Journal of Machine Learning Research 17 (2016) 1-35 Submitted 6/15; Revised 2/16; Published 3/16

DSA: Decentralized Double Stochastic Averaging Gradient Algorithm

Aryan Mokhtari ARYANM@SEAS.UPENN.EDU
Alejandro Ribeiro ARIBEIRO@SEAS.UPENN.EDU
Department of Electrical and Systems Engineering
University of Pennsylvania
Philadelphia, PA 19104, USA

Editor: Mark Schmidt

Abstract

This paper considers optimization problems where nodes of a network have access to summands of a global objective. Each of these local objectives is further assumed to be an average of a finite set of functions. The motivation for this setup is to solve large scale machine learning problems where elements of the training set are distributed to multiple computational elements. The decentralized double stochastic averaging gradient (DSA) algorithm is proposed as a solution alternative that relies on: (i) The use of local stochastic averaging gradients. (ii) Determination of descent steps as differences of consecutive stochastic averaging gradients. Strong convexity of local functions and Lipschitz continuity of local gradients is shown to guarantee linear convergence of the sequence generated by DSA in expectation. Local iterates are further shown to approach the optimal argument for almost all realizations. The expected linear convergence of DSA is in contrast to the sublinear rate characteristic of existing methods for decentralized stochastic optimization. Numerical experiments on a logistic regression problem illustrate reductions in convergence time and number of feature vectors processed until convergence relative to these other alternatives.

Keywords: decentralized optimization, stochastic optimization, stochastic averaging gradient, linear convergence, large-scale optimization, logistic regression

1. Introduction

We consider machine learning problems with large training sets that are distributed into a network of computing agents so that each of the nodes maintains a moderate number of samples. This leads to decentralized consensus optimization problems where summands of the global objective function are available at different nodes of the network. In this class of problems agents (nodes) try to optimize the global cost function by operating on their local functions and communicating with their neighbors only. Specifically, consider a variable $\mathbf{x} \in \mathbb{R}^p$ and a connected network of size N where each node n has access to a local objective function $f_n : \mathbb{R}^p \to \mathbb{R}$. The local objective function $f_n(\mathbf{x})$ is defined as the average of q_n local instantaneous functions $f_{n,i}(\mathbf{x})$ that can be individually evaluated at node n. Agents cooperate to solve the global optimization problem

$$\tilde{\mathbf{x}}^* := \operatorname*{argmin}_{\mathbf{x}} \sum_{n=1}^{N} f_n(\mathbf{x}) = \operatorname*{argmin}_{\mathbf{x}} \sum_{n=1}^{N} \frac{1}{q_n} \sum_{i=1}^{q_n} f_{n,i}(\mathbf{x}). \tag{1}$$

The formulation in (1) models a training set with a total of $\sum_{n=1}^{N} q_n$ training samples that are distributed among the N agents for parallel processing conducive to the determination of the optimal classifier $\tilde{\mathbf{x}}^*$ (Bekkerman et al. (2011); Tsianos et al. (2012a); Cevher et al. (2014)). Although we make no formal assumption, in cases of practical importance the total number of training samples $\sum_{n=1}^{N} q_n$ is very large, but the number of elements q_n available at a specific node is moderate.

Analogous formulations are also of interest in decentralized control systems (Bullo et al. (2009); Cao et al. (2013); Lopes and Sayed (2008)), wireless systems (Ribeiro (2010, 2012)), and sensor networks (Schizas et al. (2008); Khan et al. (2010); Rabbat and Nowak (2004)). Our interest here is in solving (1) with a method that has the following three properties

- Decentralized; nodes operate on their local functions and communicate with neighbors only.
- Stochastic; nodes determine a descent direction by evaluating only one out of the q_n functions $f_{n,i}$ at each iteration.
- Linear convergence rate; the expected distance to the optimum is scaled by a subunit factor at each iteration.

Decentralized optimization is relatively mature and various methods are known with complementary advantages. These methods include decentralized gradient descent (DGD) (Nedić and Ozdaglar (2009); Jakovetic et al. (2014); Yuan et al. (2013)), network Newton (Mokhtari et al. (2015a,b)), decentralized dual averaging (Duchi et al. (2012); Tsianos et al. (2012b)), the exact first order algorithm (EXTRA) (Shi et al. (2015)), as well as the alternating direction method of multipliers (ADMM) (Boyd et al. (2011); Schizas et al. (2008); Shi et al. (2014); Iutzeler et al. (2013)) and its linearized variants (Ling and Ribeiro (2014); Ling et al. (2015); Mokhtari et al. (2015c)). The ADMM, its variants, and EXTRA converge linearly to the optimal argument but DGD, network Newton, and decentralized dual averaging have sublinear convergence rates. Of particular importance to this paper, is the fact that DGD has (inexact) linear convergence to a neighborhood of the optimal argument when it uses constant stepsizes. It can achieve exact convergence by using diminishing stepsizes, but the convergence rate degrades to sublinear. This lack of linear convergence is solved by EXTRA through the use of iterations that rely on information of two consecutive steps (Shi et al. (2015)).

All of the algorithms mentioned above require the computationally costly evaluation of the local gradients $\nabla f_n(\mathbf{x}) = (1/q_n)\sum_{i=1}^{q_n} \nabla f_{n,i}(\mathbf{x})$. This cost can be avoided by stochastic decentralized algorithms that reduce computational cost of iterations by substituting all local gradients with their stochastic approximations. This reduces the computational cost per iteration but results in sublinear convergence rates of order $O(1/t)$ even if the corresponding deterministic algorithm exhibits linear convergence. This is a drawback that also exists in centralized stochastic optimization where linear convergence rates in expectation are established by decreasing the variance of the stochastic gradient approximation (Roux et al. (2012); Schmidt et al. (2013); Shalev-Shwartz and Zhang (2013); Johnson and Zhang (2013); Konečnỳ and Richtárik (2013); Defazio et al. (2014)). In this paper we build on the ideas of the stochastic averaging gradient (SAG) algorithm (Schmidt et al. (2013)) and its unbiased version SAGA (Defazio et al. (2014)). Both of these algorithms use the idea of stochastic incremental averaging gradients. At each iteration only one of the stochastic gradients is updated and the average of all of the most recent stochastic gradients is used for estimating the gradient.

The contribution of this paper is to develop the decentralized double stochastic averaging gradient (DSA) method, a novel decentralized stochastic algorithm for solving (1). The method exploits a new interpretation of EXTRA as a saddle point method and uses stochastic averaging gradients in lieu of gradients. DSA is *decentralized* because it is implementable in a network setting where nodes can communicate only with their neighbors. It is *double* because iterations utilize the information of two consecutive iterates. It is *stochastic* because the gradient of only one randomly selected function is evaluated at each iteration and it is an *averaging* method because it uses an average of stochastic gradients to approximate the local gradients. DSA is proven to converge linearly to the optimal argument $\tilde{\mathbf{x}}^*$ in expectation when the local instantaneous functions $f_{n,i}$ are strongly convex, with Lipschitz continuous gradients. This is in contrast to all other decentralized stochastic methods to solve (1) that converge at sublinear rates.

We begin the paper with a discussion of DGD, EXTRA and stochastic averaging gradient. With these definitions in place we define the DSA algorithm by replacing the gradients used in EXTRA

by stochastic averaging gradients (Section 2). We follow with a digression on the limit points of DGD and EXTRA iterations to explain the reason why DGD does not achieve exact convergence but EXTRA is expected to do so (Section 2.1). A reinterpretation of EXTRA as a saddle point method that solves for the critical points of the augmented Lagrangian of a constrained optimization problem equivalent to (1) is then introduced. It follows from this reinterpretation that DSA is a stochastic saddle point method (Section 2.2). The fact that DSA is a stochastic saddle point method is the critical enabler of the subsequent convergence analysis (Section 3). In particular, it is possible to guarantee that strong convexity and gradient Lipschitz continuity of the local instantaneous functions $f_{n,i}$ imply that a Lyapunov function associated with the sequence of iterates generated by DSA converges linearly to its optimal value in expectation (Theorem 7). Linear convergence in expectation of the local iterates to the optimal argument $\tilde{\mathbf{x}}^*$ of (1) follows as a trivial consequence (Corollary 8). We complement this result by showing convergence of all the local variables to the optimal argument $\tilde{\mathbf{x}}^*$ with probability 1 (Theorem 9).

The advantages of DSA relative to a group of stochastic and deterministic alternatives in solving a logistic regression problem are then studied in numerical experiments (Section 4). These results demonstrate that DSA is the only decentralized stochastic algorithm that reaches the optimal solution with a linear convergence rate. We further show that DSA outperforms deterministic algorithms when the metric is the number of times that elements of the training set are evaluated. The behavior of DSA for different network topologies is also evaluated. We close the paper with pertinent remarks (Section 5).

Notation Lowercase boldface $\mathbf{v}$ denotes a vector and uppercase boldface $\mathbf{A}$ a matrix. For column vectors $\mathbf{x}_1, \ldots, \mathbf{x}_N$ we use the notation $\mathbf{x} = [\mathbf{x}_1; \ldots; \mathbf{x}_N]$ to represent the stack column vector $\mathbf{x}$. We use $\|\mathbf{v}\|$ to denote the Euclidean norm of vector $\mathbf{v}$ and $\|\mathbf{A}\|$ to denote the Euclidean norm of matrix $\mathbf{A}$. For a vector $\mathbf{v}$ and a positive definite matrix $\mathbf{A}$, the $\mathbf{A}$-weighted norm is defined as $\|\mathbf{v}\|_{\mathbf{A}} := \sqrt{\mathbf{v}^T\mathbf{A}\mathbf{v}}$. The null space of matrix $\mathbf{A}$ is denoted by null($\mathbf{A}$) and the span of a vector by span($\mathbf{x}$). The operator $\mathbb{E}_{\mathbf{x}}[\cdot]$ stands for expectation over random variable $\mathbf{x}$ and $\mathbb{E}[\cdot]$ for expectation with respect to the distribution of a stochastic process.

2. Decentralized Double Stochastic Averaging Gradient

Consider a connected network that contains N nodes such that each node n can only communicate with peers in its neighborhood $\mathcal{N}_n$. Define $\mathbf{x}_n \in \mathbb{R}^p$ as a local copy of the variable $\mathbf{x}$ that is kept at node n. In decentralized optimization, agents try to minimize their local functions $f_n(\mathbf{x}_n)$ while ensuring that their local variables $\mathbf{x}_n$ coincide with the variables $\mathbf{x}_m$ of all neighbors $m \in \mathcal{N}_n$ – which, given that the network is connected, ensures that the variables $\mathbf{x}_n$ of all nodes are the same and renders the problem equivalent to (1). DGD is a well known method for decentralized optimization that relies on the introduction of nonnegative weights $w_{ij} \geq 0$ that are not null if and only if $m = n$ or if $m \in \mathcal{N}_n$. Letting $t \in \mathbb{N}$ be a discrete time index and α a given stepsize, DGD is defined by the recursion

$$\mathbf{x}_n^{t+1} = \sum_{m=1}^{N} w_{nm}\mathbf{x}_m^t - \alpha\nabla f_n(\mathbf{x}_n^t), \qquad n = 1, \ldots, N. \tag{2}$$

Since $w_{nm} = 0$ when $m \neq n$ and $m \notin \mathcal{N}_n$, it follows from (2) that node n updates $\mathbf{x}_n$ by performing an average over the variables $\mathbf{x}_m^t$ of its neighbors $m \in \mathcal{N}_n$ and its own $\mathbf{x}_n^t$, followed by descent through the negative local gradient $-\nabla f_n(\mathbf{x}_n^t)$. If a constant stepsize is used, DGD iterates $\mathbf{x}_n^t$ approach a neighborhood of the optimal argument $\tilde{\mathbf{x}}^*$ of (1) but don't converge exactly. To achieve exact convergence diminishing stepsizes are used but the resulting convergence rate is sublinear (Nedić and Ozdaglar (2009)).

EXTRA is a method that resolves either of these issues by mixing two consecutive DGD iterations with different weight matrices and opposite signs. To be precise, introduce a second set of weights

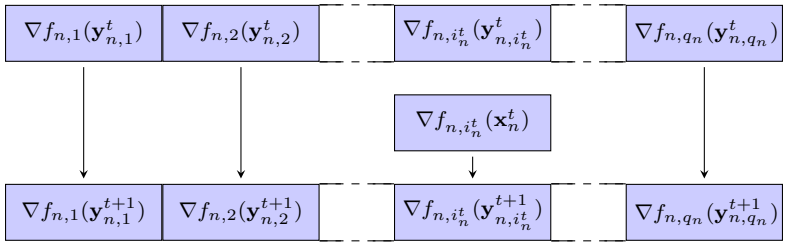


Figure 1: Stochastic averaging gradient table at node n. At each iteration t a random local instantaneous gradient $\nabla f_{n,i_n^t}(\mathbf{y}_{n,i_n^t}^t)$ is updated by $\nabla f_{n,i_n^t}(\mathbf{x}_n^t)$. The rest of the local instantaneous gradients remain unchanged, i.e., $\nabla f_{n,i}(\mathbf{y}_{n,i}^{t+1}) = \nabla f_{n,i}(\mathbf{y}_{n,i}^t)$ for $i \neq i_n^t$. This list is used to compute the stochastic averaging gradient in (7).

$\tilde{w}_{nm}$ with the same properties as the weights w_{nm} and define EXTRA through the recursion

$$\mathbf{x}_n^{t+1} = \mathbf{x}_n^t + \sum_{m=1}^N w_{nm}\mathbf{x}_m^t - \sum_{m=1}^N \tilde{w}_{nm}\mathbf{x}_m^{t-1} - \alpha\left[\nabla f_n(\mathbf{x}_n^t) - \nabla f_n(\mathbf{x}_n^{t-1})\right], \quad n = 1, \ldots, N. \tag{3}$$

Observe that (3) is well defined for $t > 0$. For $t = 0$ we utilize the regular DGD iteration in (2). In the nomenclature of this paper we say that EXTRA performs a decentralized double gradient descent step because it operates in a decentralized manner while utilizing a difference of two gradients as descent direction. Minor modification as it is, the use of this gradient difference in lieu of simple gradients, endows EXTRA with exact linear convergence to the optimal argument $\tilde{\mathbf{x}}^*$ under mild assumptions (Shi et al. (2015)).

If we recall the definitions of the local functions $f_n(\mathbf{x}_n)$ and the instantaneous local functions $f_{n,i}(\mathbf{x}_n)$ available at node n, the implementation of EXTRA requires that each node n computes the full gradient of its local objective function f_n at $\mathbf{x}_n^t$ as

$$\nabla f_n(\mathbf{x}_n^t) = \frac{1}{q_n}\sum_{i=1}^{q_n} \nabla f_{n,i}(\mathbf{x}_n^t). \tag{4}$$

This is computationally expensive when the number of instantaneous functions q_n is large. To resolve this issue, local stochastic gradients can be substituted for the local objective functions gradients in (3). These stochastic gradients approximate the gradient $\nabla f_n(\mathbf{x}_n)$ of node n by randomly choosing one of the instantaneous functions gradients $\nabla f_{n,i}(\mathbf{x}_n)$. If we let $i_n^t \in \{1, \ldots q_n\}$ denote a function index that we choose at time t at node n uniformly at random and independently of the history of the process, then the stochastic gradient is defined as

$$\hat{\mathbf{s}}_n(\mathbf{x}_n^t) := \nabla f_{n,i_n^t}(\mathbf{x}_n^t). \tag{5}$$

We can then write a stochastic version of EXTRA by replacing $\nabla f_n(\mathbf{x}_n^t)$ by $\hat{\mathbf{s}}_n(\mathbf{x}_n^t)$ and $\nabla f_n(\mathbf{x}_n^{t-1})$ by $\hat{\mathbf{s}}_n(\mathbf{x}_n^{t-1})$. Such an algorithm would have a small computational cost per iteration. On the negative side, it either has a linear convergence to a neighborhood of the optimal solution $\mathbf{x}^*$ with constant stepsize α, or it would converge sublinearly to the optimal argument when the stepisize diminishes as time passes. Here however, we want to design an algorithm with low computational complexity that converges linearly to the exact solution $\mathbf{x}^*$.

To reduce this noise we propose the use of stochastic averaging gradients instead (Defazio et al. (2014)). The idea is to maintain a list of gradients of all instantaneous functions in which one randomly chosen element is replaced at each iteration and to use an average of the elements of this

list for gradient approximation; see Figure 1. Formally, define the variable $\mathbf{y}_{n,i} \in \mathbb{R}^p$ to represent the iterate value the last time that the instantaneous gradient of function $f_{n,i}$ was evaluated. If we let $i_n^t \in \{1, \ldots, q_n\}$ denote the function index chosen at time t at node n, as we did in (5), the variables $\mathbf{y}_{n,i}$ are updated recursively as

$$\mathbf{y}_{n,i}^{t+1} = \mathbf{x}_n^t, \quad \text{if } i = i_n^t, \qquad \mathbf{y}_{n,i}^{t+1} = \mathbf{y}_{n,i}^t, \quad \text{if } i \neq i_n^t. \tag{6}$$

With these definitions in hand we can define the stochastic averaging gradient at node n as

$$\hat{\mathbf{g}}_n^t := \nabla f_{n,i_n^t}(\mathbf{x}_n^t) - \nabla f_{n,i_n^t}(\mathbf{y}_{n,i_n^t}^t) + \frac{1}{q_n} \sum_{i=1}^{q_n} \nabla f_{n,i}(\mathbf{y}_{n,i}^t). \tag{7}$$

Observe that to implement (7) the gradients $\nabla f_{n,i}(\mathbf{y}_{n,i}^t)$ are stored in the local gradient table shown in Figure 1.

