevant under (as1), the second term on the right hand side is zero; hence,

$$\begin{aligned}
&E[\mathbf{e}^T(t)\mathbf{\Phi}(t)\mathbf{e}(t)]\\
&= E[\mathbf{e}^T(t-1)(\mathbf{\Phi}(t-1) - \rho\mathbf{L} \otimes \mathbf{I}_p)^T\mathbf{\Phi}^{-1}(t)\\
&\quad \times (\mathbf{\Phi}(t-1) - \rho\mathbf{L} \otimes \mathbf{I}_p)\mathbf{e}(t-1)]\\
&\quad + E[\boldsymbol{\epsilon}^T(t)\mathbf{C}^T(t)\mathbf{H}^T(t)\mathbf{\Phi}^{-1}(t)\mathbf{H}(t)\mathbf{C}(t)\boldsymbol{\epsilon}(t)].
\end{aligned} \quad (41)$$

Stage 3: Consider the first term on the right hand side of (41). Since $\mathbf{L}$ is positive semi-definite, we can find a matrix $\mathbf{U} =$

$(\mathbf{L}\otimes\mathbf{I}_p)^{\frac{1}{2}}$ such that $\mathbf{L}\otimes\mathbf{I}_p=\mathbf{U}^T\mathbf{U}$. By the matrix inversion lemma, it holds that

$$\begin{aligned}&(\mathbf{\Phi}(t-1)-\rho\mathbf{L}\otimes\mathbf{I}_p)^{-1}\\&=(\mathbf{\Phi}(t-1)-\rho\mathbf{U}^T\mathbf{U})^{-1}\\&=\mathbf{\Phi}^{-1}(t-1)+\rho\mathbf{\Phi}^{-1}(t-1)\mathbf{U}^T\\&\quad\times(\mathbf{I}_{Jp}-\rho\mathbf{U}\mathbf{\Phi}^{-1}(t-1)\mathbf{U}^T)^{-1}\mathbf{U}\mathbf{\Phi}^{-1}(t-1).\end{aligned}\tag{42}$$

For $\lambda=1$, it follows from (2) that $\mathbf{\Phi}^{-1}(t-1)-\mathbf{\Phi}^{-1}(t)\succeq\mathbf{0}_{Jp}$. Since $\mathbf{\Phi}^{-1}(0)=\gamma\mathbf{I}_{Jp}$, it holds that $\mathbf{\Phi}^{-1}(t-1)\preceq\gamma\mathbf{I}_{Jp}$ for all $t\geq1$, and consequently

$$\mathbf{I}_{Jp}-\rho\mathbf{U}\mathbf{\Phi}^{-1}(t-1)\mathbf{U}^T\succeq\mathbf{I}_{Jp}-\rho\gamma\mathbf{U}\mathbf{U}^T=\mathbf{I}_{Jp}-\rho\gamma\mathbf{L}\otimes\mathbf{I}_p.$$

If $0<\rho<1/(\gamma\lambda_{\max}(\mathbf{L}))$, then for all $t\geq1$ it follows that

$$\mathbf{I}_{Jp}-\rho\mathbf{U}\mathbf{\Phi}^{-1}(t-1)\mathbf{U}^T\succeq\mathbf{0}_{Jp}.$$

This implies that the second term of (42) is positive definite. Thus, we have

$$\mathbf{\Phi}^{-1}(t)\preceq\mathbf{\Phi}^{-1}(t-1)\preceq(\mathbf{\Phi}(t-1)-\rho\mathbf{L}\otimes\mathbf{I}_p)^{-1}\tag{43}$$

and hence, the first term on the right hand side of (41) is bounded by

$$\begin{aligned}&E[\mathbf{e}^T(t-1)(\mathbf{\Phi}(t-1)-\rho\mathbf{L}\otimes\mathbf{I}_p)^T\mathbf{\Phi}^{-1}(t)\\&\quad\times(\mathbf{\Phi}(t-1)-\rho\mathbf{L}\otimes\mathbf{I}_p)\mathbf{e}(t-1)]\\&\leq E[\mathbf{e}^T(t-1)(\mathbf{\Phi}(t-1)-\rho\mathbf{L}\otimes\mathbf{I}_p)^T\mathbf{e}(t-1)]\\&\leq E[\mathbf{e}^T(t-1)\mathbf{\Phi}(t-1)^T\mathbf{e}(t-1)].\end{aligned}\tag{44}$$