The DSA algorithm is a variation of EXTRA that substitutes the local gradients $\nabla f_n(\mathbf{x}_n^t)$ in (3) for the local stochastic average gradients $\hat{\mathbf{g}}_n^t$ in (7),

$$\mathbf{x}_n^{t+1} = \mathbf{x}_n^t + \sum_{m=1}^N w_{nm} \mathbf{x}_m^t - \sum_{m=1}^N \tilde{w}_{nm} \mathbf{x}_m^{t-1} - \alpha \left[\hat{\mathbf{g}}_n^t - \hat{\mathbf{g}}_n^{t-1} \right]. \tag{8}$$

The DSA initial update is given by applying the same substitution for the update of DGD in (2) as

$$\mathbf{x}_n^1 = \sum_{m=1}^N w_{nm} \mathbf{x}_m^0 - \alpha \, \hat{\mathbf{g}}_n^0. \tag{9}$$

DSA is summarized in Algorithm 1 for $t \geq 0$. The DSA update in (8) is implemented in Step 9. This step requires access to the local iterates $\mathbf{x}_m^t$ of neighboring nodes $m \in \mathcal{N}_n$ which are collected in Step 2. Furthermore, implementation of the DSA update also requires access to the stochastic averaging gradients $\hat{\mathbf{g}}_n^{t-1}$ and $\hat{\mathbf{g}}_n^t$. The latter is computed in Step 4 and the former is computed and stored at the same step in the previous iteration. The computation of the stochastic averaging gradients requires the selection of the index i_n^t. This index is chosen uniformly at random in Step 3. Determination of stochastic averaging gradients also necessitates access and maintenance of the gradients table in Figure 1. The i_n^t element of this table is updated in Step 5 by replacing $\nabla f_{n,i_n^t}(\mathbf{y}_{n,i_n^t}^t)$ with $\nabla f_{n,i_n^t}(\mathbf{x}_n^t)$, while the other vectors remain unchanged. To implement the first DSA iteration at time $t = 0$ we have to perform the update in (9) instead of the update in (8) as in Step 7. Further observe that the auxiliary variables $\mathbf{y}_{n,i}^0$ are initialized to the initial iterate $\mathbf{x}_n^0$. This implies that the initial values of the stored gradients are $\nabla f_{n,i}(\mathbf{y}_{n,i}^0) = \nabla f_{n,i}(\mathbf{x}_n^0)$.

We point out that the weights w_{nm} and $\tilde{w}_{nm}$ can't be arbitrary. If we define weight matrices $\mathbf{W}$ and $\tilde{\mathbf{W}}$ with elements w_{nm} and $\tilde{w}_{nm}$, respectively, they have to satisfy conditions that we state as an assumption for future reference.

Assumption 1 *The weight matrices $\mathbf{W}$ and $\tilde{\mathbf{W}}$ must satisfy the following properties*

(a) Both are symmetric, $\mathbf{W} = \mathbf{W}^T$ and $\tilde{\mathbf{W}} = \tilde{\mathbf{W}}^T$.

(b) The null space of $\mathbf{I} - \tilde{\mathbf{W}}$ includes the span of $\mathbf{1}$, i.e., null$(\mathbf{I} - \tilde{\mathbf{W}}) \supseteq$ span$(\mathbf{1})$, the null space of $\mathbf{I} - \mathbf{W}$ is the span of $\mathbf{1}$, i.e., null$(\mathbf{I} - \mathbf{W}) =$ span$(\mathbf{1})$, and the null space of the difference $\tilde{\mathbf{W}} - \mathbf{W}$ is the span of $\mathbf{1}$, i.e., null$(\tilde{\mathbf{W}} - \mathbf{W}) =$ span$(\mathbf{1})$.

(c) They satisfy the spectral ordering $\mathbf{W} \preceq \tilde{\mathbf{W}} \preceq (\mathbf{I} + \mathbf{W})/2$ and $\mathbf{0} \prec \tilde{\mathbf{W}}$.

Algorithm 1 DSA algorithm at node n

Require: Vectors $\mathbf{x}_n^0$. Gradient table initialized with instantaneous gradients $\nabla f_{n,i}(\mathbf{y}_{n,i}^0)$ with $\mathbf{y}_{n,i}^0 = \mathbf{x}_n^0$.

1: **for** $t = 0, 1, 2, \ldots$ **do**

2: Exchange variable $\mathbf{x}_n^t$ with neighboring nodes $m \in \mathcal{N}_n$;

3: Choose i_n^t uniformly at random from the set $\{1, \ldots, q_n\}$;

4: Compute and store stochastic averaging gradient as per (7):

$$\hat{\mathbf{g}}_n^t = \nabla f_{n,i_n^t}(\mathbf{x}_n^t) - \nabla f_{n,i_n^t}(\mathbf{y}_{n,i_n^t}^t) + \frac{1}{q_n}\sum_{i=1}^{q_n} \nabla f_{n,i}(\mathbf{y}_{n,i}^t);$$

5: Take $\mathbf{y}_{n,i_n^t}^{t+1} = \mathbf{x}_n^t$ and store $\nabla f_{n,i_n^t}(\mathbf{y}_{n,i_n^t}^{t+1}) = \nabla f_{n,i_n^t}(\mathbf{x}_n^t)$ in i_n^t gradient table position. All other entries in the table remain unchanged. The vector $\mathbf{y}_{n,i_n^t}^{t+1}$ is not explicitly stored;

6: **if** $t = 0$ **then**

7: Update variable $\mathbf{x}_n^t$ as per (9): $\quad \mathbf{x}_n^{t+1} = \sum_{m=1}^{N} w_{nm}\mathbf{x}_n^t - \alpha\hat{\mathbf{g}}_n^t;$

8: **else**

9: Update variable $\mathbf{x}_n^t$ as per (8): $\quad \mathbf{x}_n^{t+1} = \mathbf{x}_n^t + \sum_{m=1}^{N} w_{nm}\mathbf{x}_n^t - \sum_{m=1}^{N} \tilde{w}_{nm}\mathbf{x}_n^{t-1} - \alpha\left[\hat{\mathbf{g}}_n^t - \hat{\mathbf{g}}_n^{t-1}\right];$

10: **end if**

11: **end for**

Requiring the matrix $\mathbf{W}$ to be symmetric and with specific null space properties is necessary to let all agents converge to the same optimal variable. Analogous properties are necessary in DGD and are not difficult to satisfy. The condition on spectral ordering is specific to EXTRA but is not difficult to satisfy either. E.g., if we have a matrix $\mathbf{W}$ that satisfies all the conditions in Assumption 1, the weight matrix $\tilde{\mathbf{W}} = (\mathbf{I} + \mathbf{W})/2$ makes Assumption 1 valid.

We also point that, as written in (7), computation of local stochastic averaging gradients $\hat{\mathbf{g}}_n^t$ is costly because it requires evaluation of the sum $\sum_{i=1}^{q_n} \nabla f_{n,i}(\mathbf{y}_{n,i}^t)$ at each iteration. To be more precise, if we implement the update in (7) naively, at each iteration we should compute the sum $\sum_{i=1}^{q_n} \nabla f_{n,i}(\mathbf{y}_{n,i}^t)$ which has a computational cost of the order $O(q_n)$. This cost can be avoided by updating the sum at each iteration with the recursive formula

$$\sum_{i=1}^{q_n} \nabla f_{n,i}(\mathbf{y}_{n,i}^t) \;=\; \sum_{i=1}^{q_n} \nabla f_{n,i}(\mathbf{y}_{n,i}^{t-1}) + \nabla f_{n,i_n^{t-1}}(\mathbf{x}_n^{t-1}) - \nabla f_{n,i_n^{t-1}}(\mathbf{y}_{n,i_n^{t-1}}^{t-1}). \tag{10}$$

Using the update in (10), we can update the sum $\sum_{i=1}^{q_n} \nabla f_{n,i}(\mathbf{y}_{n,i}^t)$ required for (7) in a computationally efficient manner. Important properties and interpretations of EXTRA and DSA are presented in the following sections after pertinent remarks.

Remark 1 The local stochastic averaging gradients in (7) are unbiased estimates of the local gradients $\nabla f_n(\mathbf{x}_n^t)$. Indeed, if we let $\mathcal{F}_t$ measure the history of the system up until time t we have that the sum in (7) is deterministic given this sigma-algebra. This observation implies that the conditional expectation $\mathbb{E}\left[(1/q_n)\sum_{i=1}^{q_n} \nabla f_{n,i}(\mathbf{y}_{n,i}^t) \mid \mathcal{F}^t\right]$ can be simplified as $(1/q_n)\sum_{i=1}^{q_n} \nabla f_{n,i}(\mathbf{y}_{n,i}^t)$. Thus, the conditional expectation of the stochastic averaging gradient is,

$$\mathbb{E}\left[\hat{\mathbf{g}}_n^t \,\middle|\, \mathcal{F}^t\right] = \mathbb{E}\Big[\nabla f_{n,i_n^t}(\mathbf{x}_n^t) \,\Big|\, \mathcal{F}^t\Big] - \mathbb{E}\Big[\nabla f_{n,i_n^t}(\mathbf{y}_{n,i_n^t}^t) \,\Big|\, \mathcal{F}^t\Big] + \frac{1}{q_n}\sum_{i=1}^{q_n} \nabla f_{n,i}(\mathbf{y}_{n,i}^t). \tag{11}$$

With the index i_n^t chosen equiprobably from the set $\{1, \ldots, q_n\}$, the expectation of the second term in (11) is the same as the sum in the last term – each of the indexes is chosen with probability $1/q_n$. In other words, we can write $\mathbb{E}\Big[\nabla f_{n,i_n^t}(\mathbf{y}_{n,i_n^t}^t) \,\Big|\, \mathcal{F}^t\Big] = (1/q_n)\sum_{i=1}^{q_n} \nabla f_{n,i}(\mathbf{y}_{n,i}^t)$. Therefore,

these two terms cancel out each other and, since the expectation of the first term in (11) is simply $\mathbb{E}\big[\nabla f_{n,i_n^t}(\mathbf{x}_n^t)\,\big|\,\mathcal{F}^t\big] = (1/q_n)\sum_{i=1}^{q_n}\nabla f_{n,i}(\mathbf{x}_n^t) = \nabla f_n(\mathbf{x}_n^t)$, we can simplify (11) to

$$\mathbb{E}\left[\hat{\mathbf{g}}_n^t\,\middle|\,\mathcal{F}^t\right] = \nabla f_n(\mathbf{x}_n^t). \tag{12}$$

The expression in (12) means, by definition, that $\hat{\mathbf{g}}_n^t$ is an unbiased estimate of $\nabla f_n(\mathbf{x}_n^t)$ when the history $\mathcal{F}^t$ is given.

Remark 2 The local stochastic averaging gradient $\hat{\mathbf{g}}_n^t$ at node n contains three terms. The first two terms $\nabla f_{n,i_n^t}(\mathbf{x}_n^t)$ and $\nabla f_{n,i_n^t}(\mathbf{y}_{n,i_n^t}^t)$ are the new and old gradients of the chosen objective function f_{n,i_n^t} at node n, respectively. The last term $(1/q_n)\sum_{i=1}^{q_n}\nabla f_{n,i}(\mathbf{y}_{n,i}^t)$ is the average of the average of all the instantaneous gradients available at node n. This update can be considered as a localized version of the stochastic averaging gradient update in the SAGA algorithm (Defazio et al. (2014)). Notice that instead of the difference $\nabla f_{n,i_n^t}(\mathbf{x}_n^t) - \nabla f_{n,i_n^t}(\mathbf{y}_{n,i_n^t}^t)$ in (7) we could use the difference $(\nabla f_{n,i_n^t}(\mathbf{x}_n^t) - \nabla f_{n,i_n^t}(\mathbf{y}_{n,i_n^t}^t))/q_n$ which would lead to stochastic averaging gradient suggested in the SAG algorithm (Schmidt et al. (2013)). As studied in (Defazio et al. (2014)), both of these approximations lead to a variance reduction method. The one suggested by SAGA is an unbiased estimator of the exact gradient $(1/q_n)\sum_{i=1}^{q_n}\nabla f_{n,i}(\mathbf{x}_n^t)$, while the one suggested by SAG is a biased estimator of the gradient with smaller variance. Since the analysis of the estimator suggested by SAGA is simpler, we use its idea to define the local stochastic averaging gradient $\hat{\mathbf{g}}_n^t$ in (7).

2.1 Limit Points of DGD and EXTRA

The derivation of EXTRA hinges on the observation that the optimal argument of (1) is not a fixed point of the DGD iteration in (2) but is a fixed point of the iteration in (3). To explain this point define $\mathbf{x} := [\mathbf{x}_1; \ldots; \mathbf{x}_N] \in \mathbb{R}^{Np}$ as a vector that concatenates the local iterates $\mathbf{x}_n$ and the aggregate function $f : \mathbb{R}^{Np} \to \mathbb{R}$ as the one that takes values $f(\mathbf{x}) = f(\mathbf{x}_1, \ldots, \mathbf{x}_N) := \sum_{n=1}^{N} f_n(\mathbf{x}_n)$. Decentralized optimization entails the minimization of $f(\mathbf{x})$ subject to the constraint that all local variables are equal,

$$\begin{aligned} \mathbf{x}^* := \operatorname{argmin}\ & f(\mathbf{x}) = f(\mathbf{x}_1, \ldots, \mathbf{x}_N) = \sum_{n=1}^{N} f_n(\mathbf{x}_n), \\ \text{s.t.}\quad & \mathbf{x}_n = \mathbf{x}_m, \quad \text{for all } n, m. \end{aligned} \tag{13}$$

The problems in (1) and (13) are equivalent in the sense that the vector $\mathbf{x}^* \in \mathbb{R}^{Np}$ is a solution of (13) if it satisfies $\mathbf{x}_n^* = \tilde{\mathbf{x}}^*$ for all n, or, equivalently, if we can write $\mathbf{x}^* = [\tilde{\mathbf{x}}^*; \ldots; \tilde{\mathbf{x}}^*]$. Regardless of interpretation, the Karush, Kuhn, Tucker (KKT) conditions of (13) dictate that that optimal argument $\mathbf{x}^*$ must sastisfy

$$\mathbf{x}^* \subset \operatorname{span}(\mathbf{1}_N \otimes \mathbf{I}_p), \qquad (\mathbf{1}_N \otimes \mathbf{I}_p)^T \nabla f(\mathbf{x}^*) = \mathbf{0}. \tag{14}$$

The first condition in (14) requires that all the local variables $\mathbf{x}_n^*$ be equal, while the second condition requires the sum of local gradients to vanish at the optimal point. This latter condition is not the same as $\nabla f(\mathbf{x}) = \mathbf{0}$. If we observe that the gradient $\nabla f(\mathbf{x}^t)$ of the aggregate function can be written as $\nabla f(\mathbf{x}) = [\nabla f_1(\mathbf{x}_1); \ldots; \nabla f_N(\mathbf{x}_N)] \in \mathbb{R}^{Np}$, the condition $\nabla f(\mathbf{x}) = \mathbf{0}$ implies that all the local gradients are null, i.e., that $\nabla f_n(\mathbf{x}_n) = \mathbf{0}$ for all n. This is stronger than having their sum being null as required by (14).

Define now the extended weight matrices as the Kronecker products $\mathbf{Z} := \mathbf{W} \otimes \mathbf{I} \in \mathbb{R}^{Np \times Np}$ and $\tilde{\mathbf{Z}} := \tilde{\mathbf{W}} \otimes \mathbf{I} \in \mathbb{R}^{Np \times Np}$. Note that the required conditions for the weight matrices $\mathbf{W}$ and $\tilde{\mathbf{W}}$ in Assumption 1 enforce some conditions on the extended weight matrices $\mathbf{Z}$ and $\tilde{\mathbf{Z}}$. Based on Assumption 1(a), the matrices $\mathbf{Z}$ and $\tilde{\mathbf{Z}}$ are also symmetric, i.e., $\mathbf{Z} = \mathbf{Z}^T$ and $\tilde{\mathbf{Z}} = \tilde{\mathbf{Z}}^T$. Conditions in

Assumption 1(b) imply that null$\{\tilde{\mathbf{Z}}-\mathbf{Z}\}$ = span$\{\mathbf{1}\otimes\mathbf{I}\}$, null$\{\mathbf{I}-\mathbf{Z}\}$ = span$\{\mathbf{1}\otimes\mathbf{I}\}$, and null$\{\mathbf{I}-\tilde{\mathbf{Z}}\} \supseteq$ span$\{\mathbf{1}\otimes\mathbf{I}\}$. Lastly, the spectral properties of matrices $\mathbf{W}$ and $\tilde{\mathbf{W}}$ in Assumption 1(c) yield that matrix $\tilde{\mathbf{Z}}$ is positive definite and the expression $\mathbf{Z} \preceq \tilde{\mathbf{Z}} \preceq (\mathbf{I}+\mathbf{Z})/2$ holds.

According to the definition of the extended weight matrix $\mathbf{Z}$, the DGD iteration in (2) is equivalent to

$$\mathbf{x}^{t+1} = \mathbf{Z}\mathbf{x}^t - \alpha\nabla f(\mathbf{x}^t), \tag{15}$$

where, according to (13), the gradient $\nabla f(\mathbf{x}^t)$ of the aggregate function can be written as $\nabla f(\mathbf{x}^t) = [\nabla f_1(\mathbf{x}_1^t); \ldots; \nabla f_N(\mathbf{x}_N^t)] \in \mathbb{R}^{Np}$. Likewise, the EXTRA iteration in (3) can be written as

$$\mathbf{x}^{t+1} = (\mathbf{I}+\mathbf{Z})\mathbf{x}^t - \tilde{\mathbf{Z}}\mathbf{x}^{t-1} - \alpha\left[\nabla f(\mathbf{x}^t) - \nabla f(\mathbf{x}^{t-1})\right]. \tag{16}$$

The fundamental difference between DGD and EXTRA is that a fixed point of (15) does not necessarily satisfy (14), whereas the fixed points of (16) are guaranteed to do so. Indeed, taking limits in (15) we see that the fixed points $\mathbf{x}^\infty$ of DGD must satisfy

$$(\mathbf{I}-\mathbf{Z})\mathbf{x}^\infty + \alpha\nabla f(\mathbf{x}^\infty) = \mathbf{0}, \tag{17}$$

which is incompatible with (14) except in peculiar circumstances – such as, e.g., when all local functions have the same minimum. The limit points of EXTRA, however, satisfy the relationship

$$\mathbf{x}^\infty - \mathbf{x}^\infty = (\mathbf{Z}-\tilde{\mathbf{Z}})\mathbf{x}^\infty - \alpha[\nabla f(\mathbf{x}^\infty) - \nabla f(\mathbf{x}^\infty)]. \tag{18}$$

Canceling out the variables on the left hand side and the gradients in the right hand side it follows that $(\mathbf{Z}-\tilde{\mathbf{Z}})\mathbf{x}^\infty = \mathbf{0}$. Since the null space of of $\mathbf{Z}-\tilde{\mathbf{Z}}$ is null$(\mathbf{Z}-\tilde{\mathbf{Z}}) = \mathbf{1}_N\otimes\mathbf{I}_p$ by assumption, we must have $\mathbf{x}^\infty \subset$ span$(\mathbf{1}_N\otimes\mathbf{I}_p)$. This is the first condition in (14). For the second condition in (14) sum the updates in (16) recursively and use the telescopic nature of the sum to write

$$\mathbf{x}^{t+1} = \tilde{\mathbf{Z}}\mathbf{x}^t - \alpha\nabla f(\mathbf{x}^t) - \sum_{s=0}^{t}(\tilde{\mathbf{Z}}-\mathbf{Z})\mathbf{x}^s. \tag{19}$$

Substituting the limit point in (19) and reordering terms, we see that $\mathbf{x}^\infty$ must satisfy

$$\alpha\nabla f(\mathbf{x}^\infty) = (\mathbf{I}-\tilde{\mathbf{Z}})\mathbf{x}^\infty - \sum_{s=0}^{\infty}(\tilde{\mathbf{Z}}-\mathbf{Z})\mathbf{x}^s. \tag{20}$$

In (20) we have that $(\mathbf{I}-\tilde{\mathbf{Z}})\mathbf{x}^\infty = \mathbf{0}$ because the null space of $(\mathbf{I}-\tilde{\mathbf{Z}})$ is null$(\mathbf{Z}-\tilde{\mathbf{Z}}) = \mathbf{1}_N\otimes\mathbf{I}_p$ by assumption and $\mathbf{x}^\infty \subset$ span$(\mathbf{1}_N\otimes\mathbf{I}_p)$ as already shown. Implementing this simplification and considering the multiplication of the resulting equality by $(\mathbf{1}_N\otimes\mathbf{I}_p)^T$ we obtain

$$(\mathbf{1}_N\otimes\mathbf{I}_p)^T\alpha\nabla f(\mathbf{x}^\infty) = -\sum_{s=0}^{\infty}(\mathbf{1}_N\otimes\mathbf{I}_p)^T(\mathbf{Z}-\tilde{\mathbf{Z}})\mathbf{x}^s. \tag{21}$$

In (21), the terms $(\mathbf{1}_N\otimes\mathbf{I}_p)^T(\mathbf{Z}-\tilde{\mathbf{Z}}) = \mathbf{0}$ because the matrices $\mathbf{Z}$ and $\tilde{\mathbf{Z}}$ are symmetric and $(\mathbf{1}_N\otimes\mathbf{I}_p)$ is in the null space of the difference $\mathbf{Z}-\tilde{\mathbf{Z}}$. This implies that $(\mathbf{1}_N\otimes\mathbf{I}_p)^T\alpha\nabla f(\mathbf{x}^\infty) = \mathbf{0}$, which is the second condition in (14). Therefore, given the assumption that the sequence of EXTRA iterates $\mathbf{x}^t$ has a limit point $\mathbf{x}^\infty$ it follows that this limit point satisfies both conditions in (14) and for this reason exact convergence with constant stepsize is achievable for EXTRA.

2.2 Stochastic Saddle Point Method Interpretation of DSA

The convergence proofs of DSA build on a reinterpretation of EXTRA as a saddle point method. To introduce this primal-dual interpretation consider the update in (19) and define the sequence of

vectors $\mathbf{v}^t = \sum_{s=0}^{t}(\tilde{\mathbf{Z}}-\mathbf{Z})^{1/2}\mathbf{x}^s$. The vector $\mathbf{v}^t$ represents the accumulation of variable dissimilarities in different nodes over time. Considering this definition of $\mathbf{v}^t$ we can rewrite (19) as

$$\mathbf{x}^{t+1} = \mathbf{x}^t - \alpha\left[\nabla f(\mathbf{x}^t) + \frac{1}{\alpha}(\mathbf{I}-\tilde{\mathbf{Z}})\mathbf{x}^t + \frac{1}{\alpha}(\tilde{\mathbf{Z}}-\mathbf{Z})^{1/2}\mathbf{v}^t\right]. \tag{22}$$

Furthermore, based on the definition of the sequence $\mathbf{v}^t = \sum_{s=0}^{t}(\tilde{\mathbf{Z}}-\mathbf{Z})^{1/2}\mathbf{x}^s$ we can write the recursive expression

$$\mathbf{v}^{t+1} = \mathbf{v}^t + \alpha\left[\frac{1}{\alpha}(\tilde{\mathbf{Z}}-\mathbf{Z})^{1/2}\mathbf{x}^{t+1}\right]. \tag{23}$$

Consider $\mathbf{x}$ as a primal variable and $\mathbf{v}$ as a dual variable. Then, the updates in (22) and (23) are equivalent to the updates of a saddle point method with stepsize α that solves for the critical points of the augmented Lagrangian

$$\mathcal{L}(\mathbf{x},\mathbf{v}) = f(\mathbf{x}) + \frac{1}{\alpha}\mathbf{v}^T(\tilde{\mathbf{Z}}-\mathbf{Z})^{1/2}\mathbf{x} + \frac{1}{2\alpha}\mathbf{x}^T(\mathbf{I}-\tilde{\mathbf{Z}})\mathbf{x}. \tag{24}$$

In the Lagrangian in (24) the factor $(1/\alpha)\mathbf{v}^T(\tilde{\mathbf{Z}}-\mathbf{Z})^{1/2}\mathbf{x}$ stems from the linear constraint $(\tilde{\mathbf{Z}}-\mathbf{Z})^{1/2}\mathbf{x}=\mathbf{0}$ and the quadratic term $(1/2\alpha)\mathbf{x}^T(\mathbf{I}-\tilde{\mathbf{Z}})\mathbf{x}$ is the augmented term added to the Lagrangian. Therefore, the optimization problem whose augmented Lagrangian is the one given in (24) is

$$\mathbf{x}^* = \underset{\mathbf{x}}{\text{argmin}}\ f(\mathbf{x}) \qquad \text{s.t. } \frac{1}{\alpha}(\tilde{\mathbf{Z}}-\mathbf{Z})^{1/2}\mathbf{x} = \mathbf{0}. \tag{25}$$

Observing that the null space of $(\tilde{\mathbf{Z}}-\mathbf{Z})^{1/2}$ is $\text{null}((\tilde{\mathbf{Z}}-\mathbf{Z})^{1/2}) = \text{null}(\tilde{\mathbf{Z}}-\mathbf{Z}) = \text{span}\{\mathbf{1}_N \otimes \mathbf{I}_p\}$, the constraint in (25) is equivalent to the consensus constraint $\mathbf{x}_n = \mathbf{x}_m$ for all n, m that appears in (13). This means that (25) is equivalent to (13), which, as already argued, is equivalent to the original problem in (1). Hence, EXTRA is a saddle point method that solves (25) which, because of their equivalence, is tantamount to solving (1). Considering that saddle point methods converge linearly, it follows that the same is true of EXTRA.

That EXTRA is a saddle point method provides a simple explanation of its convergence properties. For the purposes of this paper, however, the important fact is that if EXTRA is a saddle point method, DSA is a stochastic saddle point method. To write DSA in this form define $\hat{\mathbf{g}}^t := [\hat{\mathbf{g}}_1^t; \ldots; \hat{\mathbf{g}}_N^t] \in \mathbb{R}^{Np}$ as the vector that concatenates all the local stochastic averaging gradients at step t. Then, the DSA update in (8) can be written as

$$\mathbf{x}^{t+1} = (\mathbf{I}+\mathbf{Z})\mathbf{x}^t - \tilde{\mathbf{Z}}\mathbf{x}^{t-1} - \alpha\left[\hat{\mathbf{g}}^t - \hat{\mathbf{g}}^{t-1}\right]. \tag{26}$$

Comparing (16) and (26) we see that they differ in the latter using stochastic averaging gradients $\hat{\mathbf{g}}^t$ in lieu of the full gradients $\nabla f(\mathbf{x}^t)$. Therefore, DSA is a stochastic saddle point method in which the primal variables are updated as

$$\mathbf{x}^{t+1} = \mathbf{x}^t - \alpha\hat{\mathbf{g}}^t - (\mathbf{I}-\tilde{\mathbf{Z}})\mathbf{x}^t - (\tilde{\mathbf{Z}}-\mathbf{Z})^{1/2}\mathbf{v}^t, \tag{27}$$

and the dual variables $\mathbf{v}^t$ are updated as

$$\mathbf{v}^{t+1} = \mathbf{v}^t + (\tilde{\mathbf{Z}}-\mathbf{Z})^{1/2}\mathbf{x}^{t+1}. \tag{28}$$

Notice that the initial primal variable $\mathbf{x}^0$ is an arbitrary vector in $\mathbb{R}^{Np}$, while according to the definition $\mathbf{v}^t = \sum_{s=0}^{t}(\tilde{\mathbf{Z}}-\mathbf{Z})^{1/2}\mathbf{x}^s$. We then need to set the initial multiplier to $\mathbf{v}^0 = (\tilde{\mathbf{Z}}-\mathbf{Z})^{1/2}\mathbf{x}^0$. This is not a problem in practice because (27) and (28) are not used for implementation. In our converge analysis we utilize the (equivalent) stochastic saddle point expressions for DSA shown in (27) and (28). The expression in (8) is used for implementation because it avoids exchanging dual variables – as well as the initialization problem. The convergence analysis is presented in the following section.

3. Convergence Analysis

Our goal here is to show that as time progresses the sequence of iterates $\mathbf{x}^t$ approaches the optimal argument $\mathbf{x}^*$. To do so, in addition to the conditions on the weight matrices $\mathbf{W}$ and $\tilde{\mathbf{W}}$ in Assumption 1, we assume the instantaneous local functions $f_{n,i}$ have specific properties that we state next.

Assumption 2 The instantaneous local functions $f_{n,i}(\mathbf{x}_n)$ are differentiable and strongly convex with parameter μ.

Assumption 3 The gradient of instantaneous local functions $\nabla f_{n,i}$ are Lipschitz continuous with parameter L, i.e., for all $n \in \{1,\ldots,N\}$ and $i \in \{1,\ldots,q_n\}$ we can write

$$\|\nabla f_{n,i}(\mathbf{a}) - \nabla f_{n,i}(\mathbf{b})\| \leq L\ \|\mathbf{a}-\mathbf{b}\| \quad \mathbf{a},\mathbf{b} \in \mathbb{R}^p. \tag{29}$$

The condition imposed by Assumption 2 implies that the local functions $f_n(\mathbf{x}_n)$ and the global cost function $f(\mathbf{x}) = \sum_{n=1}^N f_n(\mathbf{x}_n)$ are also strongly convex with parameter μ. Likewise, Lipschitz continuity of the local instantaneous gradients considered in Assumption 3 enforces Lipschitz continuity of the local functions gradient $\nabla f_n(\mathbf{x}_n)$ and the aggregate function gradient $\nabla f(\mathbf{x})$ – see, e.g., (Lemma 1 of Mokhtari et al. (2015a)).

3.1 Preliminaries

In this section we study some basic properties of the sequences of primal and dual variables generated by the DSA algorithm. In the following lemma, we study the relation of the iterates $\mathbf{x}^t$ and $\mathbf{v}^t$ with the optimal primal $\mathbf{x}^*$ and dual $\mathbf{v}^*$ arguments.

Lemma 3 *Consider the DSA algorithm as defined in* (6)-(9) *and recall the updates of the primal* $\mathbf{x}^t$ *and dual* $\mathbf{v}^t$ *variables in* (27) *and* (28)*, respectively. Further, define the positive semidefinite matrix* $\mathbf{U} := (\tilde{\mathbf{Z}} - \mathbf{Z})^{1/2}$*. If Assumption 1 holds true, then the sequence of primal* $\mathbf{x}^t$ *and dual* $\mathbf{v}^t$ *variables satisfy*

$$\alpha\left[\hat{\mathbf{g}}^t - \nabla f(\mathbf{x}^*)\right] = (\mathbf{I} + \mathbf{Z} - 2\tilde{\mathbf{Z}})(\mathbf{x}^* - \mathbf{x}^{t+1}) + \tilde{\mathbf{Z}}(\mathbf{x}^t - \mathbf{x}^{t+1}) - \mathbf{U}(\mathbf{v}^{t+1} - \mathbf{v}^*). \tag{30}$$

Proof Considering the update rule for the dual variable in (28) and the definition $\mathbf{U} = (\tilde{\mathbf{Z}} - \mathbf{Z})^{1/2}$, we can substitute $\mathbf{U}\mathbf{v}^t$ in (27) by $\mathbf{U}\mathbf{v}^{t+1} - \mathbf{U}^2\mathbf{x}^{t+1}$. Applying this substitution into the DSA primal update in (27) yields

$$\alpha\hat{\mathbf{g}}^t = -(\mathbf{I} + \mathbf{Z} - \tilde{\mathbf{Z}})\mathbf{x}^{t+1} + \tilde{\mathbf{Z}}\mathbf{x}^t - \mathbf{U}\mathbf{v}^{t+1}. \tag{31}$$

By adding and subtracting $\tilde{\mathbf{Z}}\mathbf{x}^{t+1}$ to the right hand side of (31) and considering the fact that $(\mathbf{I} + \mathbf{Z} - 2\tilde{\mathbf{Z}})\mathbf{x}^* = \mathbf{0}$ we obtain

$$\alpha\hat{\mathbf{g}}^t = (\mathbf{I} + \mathbf{Z} - 2\tilde{\mathbf{Z}})(\mathbf{x}^* - \mathbf{x}^{t+1}) + \tilde{\mathbf{Z}}(\mathbf{x}^t - \mathbf{x}^{t+1}) - \mathbf{U}\mathbf{v}^{t+1}. \tag{32}$$

One of the KKT conditions of problem (25) implies that the optimal variables $\mathbf{x}^*$ and $\mathbf{v}^*$ satisfy $\alpha\nabla f(\mathbf{x}^*) + \mathbf{U}\mathbf{v}^* = \mathbf{0}$ or equivalently $-\alpha\nabla f(\mathbf{x}^*) = \mathbf{U}\mathbf{v}^*$. Adding this equality to both sides of (32) follows the claim in (30). ■

In the subsequent analyses of convergence of DSA, we need an upper bound for the expected value of squared difference between the stochastic averaging gradient $\hat{\mathbf{g}}^t$ and the optimal argument gradient $\nabla f(\mathbf{x}^*)$ given the observations until step t, i.e. $\mathbb{E}\left[\|\hat{\mathbf{g}}^t - \nabla f(\mathbf{x}^*)\|^2 \mid \mathcal{F}^t\right]$. To establish this upper bound first we define the sequence $p^t \in \mathbb{R}$ as

$$p^t := \sum_{n=1}^N \left[\frac{1}{q_n}\sum_{i=1}^{q_n}\left(f_{n,i}(\mathbf{y}_{n,i}^t) - f_{n,i}(\tilde{\mathbf{x}}^*) - \nabla f_{n,i}(\tilde{\mathbf{x}}^*)^T(\mathbf{y}_{n,i}^t - \tilde{\mathbf{x}}^*)\right)\right]. \tag{33}$$

Notice that based on the strong convexity of the local instantaneous functions $f_{n,i}$, each term $f_{n,i}(\mathbf{y}_{n,i}^t) - f_{n,i}(\tilde{\mathbf{x}}^*) - \nabla f_{n,i}(\tilde{\mathbf{x}}^*)^T(\mathbf{y}_{n,i}^t - \tilde{\mathbf{x}}^*)$ is positive and as a result the sequence p^t defined in (33) is always positive. In the following lemma, we use the result in Lemma 3 to guarantee an upper bound for the expectation $\mathbb{E}\left[\|\hat{\mathbf{g}}^t - \nabla f(\mathbf{x}^*)\|^2 \mid \mathcal{F}^t\right]$ in terms of p^t and the optimality gap $f(\mathbf{x}^t) - f(\mathbf{x}^*) - \nabla f(\mathbf{x}^*)^T(\mathbf{x}^t - \mathbf{x}^*)$.