Stage 4: Now consider the second term on the right hand side of (41). Manipulating the expectation yields

$$\begin{aligned}&E[\boldsymbol{\epsilon}^T(t)\mathbf{C}^T(t)\mathbf{H}^T(t)\mathbf{\Phi}^{-1}(t)\mathbf{H}(t)\mathbf{C}(t)\boldsymbol{\epsilon}(t)]\\&=E[\mathrm{tr}(\boldsymbol{\epsilon}^T(t)\mathbf{C}^T(t)\mathbf{H}^T(t)\mathbf{\Phi}^{-1}(t)\mathbf{H}(t)\mathbf{C}(t)\boldsymbol{\epsilon}(t))]\\&=E[\mathrm{tr}(\mathbf{C}^T(t)\mathbf{H}^T(t)\mathbf{\Phi}^{-1}(t)\mathbf{H}(t)\mathbf{C}(t)\boldsymbol{\epsilon}(t)\boldsymbol{\epsilon}^T(t))]\\&=E[\mathrm{tr}(\mathbf{C}^T(t)\mathbf{H}^T(t)\mathbf{\Phi}^{-1}(t)\mathbf{H}(t)\mathbf{C}(t)\mathrm{diag}(\{\sigma_j^2\})].\end{aligned}$$

where $\mathrm{diag}(\{\sigma_j^2\})\in\mathbb{R}^{J\times J}$ is a diagonal matrix constructed with $\{\sigma_j^2\}_{j=1}^J$ on its diagonal. Expanding the matrix multiplications and noting that $c_j(t)\leq1$, we obtain

$$\begin{aligned}&E[\boldsymbol{\epsilon}^T(t)\mathbf{C}^T(t)\mathbf{H}^T(t)\mathbf{\Phi}^{-1}(t)\mathbf{H}(t)\mathbf{C}(t)\boldsymbol{\epsilon}(t)]\\&\leq\sum_{j=1}^J\sigma_j^2E[\mathbf{h}_j^T(t)\mathbf{\Phi}_j^{-1}(t)\mathbf{h}_j(t)].\end{aligned}$$

Because $\mathbf{\Phi}_j^{-1}(t-1)\succeq\mathbf{\Phi}_j^{-1}(t)$ due to (22), we further have

$$\begin{aligned}&E[\boldsymbol{\epsilon}^T(t)\mathbf{C}^T(t)\mathbf{H}^T(t)\mathbf{\Phi}^{-1}(t)\mathbf{H}(t)\mathbf{C}(t)\boldsymbol{\epsilon}(t)]\\&\leq\sum_{j=1}^J\sigma_j^2E[\mathbf{h}_j^T(t)\mathbf{\Phi}_j^{-1}(t-1)\mathbf{h}_j(t)]\\&\leq\sum_{j=1}^J\sigma_j^2E[\lambda_{\max}(\mathbf{\Phi}_j^{-1}(t-1))\|\mathbf{h}_j(t)\|^2].\end{aligned}\tag{45}$$

Since $\mathbf{\Phi}_j^{-1}(t-1)$ and $\mathbf{h}_j(t)$ are independent, it holds $\forall t>t_0$ that

$$\begin{aligned}&E[\lambda_{\max}(\mathbf{\Phi}_j^{-1}(t-1))\|\mathbf{h}_j(t)\|^2]\\&=E[\lambda_{\max}(\mathbf{\Phi}_j^{-1}(t-1))]E[\|\mathbf{h}_j(t)\|^2]\\&<\frac{\lambda_{\max}(\mathbf{R}_{h_j}^{-1})}{2Q(\tau)(t-1)}\mathrm{tr}(\mathbf{R}_{h_j}).\end{aligned}\tag{46}$$

The inequality is due to (36) that shows $E[\lambda_{\max}(\mathbf{\Phi}_j^{-1}(t))]<\lambda_{\max}(\mathbf{R}_{h_j}^{-1})/(2Q(\tau)t)$, $\forall t\geq t_0$ and the fact $E[\|\mathbf{h}_j(t)\|^2]=\mathrm{tr}(E[\mathbf{h}_j(t)\mathbf{h}_j^T(t)])=\mathrm{tr}(\mathbf{R}_{h_j})$. Using (46) allows one to deduce from (45) that

$$\begin{aligned}&E[\boldsymbol{\epsilon}^T(t)\mathbf{C}^T(t)\mathbf{H}^T(t)\mathbf{\Phi}^{-1}(t)\mathbf{H}(t)\mathbf{C}(t)\boldsymbol{\epsilon}(t)]\\&\leq\frac{1}{2Q(\tau)(t-1)}\sum_{j=1}^J\sigma_j^2\lambda_{\max}(\mathbf{R}_{h_j}^{-1})\mathrm{tr}(\mathbf{R}_{h_j})\end{aligned}\tag{47}$$

holds $\forall t>t_0$.