Lemma 4 *Consider the DSA algorithm in* (6)-(9) *and the definition of the sequence* p^t *in* (33). *If Assumptions 1-3 hold true, then the squared norm of the difference between the stochastic averaging gradient* $\hat{\mathbf{g}}^t$ *and the optimal gradient* $\nabla f(\mathbf{x}^*)$ *in expectation is bounded above by*

$$\mathbb{E}\left[\left\|\hat{\mathbf{g}}^t - \nabla f(\mathbf{x}^*)\right\|^2 | \mathcal{F}^t\right] \leq 4Lp^t + 2\left(2L - \mu\right)\left(f(\mathbf{x}^t) - f(\mathbf{x}^*) - \nabla f(\mathbf{x}^*)^T(\mathbf{x}^t - \mathbf{x}^*)\right). \tag{34}$$

Proof See Appendix A. ■

Observe that as the sequence of iterates $\mathbf{x}^t$ approaches the optimal argument $\mathbf{x}^*$, all the local auxiliary variables $\mathbf{y}_{n,i}^t$ converge to $\tilde{\mathbf{x}}^*$ which follows convergence of p^t to null. This observation in association with the result in (34) implies that the expected value of the difference between the stochastic averaging gradient $\hat{\mathbf{g}}^t$ and the optimal gradient $\nabla f(\mathbf{x}^*)$ vanishes as the sequence of iterates $\mathbf{x}^t$ approaches the optimal argument $\mathbf{x}^*$.

3.2 Convergence

In this section we establish linear convergence of the sequence of iterates $\mathbf{x}^t$ generated by DSA to the optimal argument $\mathbf{x}^*$. To do so, define $0 < \gamma$ and $\Gamma < \infty$ as the smallest and largest eigenvalues of the positive definite matrix $\tilde{\mathbf{Z}}$, respectively. Likewise, define γ' as the smallest non-zero eigenvalue of the matrix $\tilde{\mathbf{Z}} - \mathbf{Z}$ and Γ' as the largest eigenvalue of the matrix $\tilde{\mathbf{Z}} - \mathbf{Z}$. Further, define the vectors $\mathbf{u}^t, \mathbf{u}^* \in \mathbb{R}^{2Np}$ and matrix $\mathbf{G} \in \mathbb{R}^{2Np \times 2Np}$ as

$$\mathbf{u}^* := \begin{bmatrix} \mathbf{x}^* \\ \mathbf{v}^* \end{bmatrix}, \quad \mathbf{u}^t := \begin{bmatrix} \mathbf{x}^t \\ \mathbf{v}^t \end{bmatrix}, \quad \mathbf{G} = \begin{bmatrix} \tilde{\mathbf{Z}} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}. \tag{35}$$

Observe that the vector $\mathbf{u}^* \in \mathbb{R}^{2Np}$ concatenates the optimal primal and dual variables and the vector $\mathbf{u}^t \in \mathbb{R}^{2Np}$ contains primal and dual iterates at step t. Further, $\mathbf{G} \in \mathbb{R}^{2Np \times 2Np}$ is a block diagonal positive definite matrix that we introduce since instead of tracking the value of ℓ_2 norm $\|\mathbf{u}^t - \mathbf{u}^*\|_2^2$ we study the convergence properties of $\mathbf{G}$ weighted norm $\|\mathbf{u}^t - \mathbf{u}^*\|_{\mathbf{G}}^2$. Notice that the weighted norm $\|\mathbf{u}^t - \mathbf{u}^*\|_{\mathbf{G}}^2$ is equivalent to $(\mathbf{u}^t - \mathbf{u}^*)^T\mathbf{G}(\mathbf{u}^t - \mathbf{u}^*)$. Our goal is to show that the sequence $\|\mathbf{u}^t - \mathbf{u}^*\|_{\mathbf{G}}^2$ converges linearly to null. To do this we show linear convergence of a Lyapunov function of the sequence $\|\mathbf{u}^t - \mathbf{u}^*\|_{\mathbf{G}}^2$. The Lyapunov function is defined as $\|\mathbf{u}^t - \mathbf{u}^*\|_{\mathbf{G}}^2 + cp^t$ where $c > 0$ is a positive constant.

To prove linear convergence of the sequence $\|\mathbf{u}^t - \mathbf{u}^*\|_{\mathbf{G}}^2 + cp^t$ we first show an upper bound for the expected error $\mathbb{E}\left[\|\mathbf{u}^{t+1} - \mathbf{u}^*\|_{\mathbf{G}}^2 \mid \mathcal{F}^t\right]$ in terms of $\|\mathbf{u}^t - \mathbf{u}^*\|_{\mathbf{G}}^2$ and some parameters that capture optimality gap.

Lemma 5 *Consider the DSA algorithm as defined in* (6)-(9). *Further recall the definitions of* p^t *in* (33) *and* $\mathbf{u}^t$, $\mathbf{u}^*$, *and* $\mathbf{G}$ *in* (35). *If Assumptions 1-3 hold true, then for any positive constant* $\eta > 0$ *we can write*

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{u}^{t+1} - \mathbf{u}^*\|_{\mathbf{G}}^2 \mid \mathcal{F}^t\right] \leq \|\mathbf{u}^t - \mathbf{u}^*\|_{\mathbf{G}}^2 &- 2\mathbb{E}\left[\|\mathbf{x}^{t+1} - \mathbf{x}^*\|_{\mathbf{I}+\mathbf{Z}-2\tilde{\mathbf{Z}}}^2 \mid \mathcal{F}^t\right] + \frac{\alpha 4L}{\eta}p^t \\ &- \mathbb{E}\left[\|\mathbf{x}^{t+1} - \mathbf{x}^t\|_{\tilde{\mathbf{Z}}-2\alpha\eta\mathbf{I}}^2 \mid \mathcal{F}^t\right] - \mathbb{E}\left[\|\mathbf{v}^{t+1} - \mathbf{v}^t\|^2 \mid \mathcal{F}^t\right] \\ &- \left(\frac{4\alpha\mu}{L} - \frac{2\alpha(2L-\mu)}{\eta}\right)\left(f(\mathbf{x}^t) - f(\mathbf{x}^*) - \nabla f(\mathbf{x}^*)^T(\mathbf{x}^t - \mathbf{x}^*)\right).\end{aligned} \tag{36}$$

Proof See Appendix B. ■

Lemma 5 shows an upper bound for the squared norm $\|\mathbf{u}^{t+1} - \mathbf{u}^*\|^2_{\mathbf{G}}$ which is the first part of the Lyapunov function $\|\mathbf{u}^t - \mathbf{u}^*\|^2_{\mathbf{G}} + cp^t$ at step $t+1$. Likewise, we provide an upper bound for the second term of the Lyapunov function at time $t+1$ which is p^{t+1} in terms of p^t and some parameters that capture optimality gap. This bound is studied in the following lemma.

Lemma 6 *Consider the DSA algorithm as defined in* (6)-(9) *and the definition of* p^t *in* (33)*. Further, define* $q_{\min}$ *and* $q_{\max}$ *as the smallest and largest values for the number of instantaneous functions at a node, respectively. If Assumptions 1-3 hold true, then for all* $t > 0$ *the sequence* p^t *satisfies*

$$\mathbb{E}\left[p^{t+1} \mid \mathcal{F}^t\right] \leq \left[1 - \frac{1}{q_{\max}}\right] p^t + \frac{1}{q_{\min}} \left[f(\mathbf{x}^t) - f(\mathbf{x}^*) - \nabla f(\mathbf{x}^*)^T(\mathbf{x}^t - \mathbf{x}^*)\right]. \tag{37}$$

Proof See Appendix C. ■

Lemma 6 provides an upper bound for p^{t+1} in terms of its previous value p^t and the optimality error $f(\mathbf{x}^t) - f(\mathbf{x}^*) - \nabla f(\mathbf{x}^*)^T(\mathbf{x}^t - \mathbf{x}^*)$. Combining the results in Lemmata 5 and 6 we can show that in expectation the Lyapunov function $\|\mathbf{u}^{t+1} - \mathbf{u}^*\|^2_{\mathbf{G}} + c\ p^{t+1}$ at step $t+1$ is strictly smaller than its previous value $\|\mathbf{u}^t - \mathbf{u}^*\|^2_{\mathbf{G}} + c\ p^t$ at step t.

Theorem 7 *Consider the DSA algorithm as defined in* (6)-(9)*. Further recall the definition of the sequence* p^t *in* (33)*. Define* η *as an arbitrary positive constant chosen from the interval*

$$\eta \in \left(\frac{L^2 q_{\max}}{\mu q_{\min}} + \frac{L^2}{\mu} - \frac{L}{2}\ ,\ \infty\right). \tag{38}$$

If Assumptions 1-3 hold true and the stepsize α *is chosen from the interval* $\alpha \in (0, \gamma/2\eta)$*, then for arbitrary* c *chosen from the interval*

$$c \in \left(\frac{4\alpha L q_{\max}}{\eta}\ ,\ \frac{4\alpha\mu q_{\min}}{L} - \frac{2\alpha q_{\min}(2L-\mu)}{\eta}\right), \tag{39}$$

there exits a positive constant $0 < \delta < 1$ *such that*

$$\mathbb{E}\left[\|\mathbf{u}^{t+1} - \mathbf{u}^*\|^2_{\mathbf{G}} + c\, p^{t+1} \mid \mathcal{F}^t\right] \leq (1-\delta)\left(\|\mathbf{u}^t - \mathbf{u}^*\|^2_{\mathbf{G}} + c\, p^t\right). \tag{40}$$

Proof See Appendix D. ■

We point out that the linear convergence constant δ in (40) is explicitly available – see (100) in Appendix D. It is a function of the strong convexity parameter μ, the Lipschitz continuity constant L, lower and upper bounds on the eigenvalues of the matrices $\tilde{\mathbf{Z}}$, $\tilde{\mathbf{Z}} - \mathbf{Z}$, and $\mathbf{I} + \mathbf{Z} - 2\tilde{\mathbf{Z}}$, the smallest $q_{\min}$ and largest $q_{\max}$ values for the number of instantaneous functions available at a node, and the stepsize α. Insight on the dependence of δ with problem parameters is offered in Section 3.3.

The inequality in (40) shows that the expected value of the sequence $\|\mathbf{u}^t - \mathbf{u}^*\|^2_{\mathbf{G}} + c\, p^t$ at time $t+1$ given the observation until step t is strictly smaller than the previous iterate at step t. Note that, it is not hard to verify that if the positive constant η is chosen from the interval in (38), the interval in (39) is non-empty. Computing the expected value with respect to the initial sigma field $\mathbb{E}\left[.\mid \mathcal{F}^0\right] = \mathbb{E}\left[.\right]$ implies that in expectation the sequence $\|\mathbf{u}^t - \mathbf{u}^*\|^2_{\mathbf{G}} + c\, p^t$ converges linearly to null, i.e.,

$$\mathbb{E}\left[\|\mathbf{u}^t - \mathbf{u}^*\|^2_{\mathbf{G}} + c\, p^t\right] \leq (1-\delta)^t \left(\|\mathbf{u}^0 - \mathbf{u}^*\|^2_{\mathbf{G}} + c\, p^0\right). \tag{41}$$

We use the result in (41) to establish linear convergence of the sequence of squared norm error $\|\mathbf{x}^t - \mathbf{x}^*\|^2$ in expectation.

Corollary 8 *Consider the DSA algorithm as defined in* (6)-(9) *and recall* γ *is the minimum eigenvalue of the positive definite matrix* $\tilde{\mathbf{Z}}$*. Suppose the conditions of Theorem 7 hold, then there exits a positive constant* $0 < \delta < 1$ *such that*

$$\mathbb{E}\left[\|\mathbf{x}^t - \mathbf{x}^*\|^2\right] \leq (1-\delta)^t \frac{\left(\|\mathbf{u}^0 - \mathbf{u}^*\|^2_{\mathbf{G}} + c\, p^0\right).}{\gamma} \tag{42}$$

Proof First note that according to the definitions of $\mathbf{u}$ and $\mathbf{G}$ in (35) and the definition of p^t in (33) , we can write $\|\mathbf{x}^t - \mathbf{x}^*\|^2_{\tilde{\mathbf{Z}}} \leq \|\mathbf{u}^t - \mathbf{u}^*\|^2_{\mathbf{G}} + c p^t$. Further, note that the weighted norm $\|\mathbf{x}^t - \mathbf{x}^*\|^2_{\tilde{\mathbf{Z}}}$ is lower bounded by $\gamma\|\mathbf{x}^t - \mathbf{x}^*\|^2$, since γ is a lower bound for the eigenvalues of $\tilde{\mathbf{Z}}$. Combine these two observations to obtain $\gamma\|\mathbf{x}^t - \mathbf{x}^*\|^2 \leq \|\mathbf{u}^t - \mathbf{u}^*\|^2_{\mathbf{G}} + c\, p^t$. This inequality in conjunction with the expression in (41) follows the claim in (42). ■

Corollary 8 states that the sequence $\mathbb{E}\left[\|\mathbf{x}^t - \mathbf{x}^*\|^2\right]$ linearly converges to null. Note that the sequence $\mathbb{E}\left[\|\mathbf{x}^t - \mathbf{x}^*\|^2\right]$ is not necessarily monotonically decreasing as the sequence $\mathbb{E}\left[\|\mathbf{u}^t - \mathbf{u}^*\|^2_{\mathbf{G}} + c\, p^t\right]$ is. The result in (42) shows linear convergence of the sequence of variables generated by DSA in expectation. In the following Theorem we show that the local variables $\mathbf{x}^t_n$ generated by DSA almost surely converge to the optimal argument of (1).

Theorem 9 *Consider the DSA algorithm as defined in* (6)-(9) *and suppose the conditions of Theorem 7 hold. Then, the sequences of the local variables* $\mathbf{x}^t_n$ *for all* $n = 1, \ldots, N$ *converge almost surely to the optimal argument* $\tilde{\mathbf{x}}^*$*, i.e.,*

$$\lim_{t\to\infty} \mathbf{x}^t_n \;=\; \tilde{\mathbf{x}}^* \quad a.s. \qquad \text{for all } n = 1, \ldots, N. \tag{43}$$

Proof See Appendix E. ■

Theorem 9 provides almost sure convergence of $\mathbf{x}^t$ to the optimal solution $\mathbf{x}^*$ which is stronger result than convergence in expectation as in Corollary 8.

3.3 Convergence Constant

The constant δ that controls the speed of convergence can be simplified by selecting specific values for η, α, and c. This uncovers connections to the properties of the local objective functions and the network topology. To make this clearer recall the definitions of γ and Γ as the smallest and largest eigenvalues of the positive definite matrix $\tilde{\mathbf{Z}}$, respectively, and γ' and Γ' as the smallest and largest *positive* eigenvalues of the positive semi-definite matrix $\tilde{\mathbf{Z}} - \mathbf{Z}$, respectively. Further, recall that the local objective functions are strongly convex with constant μ and their gradients are Lipschitz continuous with constant L. Then, define the condition numbers of the objective function and the graph as

$$\kappa_f = \frac{L}{\mu}, \qquad \kappa_g = \frac{\max\{\Gamma, \Gamma'\}}{\min\{\gamma, \gamma'\}}, \tag{44}$$

respectively. The condition number of the function is a measure of how difficult it is to minimize the local functions using gradient descent directions. The condition number of the graph is a measure of how slow the graph is in propagating a diffusion process. Both are known to control the speed of convergence of distributed optimization methods. The following corollary illustrates that these condition numbers also determine the convergence speed of DSA.