For $t\leq t_0$, we have $\mathbf{\Phi}_j^{-1}(t)\preceq\mathbf{\Phi}_j^{-1}(0)=\gamma\mathbf{I}_p$ because to (43), and thus

$$\begin{aligned}&\sum_{j=1}^J\sigma_j^2E[\mathbf{h}_j^T(t)\mathbf{\Phi}_j^{-1}(t)\mathbf{h}_j(t)]\\&\leq\sum_{j=1}^J\gamma\sigma_j^2E[\mathbf{h}_j^T(t)\mathbf{h}_j(t)]=\gamma\sum_{j=1}^J\sigma_j^2tr(\mathbf{R}_{h_j}).\end{aligned}$$

Therefore, for $t\leq t_0$ (45) yields

$$\begin{aligned}&E[\boldsymbol{\epsilon}^T(t)\mathbf{C}^T(t)\mathbf{H}^T(t)\mathbf{\Phi}^{-1}(t)\mathbf{H}(t)\mathbf{C}(t)\boldsymbol{\epsilon}(t)]\\&\leq\gamma\sum_{j=1}^J\sigma_j^2\mathrm{tr}(\mathbf{R}_{h_j}).\end{aligned}\tag{48}$$

Stage 5: Substituting (44), (47) and (48) into (41) implies for $t>t_0$ that

$$\begin{aligned}&E[\mathbf{e}^T(t)\mathbf{\Phi}(t)\mathbf{e}(t)]\\&\leq E[\mathbf{e}^T(t-1)\mathbf{\Phi}(t-1)\mathbf{e}(t-1)]\\&\quad+\frac{1}{2Q(\tau)(t-1)}\sum_{j=1}^J\sigma_j^2\lambda_{\max}(\mathbf{R}_{h_j}^{-1})\mathrm{tr}(\mathbf{R}_{h_j})\end{aligned}\tag{49}$$

while for $t\leq t_0$

$$\begin{aligned}&E[\mathbf{e}^T(t)\mathbf{\Phi}(t)\mathbf{e}(t)]\\&\leq E[\mathbf{e}^T(t-1)\mathbf{\Phi}(t-1)\mathbf{e}(t-1)]+\gamma\sum_{j=1}^J\sigma_j^2\mathrm{tr}(\mathbf{R}_{h_j}).\end{aligned}\tag{50}$$

Summing (49) from $r=t_0+1$ to $r=t$ and (50) from $r=1$ to $r=t_0$, applying telescopic cancellation, and noticing that

$\mathbf{\Phi}(0)=\gamma^{-1}\mathbf{I}_{Jp}$, yields for $t>t_0$

$$E[\mathbf{e}^T(t)\mathbf{\Phi}(t)\mathbf{e}(t)] \tag{51}$$
$$\le \gamma^{-1}||\mathbf{e}(0)||^2+\left(\gamma t_0+\sum_{r=t_0+1}^{t}\frac{\lambda_{\max}(\mathbf{R}_{h_j}^{-1})}{2Q(\tau)(t-1)}\right)\times\sum_{j=1}^{J}\sigma_j^2\mathrm{tr}(\mathbf{R}_{h_j})$$
$$\le \gamma^{-1}||\mathbf{e}(0)||^2+\left(\gamma t_0+\frac{\lambda_{\max}(\mathbf{R}_{h_j}^{-1})}{2Q(\tau)}\ln(t)\right)\sum_{j=1}^{J}\sigma_j^2\mathrm{tr}(\mathbf{R}_{h_j}).$$

On the other hand, it holds

$$E[\mathbf{e}^T(t)\mathbf{\Phi}(t)\mathbf{e}(t)]\ge E[\|\mathbf{e}(t)\|^2/\lambda_{\max}(\mathbf{\Phi}^{-1}(t))]\ge E[\|\mathbf{e}(t)\|]^2/E[\lambda_{\max}(\mathbf{\Phi}^{-1}(t))]$$

where the last line is due to Cauchy-Schwarz inequality

$$E[\|\mathbf{e}(t)\|^2/\lambda_{\max}(\mathbf{\Phi}^{-1}(t))]E[\lambda_{\max}(\mathbf{\Phi}^{-1}(t))]$$
$$=E[\Big(\|\mathbf{e}(t)\|/\lambda_{\max}(\mathbf{\Phi}^{-1}(t))^{\frac{1}{2}}\Big)^2]E[\Big(\lambda_{\max}(\mathbf{\Phi}^{-1}(t))^{\frac{1}{2}}\Big)^2]$$
$$\ge E[\|\mathbf{e}(t)\|]^2.$$