Corollary 10 *Consider the DSA algorithm as defined in* (6)-(9) *and suppose the conditions of Theorem 7 hold. Choose the weight matrices* $\mathbf{W}$ *and* $\tilde{\mathbf{W}}$ *as* $\tilde{\mathbf{W}} = (\mathbf{I} + \mathbf{W})/2$*, assign the same number*

of instantaneous local functions $f_{n,i}$ to each node, i.e., $q_{\min} = q_{\max} = q$, and set the constants η, α and c as

$$\eta = \frac{2L^2}{\mu}, \quad \alpha = \frac{\gamma\mu}{8L^2}, \quad c = \frac{q\gamma\mu^2}{4L^3}\left(1 + \frac{\mu}{4L}\right). \tag{45}$$

The linear convergence constant $0 < \delta < 1$ in (40) reduces to

$$\delta = \min\left[\frac{1}{16\kappa_g^2},\ \frac{1}{q[1+4\kappa_f(1+\gamma/\gamma')]},\ \frac{1}{4(\gamma/\gamma')\kappa_f + 32\kappa_g\kappa_f^4}\right]. \tag{46}$$

Proof The given values for η, α, and c satisfy the conditions in Theorem 7. Substitute then these values into the expression for δ in (100). Simplify terms and utilize the condition number definitions in (44). The second term in the minimization in (100) becomes redundant because it is dominated by the first. ■

Observe that while the choices of η, α, and c in (45) satisfy all the required conditions of Theorem 7, they are not necessarily optimal for maximizing the linear convergence constant δ. Nevertheless, the expression in (46) shows that the convergence speed of DSA decreases with increases in the graph condition number κ_g, the local functions condition number κ_f, and the number of functions assigned to each node q. For a cleaner expression observe that both, γ and γ' are the minimum eigenvalues of the weight matrix $\mathbf{W}$ and the weight matrix difference $\tilde{\mathbf{W}} - \mathbf{W}$. They can therefore be chosen to be of similar order. For reference, say that we choose $\gamma = \gamma'$ so that the ratio $\gamma/\gamma' = 1$. In that case, the constant δ in (46) reduces to

$$\delta = \min\left[\frac{1}{16\kappa_g^2},\ \frac{1}{q(1+8\kappa_f)},\ \frac{1}{4(\kappa_f + 8\kappa_f^4\kappa_g)}\right]. \tag{47}$$

The three terms in (47) establish separate regimes, problems where the graph condition number is large, problems where the number of functions at each node is large, and problems where the condition number of the local functions are large. In the first regime the first term in (47) dominates and establishes a dependence in terms of the square of the graph's condition number. In the second regime the middle term dominates and results in an inverse dependence with the number of functions available at each node. In the third regime, the third term dominates. The dependence in this case is inversely proportional to κ_f^4.

4. Numerical Experiments

We numerically study the performance of DSA in solving a logistic regression problem. In this problem we are given $Q = \sum_{n=1}^{N} q_n$ training samples that we distribute across N distinct nodes. Denote q_n as the number of samples that are assigned to node n. We assume that the samples are distributed equally over the nodes, i.e., $q_n = q_{max} = q_{min} = q = Q/N$ for $n = 1, \ldots, N$. The training points at node n are denoted by $\mathbf{s}_{n,i} \in \mathbb{R}^p$ for $i = 1, \ldots, q_n$ with associated labels $l_{n,i} \in \{-1, 1\}$. The goal is to predict the probability $\mathrm{P}(l = 1 \mid \mathbf{s})$ of having label $l = 1$ for sample point $\mathbf{s}$. The logistic regression model assumes that this probability can be computed as $\mathrm{P}(l = 1 \mid \mathbf{s}) = 1/(1+\exp(-\mathbf{s}^T\mathbf{x}))$ given a linear classifier $\mathbf{x}$ that is computed based on the training samples. It follows from this model that the regularized maximum log likelihood estimate of the classifier $\mathbf{x}$ given the training samples $(\mathbf{s}_{n,i}, l_{n,i})$ for $i = 1, \ldots, q_n$ and $n = 1, \ldots, N$ is the solution of the optimization problem

$$\tilde{\mathbf{x}}^* := \operatorname*{argmin}_{\mathbf{x}\in\mathbb{R}^p} \frac{\lambda}{2}\|\mathbf{x}\|^2 + \sum_{n=1}^{N}\sum_{i=1}^{q_n} \log\Big(1 + \exp(-l_{n,i}\mathbf{s}_{n,i}^T\mathbf{x})\Big), \tag{48}$$

where the regularization term $(\lambda/2)\|\mathbf{x}\|^2$ is added to reduce overfitting to the training set. The optimization problem in (48) can be written in the form of (1) by defining the local objective functions f_n as

$$f_n(\mathbf{x}) = \frac{\lambda}{2N}\|\mathbf{x}\|^2 + \sum_{i=1}^{q_n} \log\Big(1 + \exp(-l_{n,i}\mathbf{s}_{n,i}^T\mathbf{x})\Big). \tag{49}$$

Observe that the local functions f_n in (49) can be written as the average of a set of instantaneous functions $f_{n,i}$ defined as

$$f_{n,i}(\mathbf{x}) = \frac{\lambda}{2N}\|\mathbf{x}\|^2 + q_n \log\Big(1 + \exp\big(-l_{n,i}\mathbf{s}_{n,i}^T\mathbf{x}\big)\Big), \tag{50}$$

for all $i = 1, \ldots, q_n$. Considering the definitions of the instantaneous local functions $f_{n,i}$ in (50) and the local functions f_n in (49), problem (48) can be solved using the DSA algorithm.

In our experiments in Sections 4.1-4.4, we use a synthetic dataset where the components of the feature vectors $\mathbf{s}_{n,i}$ with label $l_{n,i} = 1$ are generated from a normal distribution with mean μ and standard deviation σ_+, while sample points with label $l_{n,i} = -1$ are generated from a normal distribution with mean $-\mu$ and standard deviation σ_-. In Section 4.5, we consider a large-scale real dataset for training the classifier.

We consider a network of size N where the edges between nodes are generated randomly with probability p_c. The weight matrix $\mathbf{W}$ is generated using the Laplacian matrix $\mathbf{L}$ of network as

$$\mathbf{W} = \mathbf{I} - \mathbf{L}/\tau, \tag{51}$$

where τ should satisfy $\tau > (1/2)\lambda_{\max}(\mathbf{L})$. In our experiments we set this parameter as $\tau = (2/3)\lambda_{\max}(\mathbf{L})$. We capture the error of each algorithm by the sum of squared differences of the local iterates $\mathbf{x}_n^t$ from the optimal solution $\tilde{\mathbf{x}}^*$ as

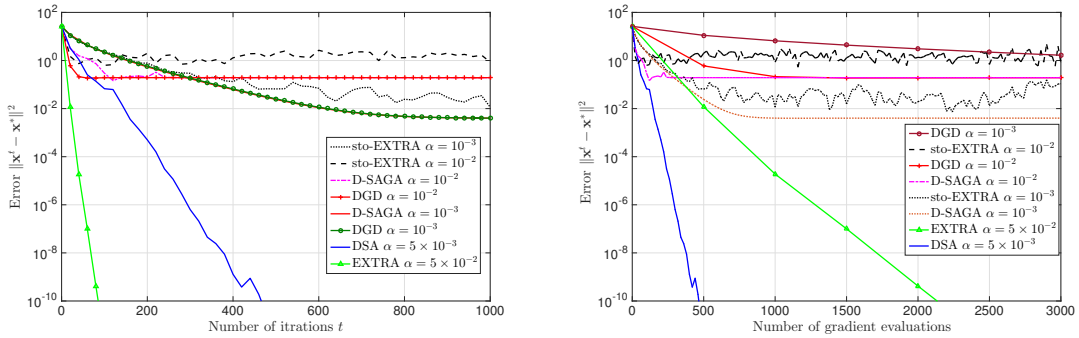
$$e^t = \|\mathbf{x}^t - \mathbf{x}^*\|^2 = \sum_{n=1}^{N} \|\mathbf{x}_n^t - \tilde{\mathbf{x}}^*\|^2. \tag{52}$$

We use a centralized algorithm for computing the optimal argument $\tilde{\mathbf{x}}^*$ in all of our experiments.

4.1 Comparison with Decentralized Methods

We provide a comparison of DSA with respect to DGD, EXTRA, stochastic EXTRA, and decentralized SAGA. The stochastic EXTRA (sto-EXTRA) is defined by using the stochastic gradient in (5) instead of using full gradient as in EXTRA or stochastic averaging gradient as in DSA. The decentralized SAGA (D-SAGA) is a stochastic version of the DGD algorithm that uses stochastic averaging gradient instead of exact gradient which is the naive approach for developing a decentralized version of the SAGA algorithm. In our experiments, the weight matrix $\tilde{\mathbf{W}}$ in EXTRA, stochastic EXTRA, and DSA is chosen as $\tilde{\mathbf{W}} = (\mathbf{I} + \mathbf{W})/2$. We use the total number of sample points $Q = 500$, feature vectors dimension $p = 2$, regularization parameter $\lambda = 10^{-4}$, probability of existence of an edge $p_c = 0.35$. To make the dataset *not* linearly separable we set the mean as $\mu = 2$ and the standard deviations to $\sigma_+ = \sigma_- = 2$. Moreover, the maximum eigenvalue of the Laplacian matrix is $\lambda_{\max}(\mathbf{L}) = 8.017$ which implies that the choice of τ in (51) is $\tau = (2/3)\lambda_{\max}(\mathbf{L}) = 5.345$. We set the total number of nodes $N = 20$ which implies that each node has access to $q = Q/N = 25$ samples.

Fig. 2 illustrates the convergence paths of DSA, EXTRA, DGD, Stochastic EXTRA, and Decentralized SAGA with constant stepsizes for $N = 20$ nodes. For EXTRA and DSA different stepsizes are chosen and the best performance for EXTRA and DSA are achieved by $\alpha = 5 \times 10^{-2}$ and $\alpha = 5 \times 10^{-3}$, respectively. It is worth mentioning that the choice of stepsize α for DSA in practice is larger than the theoretical result in Theorem 6 and Corollary 9 which suggest stepsize of the



(a) (b)

Figure 2: Convergence paths of DSA, EXTRA, DGD, Stochastic EXTRA, and Decentralized SAGA for a logistic regression problem with $Q = 500$ samples and $N = 20$ nodes. Distance to optimality $e^t = \|\mathbf{x}^t - \mathbf{x}^*\|^2$ is shown with respect to number of iterations t and number of gradient evaluations in Fig 2(a) and Fig. 2(b), respectively. DSA and EXTRA converge linearly to the optimal argument $\mathbf{x}^*$, while DGD, Stochastic EXTRA, and Decentralized SAGA with constant step sizes converge to a neighborhood of the optimal solution. Smaller choice of stepsize for DGD, Stochastic EXTRA, and Decentralized SAGA leads to a more accurate convergence, while the speed of convergence becomes slower. DSA outperforms EXTRA in terms of number of gradient evaluations to achieve a target accuracy.

order $O(\mu/L^2)$. As shown in Fig. 2, DSA is the only stochastic algorithm that converges linearly. Decentralized SAGA after a few iterations achieves the performance of DGD and they both cannot converge to the optimal argument. By choosing a smaller stepsize as $\alpha = 10^{-3}$, they reach a more accurate convergence relative to the case that the stepsize is $\alpha = 10^{-2}$; however, the speed of convergence is slower for the smaller stepsize. Stochastic EXTRA also suffers from inexact convergence, but for a different reason. DGD and decentralized SAGA have inexact convergence since they solve a penalty version of the original problem, while stochastic EXTRA can not reach the optimal solution since the noise of stochastic gradient is not vanishing. DSA resolves both issues by combining the idea of stochastic averaging from SAGA to control the noise of stochastic gradient estimation and the double descent idea of EXTRA to solve the correct optimization problem.

Fig. 2(a) illustrates convergence paths of the considered methods in terms of number of iterations t. Notice that the number of iterations t indicates the number of local iterations processed at each node. Convergence rate of EXTRA is faster than DSA in terms of number of iterations or equivalently number of communications as shown in Fig. 2(a); however, the complexity of each iteration for EXTRA is higher than DSA. Therefore, it is reasonable to compare the performances of these algorithms in terms of number of processed feature vectors or equivalently number of gradient evaluations. For instance, DSA requires $t = 380$ iterations or equivalently 380 gradient evaluations to achieve the error $e^t = 10^{-8}$, while to achieve the same accuracy EXTRA requires $t = 69$ iterations which is equivalent to $t \times q_n = 69 \times 25 = 1725$ processed feature vectors or gradient evaluations.

To illustrate this difference better, we compare the convergence paths of DSA, EXTRA, DGD, Stochastic EXTRA, and Decentralized SAGA in terms of number of gradient evaluations in Fig. 2(b). Note that the total number of gradient evaluations at each node for the stochastic methods such as DSA, sto-EXTRA, and D-SAGA is equal to the the number of iterations t, while for EXTRA and DGD – which are deterministic methods – the number of gradient evaluations is equal to the product $t \times q$. This is true since each node in the stochastic methods only evaluates 1 gradient

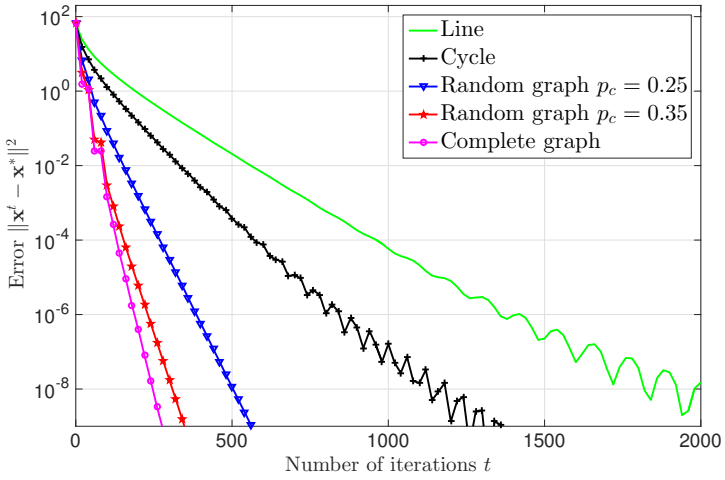


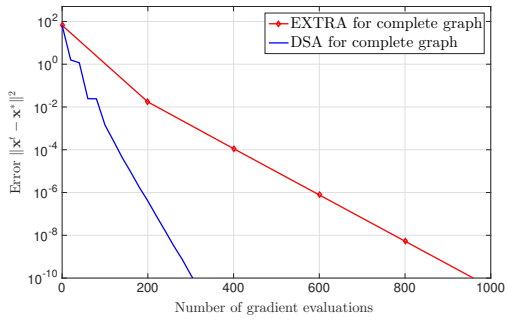
Figure 3: Convergence of DSA for different network topologies when the total number of samples is $Q = 500$ and the size of network is $N = 50$. Distance to optimality $e^t = \|\mathbf{x}_t - \mathbf{x}^*\|^2$ is shown with respect to number of iterations t. As the graph condition number κ_g becomes larger the linear convergence of DSA becomes slower. The best performance belongs to the complete graph which has the smallest condition number and the slowest convergence path belongs to the line graph which has the largest graph condition number.

per iteration, while in the deterministic methods each node requires q gradient evaluations per iteration. The convergence paths in Fig. 2(b) showcase the advantage of DSA relative to EXTRA in requiring less processed feature vectors (or equivalently gradient evaluations) for achieving a specific accuracy. It is important to mention that the initial gradient evaluations for the DSA method is not considered in Fig. 2(b) since the initial decision variable is $\mathbf{x}^0 = \mathbf{0}$ in all experiments and evaluation of the initial gradients $\nabla f_{n,i}(\mathbf{x}^0) = -(1/2)ql_{n,i}\mathbf{s}_{n,i}$ is not computationally expensive relative to the general gradient computation which is given by $\nabla f_{n,i}(\mathbf{x}) = (\lambda\mathbf{x}/N) - (ql_{n,i}\mathbf{s}_{n,i})/(1 + \exp(l_{n,i}\mathbf{x}^T\mathbf{s}_{n,i}))$. However, if we consider this initial processing the plot for DSA in Fig. 2(b) will be shifted by $q = 25$ gradient evaluations which doesn't change the conclusion that DSA outperforms EXTRA in terms of gradient evaluations

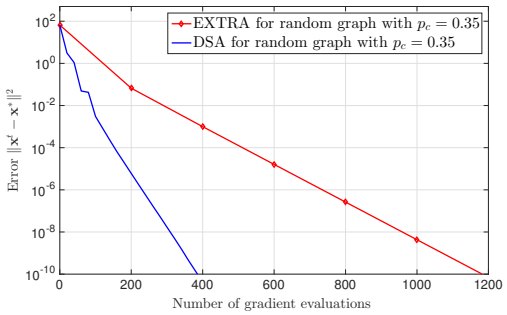
4.2 Effect of Graph Condition Number κ_g

In this section we study the effect of the graph condition number κ_g as defined in (44) on the performance of DSA. We keep the parameters in Fig. 2 except for the network size N which we set as $N = 50$. Thus, each node has access to $q = 500/50 = 10$ sample points. The convergence paths of the DSA algorithm for random networks with $p_c = 0.25$ and $p_c = 0.35$, complete graph, cycle, and line are shown in Fig. 3. Notice that the graph condition number of the line graph, cycle graph, random graph with $p_c = 0.25$, random graph with $p_c = 0.35$, and complete graph are $\kappa_g = 1.01 \times 10^3$, $\kappa_g = 2.53 \times 10^2$, $\kappa_g = 17.05$, $\kappa_g = 4.87$, and $\kappa_g = 4$, respectively. For each network topology, we have hand-optimized the stepsize α and the best choice of stepsize for the complete graph, random graph with $p_c = 0.35$, random graph with $p_c = 0.25$, cycle, and line are $\alpha = 2 \times 10^{-2}$, $\alpha = 1.5 \times 10^{-2}$, $\alpha = 10^{-2}$, $\alpha = 5 \times 10^{-3}$, and $\alpha = 3 \times 10^{-3}$, respectively.