From (36), $E[\lambda_{\max}(\mathbf{\Phi}_j^{-1}(t))]<\lambda_{\max}(\mathbf{R}_{h_j}^{-1})/(2Q(\tau)t)=1/(\lambda_{\min}(\mathbf{R}_{h_j})2Q(\tau)t)$ holds asymptotically. Defininingμ $:=\min\{\lambda_{\min}(\mathbf{R}_{h_j}),j\in\mathcal{V}\}$, we establish that

$$2Q(\tau)\mu t E[||\mathbf{e}(t)||^2]\le E[\mathbf{e}^T(t)\mathbf{\Phi}(t)\mathbf{e}(t)],\quad t>t_0. \tag{52}$$

Combining (51) and (52) implies

$$2Q(\tau)\mu t E[||\mathbf{e}(t)||]^2$$
$$\le\gamma^{-1}||\mathbf{e}(0)||^2+(\gamma t_0+\frac{\lambda_{\max}(\mathbf{R}_{h_j}^{-1})}{2Q(\tau)}\ln(t))\sum_{j=1}^{J}\sigma_j^2\mathrm{tr}(\mathbf{R}_{h_j}). \tag{53}$$

Finally, with $||\mathbf{e}(t)||^2:=\sum_{j=1}^{J}||\mathbf{e}_j(t)||^2=\sum_{j=1}^{J}||\mathbf{s}_j(t)-\mathbf{s}_0||^2$ this leads to (24), which completes the proof of CD-RLS-1.

Consider next CD-RLS-2. Stage 1 of the proof remains the same, while for Stage 2, $\mathbf{e}_j(t-1)-\mathbf{e}_{j'}(t-1)$ is replaced by $c_j(t)\big[\mathbf{e}_j(t-1)-\mathbf{e}_{j'}(t-1)\big]$ in (38) to arrive at

$$\mathbf{\Phi}_j(t)\mathbf{e}_j(t) \tag{54}$$
$$=\mathbf{\Phi}_j(t-1)\mathbf{e}_j(t-1)-c_j(t)\mathbf{h}_j(t)\mathbf{h}_j^T(t)\mathbf{e}_j(t-1)+c_j(t)\mathbf{h}_j(t)\epsilon_j(t)-\rho\sum_{j'\in\mathcal{N}_j}c_j(t)\big[\mathbf{e}_j(t-1)-\mathbf{e}_{j'}(t-1)\big].$$

Its matrix form (41) can be expressed as

$$E[\mathbf{e}^T(t)\mathbf{\Phi}(t)\mathbf{e}(t)]$$
$$=E[\mathbf{e}^T(t-1)(\mathbf{\Phi}(t-1)-\rho(\mathbf{C}(t)\mathbf{L})\otimes\mathbf{I}_p)^T\mathbf{\Phi}^{-1}(t)\times(\mathbf{\Phi}(t-1)-\rho(\mathbf{C}(t)\mathbf{L})\otimes\mathbf{I}_p)\mathbf{e}(t-1)]$$
$$+E[\boldsymbol{\epsilon}^T(t)\mathbf{C}^T(t)\mathbf{H}^T(t)\mathbf{\Phi}^{-1}(t)\mathbf{H}(t)\mathbf{C}(t)\boldsymbol{\epsilon}(t)]. \tag{55}$$

Observe that the right hand sides of (41) and (55) are only different in their first terms. Similar to Stage 3 (cf. (44)), we need to show that the first term satisfies

$$E[\mathbf{e}^T(t-1)(\mathbf{\Phi}(t-1)-\rho(\mathbf{C}(t)\mathbf{L})\otimes\mathbf{I}_p)^T\mathbf{\Phi}^{-1}(t)\times(\mathbf{\Phi}(t-1)-\rho(\mathbf{C}(t)\mathbf{L})\otimes\mathbf{I}_p)\mathbf{e}(t-1)]$$
$$\le E[\mathbf{e}^T(t-1)\mathbf{\Phi}(t-1)\mathbf{e}(t-1)]. \tag{56}$$