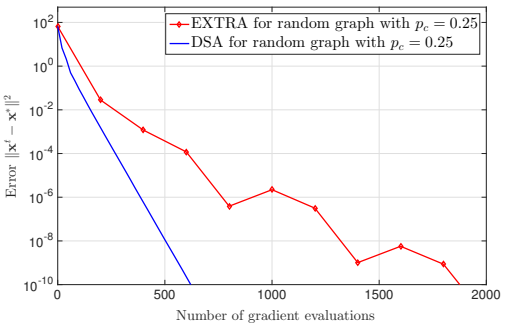
As we expect for the topologies that the graph has more edges and the graph condition number κ_g is smaller we observe a faster linear convergence for DSA. The best performance belongs to the



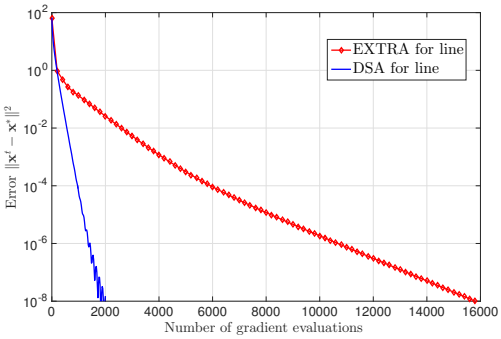
(a) complete graph



(b) random graph $p_c = 0.35$



(c) random graph $p_c = 0.25$



(d) line

Figure 4: Convergence paths of DSA and EXTRA for different network topologies when the total number of samples is $Q = 500$ and the size of network is $N = 50$. Distance to optimality $e^t = \|\mathbf{x}^t - \mathbf{x}^*\|^2$ is shown with respect to number of gradient evaluations. DSA converges faster relative to EXTRA in all of the considered networks. The difference between the convergence paths of DSA and EXTRA is more substantial when the graph has a large condition number κ_g. The stepsize α for DSA and EXTRA in all the considered cases is hand-optimized and the results for the best choice of α are reported.

complete graph which requires $t = 247$ iterations to achieve the relative error $e^t = 10^{-8}$. In the random graphs with connectivity probabilities $p_c = 0.35$ and $p_c = 0.25$, DSA achieves the relative error $e^t = 10^{-8}$ after $t = 310$ and $t = 504$ iterations, respectively. For the cycle and line graphs the numbers of required iterations for reaching the relative error $e^t = 10^{-8}$ are $t = 1133$ and $t = 1819$, respectively. These observations match the theoretical result in (47) that DSA converges faster when the graph condition number κ_g is smaller.

We also compare the performances of DSA and EXTRA over different topologies to verify the claim that DSA is more efficient than EXTRA in terms of number of gradient evaluations over different network topologies. The parameters are as in Fig. 3 and the stepsize α for EXTRA in different topologies are optimized separately. In particular, the best stepsize for the complete graph, random graph with $p_c = 0.35$, random graph with $p_c = 0.25$, and line are $\alpha = 6\times10^{-2}$, $\alpha = 5\times10^{-2}$, $\alpha = 3 \times 10^{-2}$, and $\alpha = 5 \times 10^{-2}$, respectively. Fig. 4 shows the convergence paths of DSA and EXTRA versus number of gradient evaluations for four different network topologies. We observe

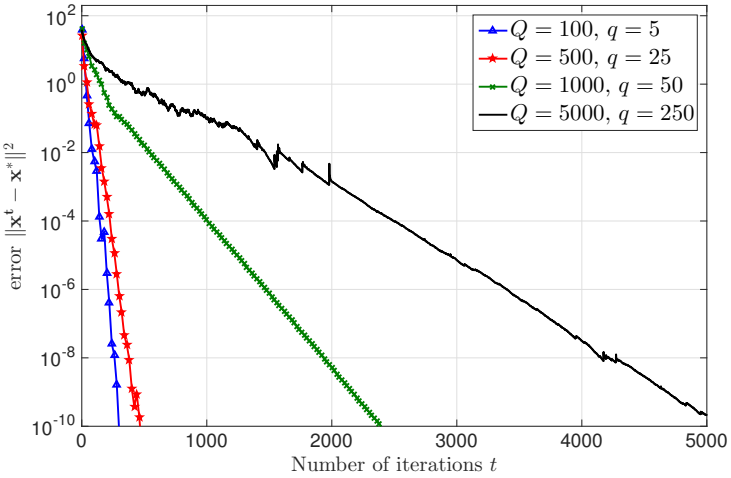


Figure 5: Comparison of convergence paths of DSA for different number of samples Q when the network size is $N = 20$ and the graph is randomly generated with the connectivity ratio $p_c = 0.35$. Convergence time for DSA increases by increasing the total number of sample points Q which is equivalent to increasing the number of samples at each node $q = Q/N$.

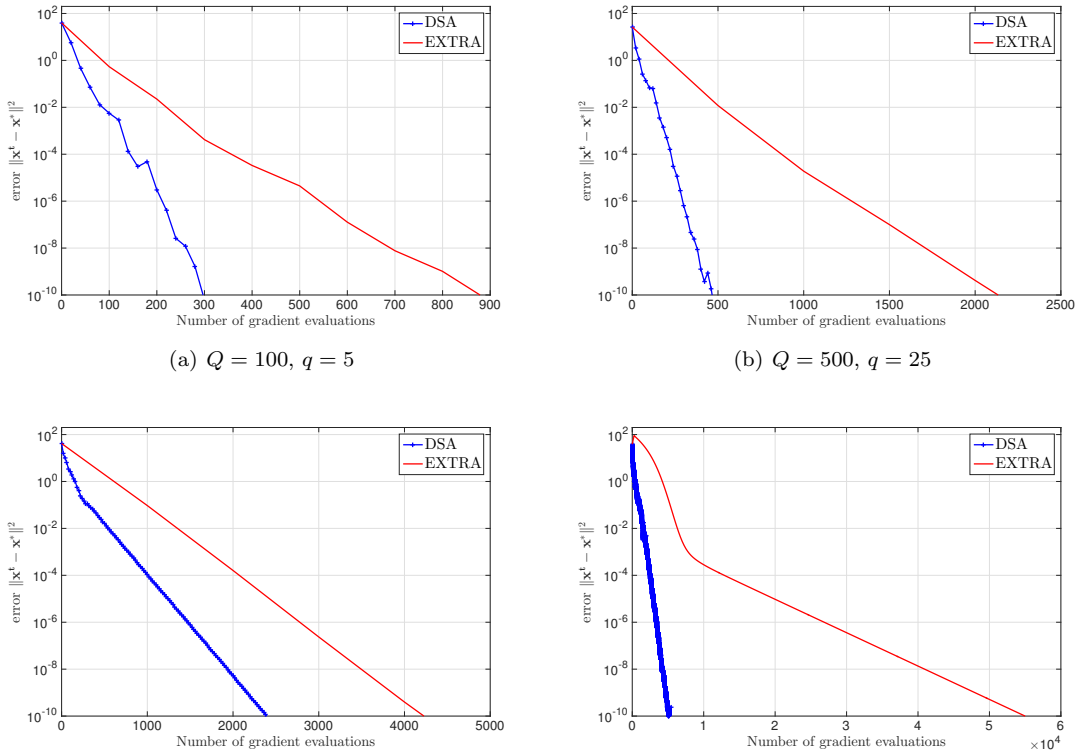
that in the considered graphs, DSA achieves a target accuracy $\|\mathbf{x}^t - \mathbf{x}^*\|^2$ faster than EXTRA. In other words, to achieve a specific accuracy $\|\mathbf{x}^t - \mathbf{x}^*\|^2$ DSA requires less number of local gradient evaluations relative to EXTRA. In addition, the gap between the performance of DSA and EXTRA is more substantial when the graph condition number κ_g is larger. In particular, in the case that we have a complete graph, which has a small graph condition number, the difference between the convergence paths of DSA and EXTRA is less significant comparing to the line graph which has a large graph condition number.

4.3 Effect of Number of Functions (Samples) at Each Node q

To evaluate performance for different number of functions (sample points) available at each node which is indicated by q, we use the same setting as in Fig. 2; however, we consider scenarios with different number of samples Q which leads to different number of samples at each node q. To be more precise, we fix the total number of nodes in the network as $N = 20$ and we consider the cases that the total number of samples are $Q = 100$, $Q = 500$, $Q = 1000$, and $Q = 5000$ where the corresponding number of samples at each node are $q = 5$, $q = 25$, $q = 50$, and $q = 250$, respectively. Similar to the experiment in Fig. 2, the graph is generated randomly with connectivity ratio $p_c = 0.35$.

For each of these scenarios the DSA stepsize α is hand-optimized and the best choice is used for comparison with others. The results are reported for $\alpha = 10^{-4}$, $\alpha = 10^{-3}$, $\alpha = 5 \times 10^{-3}$, and $\alpha = 10^{-1}$ when the total number of samples are $Q = 5000$, $Q = 1000$, $Q = 500$, $Q = 100$, respectively. The resulting convergence paths are shown in Fig. 5.

The convergence paths in Fig. 5 show that as we increase the total number of samples Q and consequently the number of assigned samples to each node q, we observe that DSA converges slower to the optimal argument. This conclusion is expected from the theoretical result in (47) which shows that the linear convergence rate of DSA becomes slower by increasing q. In particular, to achieve the target accuracy of $\|\mathbf{x}^t - \mathbf{x}^*\|^2 = 10^{-8}$ DSA requires $t = 260$, $t = 380$, $t = 1960$, and $t = 4218$



(a) $Q = 100$, $q = 5$

(b) $Q = 500$, $q = 25$

(c) $Q = 1000$, $q = 50$

(d) $Q = 5000$, $q = 250$

Figure 6: Convergence paths of DSA and EXTRA for the cases that $(Q = 100, q = 5)$, $(Q = 500, q = 25)$, $(Q = 1000, q = 50)$, and $(Q = 5000, q = 250)$ are presented. Distance to optimality $e^t = \|\mathbf{x}^t - \mathbf{x}^*\|^2$ is shown with respect to number of gradient evaluations. The total number of nodes in the network is fixed and equal to $N = 20$ and the graph is randomly generated with the connectivity ratio $p_c = 0.35$. DSA converges faster relative to EXTRA and they both converge slower when the total number of samples Q increases.

iterations (or equivalently gradient evaluations) for the cases that $q = 5$, $q = 25$, $q = 50$, $q = 250$, respectively.

To have a more comprehensive comparison of DSA and EXTRA, we also compare their performances under the four different settings considered in Fig. 5. The convergence paths of these methods in terms of number of gradient evaluations for $(Q = 100, q = 5)$, $(Q = 500, q = 25)$, $(Q = 1000, q = 50)$, and $(Q = 5000, q = 250)$ are presented in Fig 6. The optimal stepsizes for EXTRA in the considered settings are $\alpha = 4 \times 10^{-1}$, $\alpha = 5 \times 10^{-2}$, $\alpha = 3 \times 10^{-2}$, and $\alpha = \times 10^{-2}$, respectively. An interesting observation is the effect of q on the convergence rate of EXTRA. We observe that EXTRA converges slower as the number of samples at each node q increases which is identical to the observation for DSA in Fig. 5. Moreover, for all of the settings considered in Fig. 6, DSA outperforms EXTRA in terms of number of required gradient evaluations until convergence. Moreover, by increasing the total number of samples Q and subsequently the number of assigned samples to each node q the advantage of DSA with respect to EXTRA in terms of computational complexity becomes more significant. This observation justifies the use of DSA for large-scale optimization problems as we consider in Section 4.5.

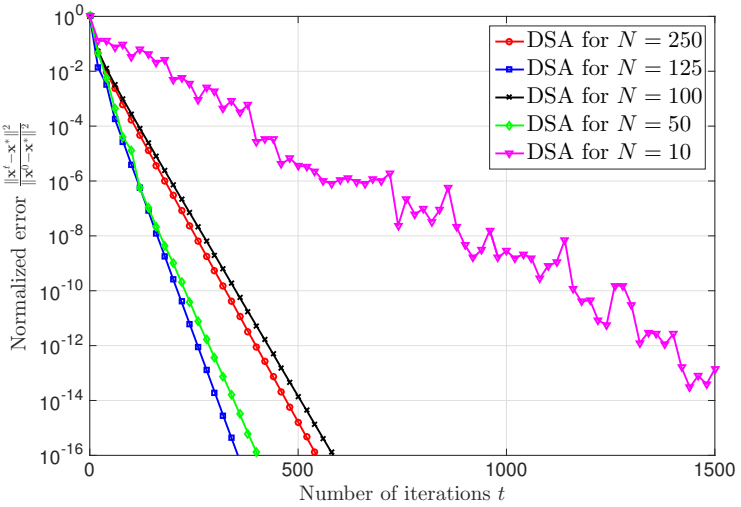
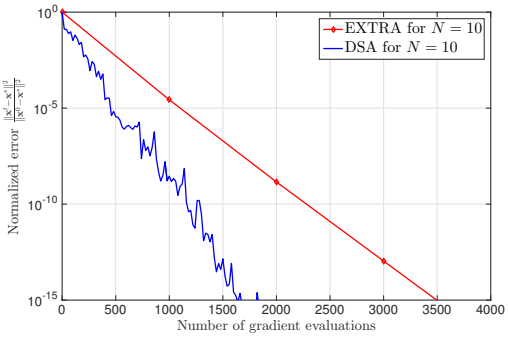


Figure 7: Normalized error $\|\mathbf{x}^t-\mathbf{x}^*\|^2/\|\mathbf{x}^0-\mathbf{x}^*\|^2$ of DSA versus number of iterations t for networks with different number of nodes N when the total number of samples is fixed $Q = 500$. The graphs are randomly generated with the connectivity ratio $p_c = 0.35$. Picking a very small or large value for N which leads to a very large or small value for q, respectively, is not preferable. The best performance belongs to the case that $N = 125$ and $q = 4$.

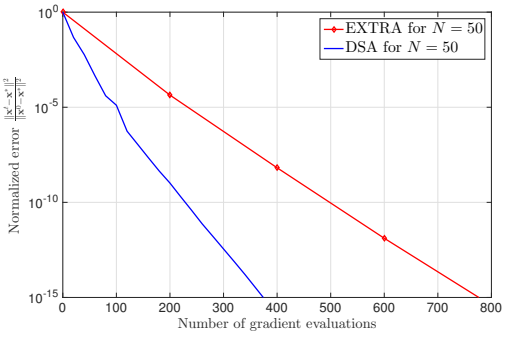
4.4 Effect of Number of Nodes N

In some settings, we can choose the number of nodes (processors) N for training the dataset. In this section, we study the effect of network size N on the convergence path of DSA when a fixed number of samples Q is given to train the classifier $\mathbf{x}$. Notice that when Q is fixed, by changing the number of nodes N, the number of assigned samples to each node $q = Q/N$ changes proportionally. Then, we may want to pick the number of nodes N or equivalently the number of assigned samples to each node q which leads to the best performance of DSA for training Q samples. Hence, we fix the total number of sample points as $Q = 500$ and assign the same amount of sample points q to each node. We consider 5 different settings with $N = 10$, $N = 50$, $N = 100$, $N = 125$, and $N = 250$ which their corresponding number of assigned samples to each node are $q = 50$, $q = 10$, $q = 5$, $q = 4$, and $q = 2$, respectively. The DSA stepsize for each of the considered settings is hand-optimized. The stepsizes $\alpha = 5 \times 10^{-3}$, $\alpha = 2 \times 10^{-2}$, $\alpha = 6 \times 10^{-2}$, and $\alpha = 8 \times 10^{-2}$ are considered for the cases that the number of assigned samples to each node are $q = 50$, $q = 10$, $q = 5$, $q = 4$, and $q = 2$, respectively.