Substituting the update (22) with $\lambda=1$ into (56), it suffices to prove that

$$E[\mathbf{e}^T(t-1)\mathbf{C}(t)\otimes\mathbf{I}_p\mathbf{H}(t)\mathbf{H}^T(t)\times(\mathbf{I}_J+\mathbf{H}^T(t)\mathbf{\Phi}^{-1}(t-1)\mathbf{H}(t))^{-1}\otimes\mathbf{I}_p\mathbf{e}(t-1)]$$
$$\ge\rho E[\mathbf{e}^T(t-1)\mathbf{W}\mathbf{e}(t-1)] \tag{57}$$

where

$$\mathbf{W}:=\mathbf{W}_1+\mathbf{W}_1^T-\mathbf{W}_2-(\mathbf{L}\mathbf{C}(t))\otimes\mathbf{I}_p-(\mathbf{C}(t)\mathbf{L})\otimes\mathbf{I}_p+\rho\mathbf{L}\otimes\mathbf{I}_p\mathbf{\Phi}^{-1}(t-1)(\mathbf{C}(t)\mathbf{L})\otimes\mathbf{I}_p$$
$$\mathbf{W}_1:=\mathbf{C}(t)\otimes\mathbf{I}_p\mathbf{H}(t)\mathbf{H}^T(t)\mathbf{\Phi}^{-1}(t-1)\times((\mathbf{I}_J+\mathbf{H}^T(t)\mathbf{\Phi}^{-1}(t-1)\mathbf{H}(t))^{-1}\mathbf{L})\otimes\mathbf{I}_p$$
$$\mathbf{W}_2:=(\mathbf{L}\mathbf{C}(t))\otimes\mathbf{I}_p\mathbf{\Phi}^{-1}(t-1)\mathbf{H}(t)\mathbf{H}^T(t)\mathbf{\Phi}^{-1}(t-1)\times((\mathbf{I}_J+\mathbf{H}^T(t)\mathbf{\Phi}^{-1}(t-1)\mathbf{H}(t))^{-1}\mathbf{L})\otimes\mathbf{I}_p.$$

For the left hand side of (57), use the lower bound of the conditional expectation $2Q(\tau)\le E[c_j(t)|\mathbf{h}_j(t),\mathbf{s}_j(t-1)]$ to eliminate $\mathbf{C}(t)$, and arrive at

$$E[\mathbf{e}^T(t-1)\mathbf{C}(t)\otimes\mathbf{I}_p\mathbf{H}(t)\mathbf{H}^T(t)\times(\mathbf{I}_J+\mathbf{H}^T(t)\mathbf{\Phi}^{-1}(t-1)\mathbf{H}(t))^{-1}\otimes\mathbf{I}_p\mathbf{e}(t-1)]$$
$$\ge 2Q(\tau)E[\mathbf{e}^T(t-1)\mathbf{H}(t)\mathbf{H}^T(t)\times(\mathbf{I}_J+\mathbf{H}^T(t)\mathbf{\Phi}^{-1}(t-1)\mathbf{H}(t))^{-1}\otimes\mathbf{I}_p\mathbf{e}(t-1)]. \tag{58}$$

By (43), it holds that $\mathbf{\Phi}^{-1}(t-1)\preceq\mathbf{\Phi}^{-1}(0)=\gamma\mathbf{I}_{Jp}$, and thus

$$\big[\mathbf{I}_J+\mathbf{H}^T(t)\mathbf{\Phi}^{-1}(t-1)\mathbf{H}(t)\big]^{-1}\succeq\big[\mathbf{I}_J+\gamma\mathbf{H}^T(t)\mathbf{H}(t)\big]^{-1}.$$

By assumption $\{\mathbf{h}_j(t)\}$ are uniformly bounded. If $\mathbf{h}_j^T(t)\mathbf{h}_j(t)\le K$ for all $j=1,\ldots,J$, we find

$$\big[\mathbf{I}_J+\mathbf{H}^T(t)\mathbf{\Phi}^{-1}(t-1)\mathbf{H}(t)\big]^{-1}\succeq\frac{1}{1+\gamma K^2}\mathbf{I}_J. \tag{59}$$

Substituting (59) into (58), we obtain a lower bound for the left hand side of (57) given by

$$E[\mathbf{e}^T(t-1)\mathbf{C}(t)\otimes\mathbf{I}_p\mathbf{H}(t)\mathbf{H}^T(t)\times(\mathbf{I}_J+\mathbf{H}^T(t)\mathbf{\Phi}^{-1}(t-1)\mathbf{H}(t))^{-1}\otimes\mathbf{I}_p\mathbf{e}(t-1)]$$
$$\ge\frac{2Q(\tau)}{1+\gamma K^2}E[\mathbf{e}^T(t-1)\mathbf{H}(t)\mathbf{H}^T(t)\mathbf{e}(t-1)]$$
$$=\frac{2Q(\tau)}{1+\gamma K^2}E[\mathbf{e}^T(t-1)\mathrm{diag}\{\mathbf{R}_{h_j}\}\mathbf{e}(t-1)]$$
$$\ge\frac{2Q(\tau)\mu}{1+\gamma K^2}E[||\mathbf{e}(t-1)||^2]. \tag{60}$$

As for the right hand side of (57), it is upper bounded by

$$\begin{aligned}&\rho E[\mathbf{e}^T(t-1)\mathbf{W}\mathbf{e}(t-1)]\\&\le \rho E[(2||\mathbf{W}_1||_2+||\mathbf{W}_2||_2+2||\mathbf{L}||_2\\&\quad+\ \rho||\mathbf{L}||_2^2||\mathbf{\Phi}^{-1}(t-1)||_2)||\mathbf{e}(t-1)||^2] \qquad (61)\end{aligned}$$

where we used that all the diagonal elements $c_j(t)$ of $\mathbf{C}(t)$ are within the range $[0,1]$ while $||\mathbf{W}_1||_2$ is upper bounded by

$$\begin{aligned}||\mathbf{W}_1||_2 &\le ||\mathbf{C}(t)||_2||\mathbf{H}(t)||_2^2||\mathbf{\Phi}^{-1}(t-1)||_2\\&\quad\times\ ||(\mathbf{I}_J+\mathbf{H}^T(t)\mathbf{\Phi}^{-1}(t-1)\mathbf{H}(t))^{-1}||_2||\mathbf{L}||_2.\end{aligned}$$

Noticing that $||\mathbf{C}(t)||_2\le 1$, $||\mathbf{H}(t)||_2^2\le K^2$ by assumption, $||\mathbf{\Phi}^{-1}(t-1)||_2\le||\mathbf{\Phi}^{-1}(0)||_2=\gamma$, $||(\mathbf{I}_J+\mathbf{H}^T(t)\mathbf{\Phi}^{-1}(t-1)\mathbf{H}(t))^{-1}||_2\le 1$ and $||\mathbf{L}||_2\le\lambda_{\max}(\mathbf{L})$, we find that

$$||\mathbf{W}_1||_2\le\gamma\lambda_{\max}(\mathbf{L})K^2.$$

Similarly, $||\mathbf{W}_2||_2$ is upper bounded by

$$||\mathbf{W}_2||_2\le\gamma^2\lambda_{\max}(\mathbf{L})^2K^2.$$

Therefore, (61) reduces to

$$\begin{aligned}&\rho E[\mathbf{e}^T(t-1)\mathbf{W}\mathbf{e}(t-1)]\\&\le\rho(2\gamma\lambda_{\max}(\mathbf{L})K^2+\gamma^2\lambda_{\max}(\mathbf{L})^2K^2+2\lambda_{\max}(\mathbf{L})\\&\quad+\ \rho\gamma\lambda_{\max}(\mathbf{L})^2)E[||\mathbf{e}(t-1)||^2]. \qquad (62)\end{aligned}$$

Considering a positive constant

$$\begin{aligned}\rho_0 :=&\sqrt{\frac{2Q(\tau)\mu}{\gamma\lambda_{\max}(\mathbf{L})^2(1+\gamma K^2)}+(\frac{\gamma K^2}{2}+\frac{\gamma K^2+1}{\gamma\lambda_{\max}(\mathbf{L})})^2}\\&-\left(\frac{\gamma K^2}{2}+\frac{\gamma K^2+1}{\gamma\lambda_{\max}(\mathbf{L})}\right)\end{aligned}$$

and combining (60) with (62), we see that if ρ is chosen within $[0,\rho_0]$, then (57) holds for all $t\ge 1$; and so does (56).

Following Stages 4 and 5 in the proof for CD-RLS-1, we can show that (24) holds almost surely for CD-RLS-2 $\forall t>t_0$. This completes the proof of the entire theorem. ■

REFERENCES

[1] S. I. Amari., "Natural gradient works efficiently in learning," *Neural Comput.*, vol. 10, pp. 251–276, 1998.

[2] S. Appadwedula, V. V. Veeravalli, and D. L. Jones, "Decentralized detection with censoring sensors," *IEEE Trans. Signal Process.*, vol. 56, no. 4, pp. 1362–1373, Apr. 2008.

[3] R. Arroyo-Valles, S. Maleki, and G. Leus, "A censoring strategy for decentralized estimation in energy-constrained adaptive diffusion networks," in *Proc. Int. Workshop Signal Process. Adv. Wireless Commun.*, Darmstadt, Germany, Jun. 2013, pp. 155–159.

[4] D. Berberidis, V. Kekatos, and G. B. Giannakis, "Online censoring for large-scale regressions with application to streaming big data," *IEEE Trans. Signal Process.*, vol. 64, no. 15, pp. 3854–3867, Aug. 2016.