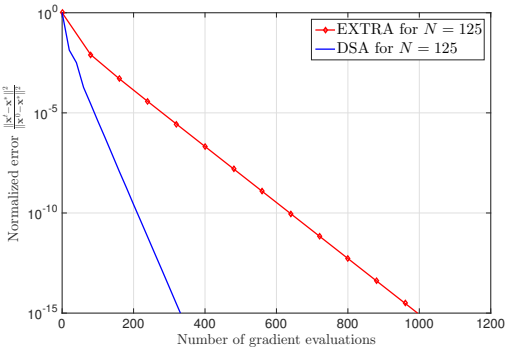
Fig. 7 shows the convergence paths of DSA for networks with different number of nodes. Notice that the normalized error $\tilde{e}^t = \|\mathbf{x}^t-\mathbf{x}^*\|^2/\|\mathbf{x}^0-\mathbf{x}^*\|^2$ is reported, since the dimension of the vector $\mathbf{x}$ is different for different choices of N. Comparison of the convergence paths in Fig. 7 shows that the best performance belongs to the case that $N = 125$ and each node has access to $q = 4$ sample points. The performance of DSA becomes worse for the case that there are $N = 5$ nodes in the network and each node has $q = 100$ sample points. This observation implies that the DSA algorithm is also preferable to SAGA which corresponds to the case that $N = 1$. Moreover, we observe that when the number of nodes is large as $N = 250$ and each node has access to $q = 2$ samples, DSA doesn't perform well. Thus, increasing the size of network N doesn't always lead to a better performance for DSA. The best performance is observed when a moderate subset of the samples is assigned to each node.



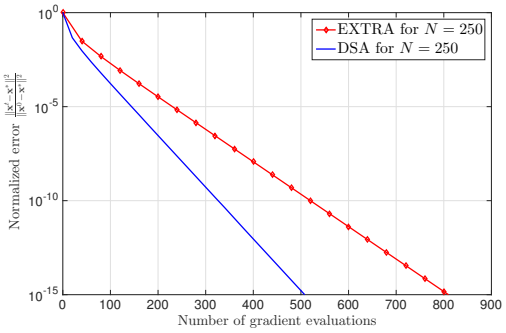
(a) $N = 10$ and $q = 50$



(b) $N = 50$ and $q = 10$



(c) $N = 125$ and $q = 4$



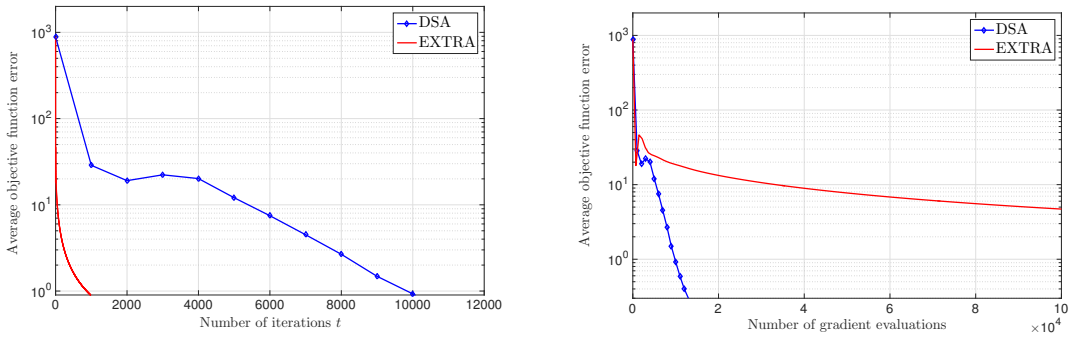
(d) $N = 250$ and $q = 2$

Figure 8: Convergence paths of DSA and EXTRA for different different number of nodes N when the total number of sample points is fixed as $Q = 500$. The graphs are randomly generated with the connectivity ratio $p_c = 0.35$. Normalized distance to optimality $\tilde{e}^t = \|\mathbf{x}^t - \mathbf{x}^*\|^2/\|\mathbf{x}^0 - \mathbf{x}^*\|^2$ is shown with respect to number of gradient evaluations. DSA converges faster relative to EXTRA in all of the considered settings.

We also study the convergence rates of DSA and EXTRA in terms of number of gradient evaluations for networks with different number of nodes N. Fig. 8 demonstrates the convergence paths of DSA and EXTRA for the cases that $N = 10$, $N = 50$, $N = 125$, and $N = 250$. Similar to DSA, we report the best performance of EXTRA for each setting which is achieved by the stepsizes $\alpha = 5 \times 10^{-2}$, $\alpha = 8 \times 10^{-2}$, $\alpha = 8 \times 10^{-2}$, and $\alpha = 10^{-1}$ for $N = 10$, $N = 50$, $N = 125$, and $N = 250$, respectively. Observe that in all settings DSA is more efficient relative to EXTRA and it requires less number of gradient evaluations for convergence.

4.5 Large-scale Classification Application

In this section we solve the logistic regression problem in (48) for the protein homology dataset provided in KDD Cup 2004. The dataset contains $Q = 1.45 \times 10^5$ sample points and each sample point has $p = 74$ features. We consider the case that the sample points are distributed over $N = 200$ nodes which implies that each node has access to $q = 725$ samples. We set the connectivity ratio $p_c = 0.35$ and hand optimize the stepsize α for DSA and EXTRA separately. The best performance of DSA is observed for $\alpha = 2 \times 10^{-7}$ and the best choice of stepize for EXTRA is $\alpha = 6 \times 10^{-7}$. We capture the error in terms of the average objective function error e^t_{avg} of the network which is



(a) (b)

Figure 9: Convergence paths of DSA and EXTRA for the protein homology classification problem with $Q = 1.45 \times 10^5$ samples. The graph has $N = 200$ nodes and it is randomly generated with the connectivity ratio $p_c = 0.35$. The average objective function error is shown with respect to number of iterations t and number of gradient evaluations, respectively.

defined as

$$e_{avg}^t := \frac{1}{N} \sum_{m=1}^{N} \left[\sum_{n=1}^{N} f_n(\mathbf{x}_m^t) - \sum_{n=1}^{N} f_n(\mathbf{x}^*) \right]. \tag{53}$$

Note that the difference $\sum_{n=1}^{N} f_n(\mathbf{x}_m^t) - \sum_{n=1}^{N} f_n(\mathbf{x}^*)$ shows the objective function error associated with the decision variable of node m at time t. Thus, the expression in (53) indicates the average objective function error of the network at step t.

The average objective function error for DSA and EXTRA in terms of number of iterations t and number of gradient evaluations are presented in Fig. 9(a) and Fig. 9(b), respectively. As we observe, the results in Fig. 9 for the large-scale classification problem match the observations in Fig. 2 for the classification problem with a synthetic dataset. In particular, both algorithms converge linearly, while EXTRA converges faster than DSA in terms of number of iterations or equivalently in terms of communication cost. On the other hand, DSA outperforms EXTRA in terms of computational complexity or number of required gradients to reach a target accuracy. Moreover, notice that the difference between the performances of DSA and EXTRA in terms of number of gradient evaluations is more significant in Fig. 9(b) relative to the one in Fig. 2(b). Thus, by increasing the problem dimension we obtain more computational complexity benefit by using DSA instead of EXTRA.

5. Conclusions

Decentralized double stochastic averaging gradient (DSA) is proposed as an algorithm for solving decentralized optimization problems where the local functions can be written as an average of a set of local instantaneous functions. DSA exploits stochastic averaging gradients in lieu of gradients and mixes information of two consecutive iterates to determine the descent direction. By assuming strongly convex local instantaneous functions with Lipschitz continuous gradients, the DSA algorithm converges linearly to the optimal arguments in expectation. In addition, the sequence of local iterates $\mathbf{x}_n^t$ for each node in the network almost surely converges to the optimal argument $\tilde{\mathbf{x}}^*$. A comparison between the DSA algorithm and a group of stochastic and deterministic alternatives are provided for solving a logistic regression problem. The numerical results show DSA is the only stochastic decentralized algorithm to reach linear convergence. DSA outperforms decentralized stochastic alternatives in terms of number of required iteration for convergence, and exhibits faster

convergence relative to deterministic alternatives in terms of number feature vectors processed until convergence.

DSA utilizes the idea of stochastic averaging gradient suggested in SAGA to reduce the computational cost of EXTRA. Although, this modification is successful in reducing the computational complexity of EXTRA and remaining the convergence rate linear, it requires stronger assumptions to prove the linear convergence. In DSA, the local instantaneous functions are required to be strongly convex which is a stricter assumption relative to the required condition for EXTRA that the global objective function should be strongly convex. This assumption for the linear convergence of DSA is inherited from the SAGA algorithm and it can be relaxed by using SVRG (Johnson and Zhang (2013)) instead of SAGA for estimating the gradients of the local functions. This modification in the update of DSA is an obvious extension of the current work and can be considered as a future research direction.

Acknowledgments

We acknowledge the support of the National Science Foundation (NSF CAREER CCF-0952867) and the Office of Naval Research (ONR N00014-12-1-0997).

Appendix A. Proof of Lemma 4

According to the definition of $\hat{\mathbf{g}}^t$ which is the concatenation of the local stochastic averaging gradients $\hat{\mathbf{g}}_n^t$ and the fact that the expected value of sum is equal to the sum of expected values, we can write the expected value $\mathbb{E}\left[\|\hat{\mathbf{g}}^t - \nabla f(\mathbf{x}^*)\|^2 \mid \mathcal{F}^t\right]$ as

$$\mathbb{E}\left[\left\|\hat{\mathbf{g}}^t - \nabla f(\mathbf{x}^*)\right\|^2 \mid \mathcal{F}^t\right] = \sum_{n=1}^{N} \mathbb{E}\left[\left\|\hat{\mathbf{g}}_n^t - \nabla f_n(\tilde{\mathbf{x}}^*)\right\|^2 \mid \mathcal{F}^t\right]. \tag{54}$$

We proceed by finding upper bounds for the summands of (54). Observe that using the standard variance decomposition for any random variable vector $\mathbf{a}$ we can write $\mathbb{E}\left[\|\mathbf{a}\|^2\right] = \|\mathbb{E}[\mathbf{a}]\|^2 + \mathbb{E}\left[\|\mathbf{a} - \mathbb{E}[\mathbf{a}]\|^2\right]$. Notice that the same relation holds true when the expectations are computed with respect to a specific field $\mathcal{F}$. By setting $\mathbf{a} = \hat{\mathbf{g}}_n^t - \nabla f_n(\tilde{\mathbf{x}}^*)$ and considering that $\mathbb{E}[\mathbf{a} \mid \mathcal{F}^t] = \nabla f_n(\mathbf{x}_n^t) - \nabla f_n(\tilde{\mathbf{x}}^*)$, the variance decomposition implies

$$\begin{aligned}\mathbb{E}\left[\left\|\hat{\mathbf{g}}_n^t - \nabla f_n(\tilde{\mathbf{x}}^*)\right\|^2 \mid \mathcal{F}^t\right] &= \left\|\nabla f_n(\mathbf{x}_n^t) - \nabla f_n(\tilde{\mathbf{x}}^*)\right\|^2 \\ &\quad + \mathbb{E}\left[\left\|\hat{\mathbf{g}}_n^t - \nabla f_n(\tilde{\mathbf{x}}^*) - \nabla f_n(\mathbf{x}_n^t) + \nabla f_n(\tilde{\mathbf{x}}^*)\right\|^2 \mid \mathcal{F}^t\right]. \end{aligned}\tag{55}$$

The next step is to find an upper bound for the last term in (55). Adding and subtracting $\nabla f_{n,i_n^t}(\tilde{\mathbf{x}}^*)$ and using the inequality $\|\mathbf{a}+\mathbf{b}\|^2 \leq 2\|\mathbf{a}\|^2 + 2\|\mathbf{b}\|^2$ for $\mathbf{a} = \nabla f_{n,i_n^t}(\mathbf{x}_n^t) - \nabla f_{n,i_n^t}(\tilde{\mathbf{x}}^*) - \nabla f_n(\mathbf{x}_n^t) + \nabla f_n(\tilde{\mathbf{x}}^*)$ and $\mathbf{b} = -(\nabla f_{n,i_n^t}(\mathbf{y}_{n,i_n^t}^t) - \nabla f_{n,i_n^t}(\tilde{\mathbf{x}}^*) - (1/q_n)\sum_{i=1}^{q_n} \nabla f_{n,i}(\mathbf{y}_{n,i}^t) + \nabla f_n(\tilde{\mathbf{x}}^*))$ lead to

$$\begin{aligned}&\mathbb{E}\left[\left\|\hat{\mathbf{g}}_n^t - \nabla f_n(\tilde{\mathbf{x}}^*) - \nabla f_n(\mathbf{x}_n^t) + \nabla f_n(\tilde{\mathbf{x}}^*)\right\|^2 \mid \mathcal{F}^t\right] \\ &\quad \leq 2\mathbb{E}\left[\left\|\nabla f_{n,i_n^t}(\mathbf{x}_n^t) - \nabla f_{n,i_n^t}(\tilde{\mathbf{x}}^*) - \nabla f_n(\mathbf{x}_n^t) + \nabla f_n(\tilde{\mathbf{x}}^*)\right\|^2 \mid \mathcal{F}^t\right] \\ &\qquad + 2\mathbb{E}\left[\left\|\nabla f_{n,i_n^t}(\mathbf{y}_{n,i_n^t}^t) - \nabla f_{n,i_n^t}(\tilde{\mathbf{x}}^*) - \frac{1}{q_n}\sum_{i=1}^{q_n} \nabla f_{n,i}(\mathbf{y}_{n,i}^t) + \nabla f_n(\tilde{\mathbf{x}}^*)\right\|^2 \mid \mathcal{F}^t\right].\end{aligned}\tag{56}$$

In this step we use the standard variance decomposition twice to simplify the two expectations in the right hand side of (56). Based on the standard variance decomposition $\mathbb{E}\left[\|\mathbf{a} - \mathbb{E}[\mathbf{a}]\|^2\right] =$

$\mathbb{E}\left[\|\mathbf{a}\|^2\right] - \|\mathbb{E}\left[\mathbf{a}\right]\|^2$ we obtain $\mathbb{E}\left[\|\mathbf{a} - \mathbb{E}\left[\mathbf{a}\right]\|^2\right] \leq \mathbb{E}\left[\|\mathbf{a}\|^2\right]$. Therefore, by setting $\mathbf{y} = \nabla f_{n,i_n^t}(\mathbf{y}_{n,i_n^t}^t) - \nabla f_{n,i_n^t}(\tilde{\mathbf{x}}^*)$ and observing that the expected value $\mathbb{E}\left[\nabla f_{n,i_n^t}(\mathbf{y}_{n,i_n^t}^t) - \nabla f_{n,i_n^t}(\tilde{\mathbf{x}}^*) \mid \mathcal{F}^t\right]$ is equal to $(1/q_n)\sum_{i=1}^{q_n} \nabla f_{n,i}(\mathbf{y}_{n,i}^t) - \nabla f_n(\tilde{\mathbf{x}}^*)$ we obtain that

$$\mathbb{E}\left[\left\|\nabla f_{n,i_n^t}(\mathbf{y}_{n,i_n^t}^t) - \nabla f_{n,i_n^t}(\tilde{\mathbf{x}}^*) - \frac{1}{q_n}\sum_{i=1}^{q_n} \nabla f_{n,i}(\mathbf{y}_{n,i}^t) + \nabla f_n(\tilde{\mathbf{x}}^*)\right\|^2 \mid \mathcal{F}^t\right] \leq \mathbb{E}\left[\left\|\nabla f_{n,i_n^t}(\mathbf{y}_{n,i_n^t}^t) - \nabla f_{n,i_n^t}(\tilde{\mathbf{x}}^*)\right\|^2 \mid \mathcal{F}^t\right]. \tag{57}$$

Moreover, by choosing $\mathbf{a} = \nabla f_{n,i_n^t}(\mathbf{x}_n^t) - \nabla f_{n,i_n^t}(\tilde{\mathbf{x}}^*)$ and noticing the relation for the expected value which is $\mathbb{E}\left[\nabla f_{n,i_n^t}(\mathbf{x}_n^t) - \nabla f_{n,i_n^t}(\tilde{\mathbf{x}}^*) \mid \mathcal{F}^t\right] = \nabla f_n(\mathbf{x}_n^t) - \nabla f_n(\tilde{\mathbf{x}}^*)$, the equality $\mathbb{E}\left[\|\mathbf{a} - \mathbb{E}\left[\mathbf{a}\right]\|^2\right] = \mathbb{E}\left[\|\mathbf{a}\|^2\right] - \|\mathbb{E}\left[\mathbf{a}\right]\|^2$ yields

$$\mathbb{E}\left[\left\|\nabla f_{n,i_n^t}(\mathbf{x}_n^t) - \nabla f_{n,i_n^t}(\tilde{\mathbf{x}}^*) - \nabla f_n(\mathbf{x}_n^t) + \nabla f_n(\tilde{\mathbf{x}}^*)\right\|^2 \mid \mathcal{F}^t\right] = \mathbb{E}\left[\left\|\nabla f_{n,i_n^t}(\mathbf{x}_n^t) - \nabla f_{n,i_n^t}(\tilde{\mathbf{x}}^*)\right\|^2 \mid \mathcal{F}^t\right] - \left\|\nabla f_n(\mathbf{x}_n^t) - \nabla f_n(\tilde{\mathbf{x}}^*)\right\|^2. \tag{58}$$