[5] P. Bianchi, G. Fort, and W. Hachem, "Performance of a distributed stochastic approximation algorithm," *IEEE Trans. Inf. Theory*, vol. 59, no. 11, pp. 7405–7418, Nov. 2013.

[6] V. Cevher, S. Becker, and M. Schmidt, "Convex optimization for big data: Scalable, randomized, and parallel algorithms for big data analytics," *IEEE Signal Process. Mag.*, vol. 31, no. 5, pp. 32–43, Sep. 2014.

[7] G. Grimmett and D. Stirzaker, *Probability and Random Processes*. London, U.K.: Oxford Univ. Press, 2011.

[8] G. B. Giannakis, Q. Ling, G. Mateos, I. D. Schizas, and H. Zhu, "Decentralized learning for wireless communications and networking," in *Splitting Methods in Communication and Imaging, Science and Engineering*, R. Glowinski, S. Osher, and W. Yin, Eds. New York, NY, USA: Springer, 2016.

[9] R. Jiang, Y. Lin, B. Chen, and B. Suter, "Distributed sensor censoring for detection in sensor networks under communication constraints," in *Proc. Conf. Rec. 39th Asilomar Conf. Signals, Syst. Comput.*, Pacific Grove, CA, USA, Nov. 2005, pp. 946–950.

[10] R. Jiang and B. Chen, "Fusion of censored decisions in wireless sensor networks," *IEEE Trans. Wireless Commun.*, vol. 4, no. 6, pp. 2668–2673, Nov. 2005.

[11] H. J. Kushner and G. G. Yin, *Stochastic Approximation Algorithms and Applications*. New York, NY, USA: Springer, 1997.

[12] M. Lichman, "UCI machine learning repository," 2013. [Online]. Available: http://archive.ics.uci.edu/ml

[13] Z. Liu, C. Li, and Y. Liu, "Distributed censored regression over networks," *IEEE Trans. Signal Process.*, vol. 63, no. 20, pp. 5437–5449, Oct. 2015.

[14] C. G. Lopes and A. H. Sayed, "Diffusion least-mean squares over adaptive networks: Formulation and performance analysis," *IEEE Trans. Signal Process.*, vol. 56, no. 7, pp. 3122–3136, Jul. 2008.

[15] G. Mateos and G. B. Giannakis, "Distributed recursive least-squares: Stability and performance analysis," *IEEE Trans. Signal Process.*, vol. 60, no. 7, pp. 3740–3754, Jul. 2012.

[16] G. Mateos, I. D. Schizas, and G. B. Giannakis, "Performance analysis of the consensus-based distributed LMS algorithm," *EURASIP J. Adv. Signal Process.*, vol. 2009, 2009, Art. no. 981030.

[17] G. Mateos, I. D. Schizas, and G. B. Giannakis, "Distributed recursive least-squares for consensus-based in-network adaptive estimation," *IEEE Trans. Signal Process.*, vol. 57, no. 11, pp. 4583–4588, Nov. 2009.

[18] G. Morral, P. Bianchi, and G. Fort, "Success and failure of adaptation-diffusion algorithms with decaying step size in multiagent networks," *IEEE Trans. Signal Process.*, vol. 65, no. 11, pp. 2798–2813, Jun. 2017.

[19] E. Msechu and G. B. Giannakis, "Sensor-centric data reduction for estimation with WSNs via censoring and quantization," *IEEE Trans. Signal Process.*, vol. 60, no. 1, pp. 400–414, Jan. 2012.

[20] N. Patwari and A. O. Hero, "Hierarchical censoring for distributed detection in wireless sensor networks," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Hong Kong, 2003, pp. IV-848–IV-851.

[21] J. Predd, S. Kulkarni, and H. V. Poor, "Distributed learning in wireless sensor networks," *IEEE Signal Process. Mag.*, vol. 23, no. 4, pp. 56–69, Jul. 2006.

[22] M. Rabbat and R. Nowak, "Distributed optimization in sensor networks," in *Proc. Int. Conf. Inf. Process. Sensor Netw.*, Berkeley, CA, Apr. 2004, pp. 20–27.

[23] C. Rago, P. Willett, and Y. Bar-Shalom, "Censoring sensors: A low-communication-rate scheme for distributed detection," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 32, no. 2, pp. 554–568, Apr. 1996.

[24] M. A. Sharkh, M. Jammal, A. Shami, and A. Ouda, "Resource allocation in a network-based cloud computing environment: Design challenges," *IEEE Commun. Mag.*, vol. 51, no. 11, pp. 46–52, Nov. 2013.

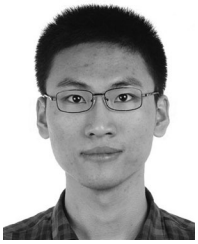
[25] K. Slavakis, S. J. Kim, G. Mateos, and G. B. Giannakis, "Stochastic approximation vis-a-vis online learning for big data analytics," *IEEE Signal Process. Mag.*, vol. 31, no. 6, pp. 124–129, Nov. 2014.

[26] V. Solo and X. Kong, *Adaptive Signal Processing Algorithms: Stability and Performance*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1995.

[27] P. Tseng, "Applications of a splitting algorithm to decomposition in convex programming and variational inequalities," *SIAM J. Control Optim.*, vol. 29, pp. 119–138, Jan. 1991.

[28] Z. Wang, Z. Yu, Q. Ling, D. Berberidis, and G. B. Giannakis, "Distributed recursive least-squares with data-adaptive censoring," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, New Orleans, LA, USA, Mar. 2017, pp. 5860–5864.

Zifeng Wang received the Bachelor's degree in computational mathematics from the Special Class for the Gifted Young, University of Science and Technology of China, Hefei, China, in 2018. He is currently working toward the Ph.D. degree in applied mathematics at the University of California, Los Angeles, Los Angeles, CA, USA. His research interests include optimization and machine learning.



Zheng Yu received the Bachelor's degree in computational mathematics from the Special Class for the Gifted Young, University of Science and Technology of China, Hefei, China, in 2017. He is currently working toward the Ph.D. degree at the Department of Operation Research and Financial Engineering, Princeton University, Princeton, NJ, USA. His research interests include stochastic optimization and machine learning algorithms.

Qing Ling (SM'15) received the B.E. degree in automation and the Ph.D. degree in control theory and control engineering from the University of Science and Technology of China, Hefei, China, in 2001 and 2006, respectively. He was a Postdoctoral Research Fellow with the Department of Electrical and Computer Engineering, Michigan Technological University, Houghton, MI, USA, from 2006 to 2009, and an Associate Professor with the Department of Automation, University of Science and Technology of China, from 2009 to 2017. He is currently a Professor with the School of Data and Computer Science, Sun Yat-Sen University, Guangzhou, China. His research interests include decentralized network optimization and its applications. He is an Associate Editor for the IEEE TRANSACTIONS ON NETWORK AND SERVICE MANAGEMENT and the IEEE SIGNAL PROCESSING LETTERS. He was the recipient of the 2017 IEEE Signal Processing Society Young Author Best Paper Award as a Supervisor, and the 2017 International Consortium of Chinese Mathematicians Distinguished Paper Award.

Dimitris Berberidis (S'15) received the Diploma in electrical and computer engineering (ECE) from the University of Patras, Patras, Greece, in 2012, and the M.Sc. degree in ECE from the University of Minnesota, Minneapolis, MN, USA, where he is currently working toward the Ph.D. degree. His research interests include statistical signal processing, sketching and tracking of large-scale processes as well as active and semisupervised learning over graphs.

Georgios B. Giannakis (F'97) received the Diploma in electrical engineering from the National Technical University of Athens, Athens, Greece, in 1981. From 1982 to 1986, he was with the University of Southern California, Los Angeles, CA, USA, where he received the M.Sc. degree in electrical engineering in 1983, the M.Sc. degree in mathematics in 1986, and the Ph.D. degree in electrical engineering in 1986.

He was with the University of Virginia from 1987 to 1998, and since 1999, he has been a Professor with the University of Minnesota, Minneapolis, MN, USA, where he is currently the Endowed Chair in wireless telecommunications, the University of Minnesota McKnight Presidential Chair in electrical and computer engineering, and the Director of the Digital Technology Center. His research interests include communications, networking, and statistical learning—subjects on which he authored or coauthored more than 400 journal papers, 700 conference papers, 25 book chapters, 2 edited books, and 2 research monographs (H-index 128). His current research focuses on learning from Big Data, wireless cognitive radios, and network science with applications to social, brain, and power networks with renewables. He is the (co-) inventor of 32 patents issued, and the (co-) recipient of 9 best journal paper awards from the IEEE Signal Processing (SP) and Communications Societies, including the G. Marconi Prize Paper Award in Wireless Communications. He was also the recipient of Technical Achievement Awards from the SP Society (2000), from EURASIP (2005), the Young Faculty Teaching Award, the G. W. Taylor Award for Distinguished Research from the University of Minnesota, and the IEEE Fourier Technical Field Award (inaugural recipient in 2015). He is a Fellow of EURASIP, and has served the IEEE in a number of posts, including that of a Distinguished Lecturer for the IEEE-SP Society.