By substituting the upper bound in (57) and the simplification in (58) into (56), and considering the expression in (55) we obtain that

$$\mathbb{E}\left[\left\|\hat{\mathbf{g}}^t - \nabla f(\mathbf{x}^*)\right\|^2 \mid \mathcal{F}^t\right] \leq 2\sum_{n=1}^{N} \mathbb{E}\left[\left\|\nabla f_{n,i_n^t}(\mathbf{y}_{n,i_n^t}^t) - \nabla f_{n,i_n^t}(\tilde{\mathbf{x}}^*)\right\|^2 \mid \mathcal{F}^t\right] - \sum_{n=1}^{N} \left\|\nabla f_n(\mathbf{x}_n^t) - \nabla f_n(\tilde{\mathbf{x}}^*)\right\|^2 + 2\sum_{n=1}^{N} \mathbb{E}\left[\left\|\nabla f_{n,i_n^t}(\mathbf{x}_n^t) - \nabla f_{n,i_n^t}(\tilde{\mathbf{x}}^*)\right\|^2 \mid \mathcal{F}^t\right]. \tag{59}$$

We proceed by finding an upper bound for the first sum in the right hand side of (59). Notice that if the gradients of the function g are Lipschitz continuous with parameter L, then for any two vectors $\mathbf{a}_1$ and $\mathbf{a}_2$ we can write $g(\mathbf{a}_1) \geq g(\mathbf{a}_2) + \nabla g(\mathbf{a}_2)^T(\mathbf{a}_1 - \mathbf{a}_2) + (1/2L)\|\nabla g(\mathbf{a}_1) - \nabla g(\mathbf{a}_2)\|^2$. According to the Lipschitz continuity of the instantaneous local functions gradient $\nabla f_{n,i}(\mathbf{x}_n)$, we can write the inequality for $g = f_{n,i}$, $\mathbf{a}_1 = \mathbf{y}_{n,i}^t$ and $\mathbf{a}_2 = \tilde{\mathbf{x}}^*$ which is equivalent to

$$\frac{1}{2L}\left\|\nabla f_{n,i}(\mathbf{y}_{n,i}^t) - \nabla f_{n,i}(\tilde{\mathbf{x}}^*)\right\|^2 \leq f_{n,i}(\mathbf{y}_{n,i}^t) - f_{n,i}(\tilde{\mathbf{x}}^*) - \nabla f_{n,i}(\tilde{\mathbf{x}}^*)^T(\mathbf{y}_{n,i}^t - \tilde{\mathbf{x}}^*). \tag{60}$$

Summing up both sides of (60) for all $i = 1, \ldots, q_n$ and dividing both sides of the implied inequality by q_n yield

$$\frac{1}{q_n}\sum_{i=1}^{q_n}\left\|\nabla f_{n,i}(\mathbf{y}_{n,i}^t) - \nabla f_{n,i}(\tilde{\mathbf{x}}^*)\right\|^2 \leq 2L\left[\frac{1}{q_n}\sum_{i=1}^{q_n} f_{n,i}(\mathbf{y}_{n,i}^t) - f_{n,i}(\tilde{\mathbf{x}}^*) - \nabla f_{n,i}(\tilde{\mathbf{x}}^*)^T(\mathbf{y}_{n,i}^t - \tilde{\mathbf{x}}^*)\right]. \tag{61}$$

Since the random functions f_{n,i_n^t} has a uniform distribution over the set $\{f_{n,1}, \ldots, f_{n,q_n}\}$, we can substitute the left hand side of (61) by $\mathbb{E}\left[\left\|\nabla f_{n,i_n^t}(\mathbf{y}_{n,i_n^t}^t) - \nabla f_{n,i_n^t}(\tilde{\mathbf{x}}^*)\right\|^2 \mid \mathcal{F}^t\right]$. Apply this substitution and sum up both sides of (61) for $n = 1, \ldots, N$. According to the definition of sequence p^t in (33), if we sum up the right hand side of (61) over n it can be simplified as $2Lp^t$. Applying these simplifications we obtain

$$\sum_{n=1}^{N} \mathbb{E}\left[\left\|\nabla f_{n,\boldsymbol{\theta}_n^t}(\mathbf{y}_n^t) - \nabla f_{n,\boldsymbol{\theta}_n^t}(\tilde{\mathbf{x}}^*)\right\|^2 \mid \mathcal{F}^t\right] \leq 2Lp^t. \tag{62}$$

Substituting the upper bound in (62) into (59) and simplifying the sum $\sum_{n=1}^{N} \|\nabla f_n(\mathbf{x}_n^t) - \nabla f_n(\tilde{\mathbf{x}}^*)\|^2$ as $\|\nabla f(\mathbf{x}^t) - \nabla f(\mathbf{x}^*)\|^2$ yield

$$\mathbb{E}\left[\left\|\hat{\mathbf{g}}^t - \nabla f(\mathbf{x}^*)\right\|^2 \mid \mathcal{F}^t\right] \leq 2\sum_{n=1}^{N} \mathbb{E}\left[\left\|\nabla f_{n,i_n^t}(\mathbf{x}_n^t) - \nabla f_{n,i_n^t}(\tilde{\mathbf{x}}^*)\right\|^2 \mid \mathcal{F}^t\right] - \left\|\nabla f(\mathbf{x}^t) - \nabla f(\mathbf{x}^*)\right\|^2 + 4Lp^t. \tag{63}$$

To show that the sum in the right hand side of (63) is bounded above we use the Lipschitz continuity of the instantaneous functions gradients $\nabla f_{n,i}$. Using the same argument from (60) to (62) we can write

$$\sum_{n=1}^{N} \mathbb{E}\left[\left\|\nabla f_{n,i_n^t}(\mathbf{x}_n^t) - \nabla f_{n,i_n^t}(\tilde{\mathbf{x}}^*)\right\|^2 \mid \mathcal{F}^t\right] \leq 2L\sum_{n=1}^{N} \frac{1}{q_n}\left[\sum_{i=1}^{q_n} f_{n,i}(\mathbf{x}_n^t) - f_{n,i}(\tilde{\mathbf{x}}^*) - \nabla f_{n,i}(\tilde{\mathbf{x}}^*)^T(\mathbf{x}_n^t - \tilde{\mathbf{x}}^*)\right]. \tag{64}$$

Considering the definition of the local objective functions $f_n(\mathbf{x}_n) = (1/q_n)\sum_{i=1}^{q_n} f_{n,i}(\mathbf{x}_n)$ and the aggregate function $f(\mathbf{x}) := \sum_{n=1}^{N} f_n(\mathbf{x}_n)$, the right hand side of (64) can be simplified as

$$\sum_{n=1}^{N} \mathbb{E}\left[\left\|\nabla f_{n,i_n^t}(\mathbf{x}_n^t) - \nabla f_{n,i_n^t}(\tilde{\mathbf{x}}^*)\right\|^2 \mid \mathcal{F}^t\right] \leq 2L\left(f(\mathbf{x}^t) - f(\mathbf{x}^*) - \nabla f(\mathbf{x}^*)^T(\mathbf{x}^t - \mathbf{x}^*)\right). \tag{65}$$

Replacing the sum in (63) by the upper bound in (65) implies

$$\mathbb{E}\left[\left\|\hat{\mathbf{g}}^t - \nabla f(\mathbf{x}^*)\right\|^2 \mid \mathcal{F}^t\right] \leq 4Lp^t - \left\|\nabla f(\mathbf{x}^t) - \nabla f(\mathbf{x}^*)\right\|^2 + 4L\left(f(\mathbf{x}^t) - f(\mathbf{x}^*) - \nabla f(\mathbf{x}^*)^T(\mathbf{x}^t - \mathbf{x}^*)\right). \tag{66}$$

Considering the strong convexity of the global objective function f with constant μ we can write

$$\left\|\nabla f(\mathbf{x}^t) - \nabla f(\mathbf{x}^*)\right\|^2 \geq 2\mu\left(f(\mathbf{x}^t) - f(\mathbf{x}^*) - \nabla f(\mathbf{x}^*)^T(\mathbf{x}^t - \mathbf{x}^*)\right). \tag{67}$$

Therefore, we can substitute $\|\nabla f(\mathbf{x}^t) - \nabla f(\mathbf{x}^*)\|^2$ in (65) by the lower bound in (67) and the claim in (34) follows.

Appendix B. Proof of Lemma 5

According to the Lipschitz continuity of the aggregate function gradients $\nabla f(\mathbf{x})$, we can write $(1/L)\|\nabla f(\mathbf{x}^t) - \nabla f(\mathbf{x}^*)\|^2 \leq (\mathbf{x}^t - \mathbf{x}^*)^T(\nabla f(\mathbf{x}^t) - \nabla f(\mathbf{x}^*))$. By adding and subtracting $\mathbf{x}^{t+1}$ to the term $\mathbf{x}^t - \mathbf{x}^*$ and multiplying both sides of the inequality by 2α we obtain

$$\frac{2\alpha}{L}\left\|\nabla f(\mathbf{x}^t) - \nabla f(\mathbf{x}^*)\right\|^2 \leq 2\alpha(\mathbf{x}^{t+1} - \mathbf{x}^*)^T(\nabla f(\mathbf{x}^t) - \nabla f(\mathbf{x}^*)) + 2\alpha(\mathbf{x}^t - \mathbf{x}^{t+1})^T(\nabla f(\mathbf{x}^t) - \nabla f(\mathbf{x}^*)). \tag{68}$$

Expanding the difference $\nabla f(\mathbf{x}^t) - \nabla f(\mathbf{x}^*)$ as $\hat{\mathbf{g}}^t - \nabla f(\mathbf{x}^*) + \nabla f(\mathbf{x}^t) - \hat{\mathbf{g}}^t$ for the first inner product in the right hand side of (68) implies

$$\frac{2\alpha}{L}\left\|\nabla f(\mathbf{x}^t) - \nabla f(\mathbf{x}^*)\right\|^2 \leq 2\alpha(\mathbf{x}^t - \mathbf{x}^{t+1})^T(\nabla f(\mathbf{x}^t) - \nabla f(\mathbf{x}^*)) + 2\alpha(\mathbf{x}^{t+1} - \mathbf{x}^*)^T(\hat{\mathbf{g}}^t - \nabla f(\mathbf{x}^*)) + 2\alpha(\mathbf{x}^{t+1} - \mathbf{x}^*)^T(\nabla f(\mathbf{x}^t) - \hat{\mathbf{g}}^t). \tag{69}$$

We proceed to simplify the inner product $2\alpha(\mathbf{x}^{t+1}-\mathbf{x}^*)^T(\hat{\mathbf{g}}^t-\nabla f(\mathbf{x}^*))$ in the right hand side of (69) by substituting $\alpha(\hat{\mathbf{g}}^t-\nabla f(\mathbf{x}^*))$ with its equivalent as introduced in (30). By applying this substitution the inner product $2\alpha(\mathbf{x}^{t+1}-\mathbf{x}^*)^T(\hat{\mathbf{g}}^t-\nabla f(\mathbf{x}^*))$ can be simplified as

$$2\alpha(\mathbf{x}^{t+1}-\mathbf{x}^*)^T(\hat{\mathbf{g}}^t-\nabla f(\mathbf{x}^*)) = -2\|\mathbf{x}^{t+1}-\mathbf{x}^*\|^2_{\mathbf{I}+\mathbf{Z}-2\tilde{\mathbf{Z}}} + 2(\mathbf{x}^{t+1}-\mathbf{x}^*)^T\tilde{\mathbf{Z}}(\mathbf{x}^t-\mathbf{x}^{t+1}) - 2(\mathbf{x}^{t+1}-\mathbf{x}^*)^T\mathbf{U}(\mathbf{v}^{t+1}-\mathbf{v}^*). \tag{70}$$

Based on the KKT condition of problem (25), the optimal primal variable satisfies $(\tilde{\mathbf{Z}}-\mathbf{Z})^{1/2}\mathbf{x}^*=\mathbf{0}$ which by considering the definition of the matrix $\mathbf{U}=(\tilde{\mathbf{Z}}-\mathbf{Z})^{1/2}$ we obtain that $\mathbf{U}\mathbf{x}^*=\mathbf{0}$. This observation in conjunction with the update rule of the dual variable $\mathbf{v}^t$ in (28) implies that we can substitute $\mathbf{U}(\mathbf{x}^{t+1}-\mathbf{x}^*)$ by $\mathbf{v}^{t+1}-\mathbf{v}^t$. Making this substitution into the last summand of the right hand side of (70) and considering the symmetry of the matrix $\mathbf{U}$ yield

$$2\alpha(\mathbf{x}^{t+1}-\mathbf{x}^*)^T(\hat{\mathbf{g}}^t-\nabla f(\mathbf{x}^*)) = -2\|\mathbf{x}^{t+1}-\mathbf{x}^*\|^2_{\mathbf{I}+\mathbf{Z}-2\tilde{\mathbf{Z}}} + 2(\mathbf{x}^{t+1}-\mathbf{x}^*)^T\tilde{\mathbf{Z}}(\mathbf{x}^t-\mathbf{x}^{t+1}) - 2(\mathbf{v}^{t+1}-\mathbf{v}^t)^T(\mathbf{v}^{t+1}-\mathbf{v}^*). \tag{71}$$

According to the definition of the vector $\mathbf{u}$ and matrix $\mathbf{G}$ in (35), the last two summands of (71) can be simplified as $2(\mathbf{u}^{t+1}-\mathbf{u}^t)^T\mathbf{G}(\mathbf{u}^*-\mathbf{u}^{t+1})$. Moreover, observe that the inner product $2(\mathbf{u}^{t+1}-\mathbf{u}^t)^T\mathbf{G}(\mathbf{u}^*-\mathbf{u}^{t+1})$ can be simplified as $\|\mathbf{u}^t-\mathbf{u}^*\|^2_{\mathbf{G}}-\|\mathbf{u}^{t+1}-\mathbf{u}^*\|^2_{\mathbf{G}}-\|\mathbf{u}^{t+1}-\mathbf{u}^t\|^2_{\mathbf{G}}$. Applying this simplification into (71) implies

$$2\alpha(\mathbf{x}^{t+1}-\mathbf{x}^*)^T(\hat{\mathbf{g}}^t-\nabla f(\mathbf{x}^*)) = -2\|\mathbf{x}^{t+1}-\mathbf{x}^*\|^2_{\mathbf{I}+\mathbf{Z}-2\tilde{\mathbf{Z}}} + \|\mathbf{u}^t-\mathbf{u}^*\|^2_{\mathbf{G}} - \|\mathbf{u}^{t+1}-\mathbf{u}^*\|^2_{\mathbf{G}} - \|\mathbf{u}^{t+1}-\mathbf{u}^t\|^2_{\mathbf{G}}. \tag{72}$$

The next step is to find an upper bound for the inner product $2\alpha(\mathbf{x}^t-\mathbf{x}^{t+1})^T(\nabla f(\mathbf{x}^t)-\nabla f(\mathbf{x}^*))$. Note that for any two vectors $\mathbf{a}$ and $\mathbf{b}$, and any positive scalar η the inequality $2\mathbf{a}^T\mathbf{b}\leq\eta\|\mathbf{a}\|^2+\eta^{-1}\|\mathbf{b}\|^2$ holds. Thus, by setting $\mathbf{a}=\mathbf{x}^t-\mathbf{x}^{t+1}$ and $\mathbf{b}=\nabla f(\mathbf{x}^t)-\nabla f(\mathbf{x}^*)$ we obtain that

$$2\alpha(\mathbf{x}^t-\mathbf{x}^{t+1})^T(\nabla f(\mathbf{x}^t)-\nabla f(\mathbf{x}^*)) \leq \frac{\alpha}{\eta}\|\nabla f(\mathbf{x}^t)-\nabla f(\mathbf{x}^*)\|^2+\alpha\eta\|\mathbf{x}^t-\mathbf{x}^{t+1}\|^2. \tag{73}$$

Now we substitute the terms in the right hand side of (69) by their simplifications or upper bounds. Replacing the inner product $2\alpha(\mathbf{x}^{t+1}-\mathbf{x}^*)^T(\hat{\mathbf{g}}^t-\nabla f(\mathbf{x}^*))$ by the simplification in (72), substituting expression $2\alpha(\mathbf{x}^t-\mathbf{x}^{t+1})^T(\nabla f(\mathbf{x}^t)-\nabla f(\mathbf{x}^*))$ by the upper bound in (73), and substituting inner product $2\alpha(\mathbf{x}^{t+1}-\mathbf{x}^*)^T(\nabla f(\mathbf{x}^t)-\hat{\mathbf{g}}^t)$ by the sum $2\alpha(\mathbf{x}^t-\mathbf{x}^*)^T(\nabla f(\mathbf{x}^t)-\hat{\mathbf{g}}^t)+2\alpha(\mathbf{x}^{t+1}-\mathbf{x}^t)^T(\nabla f(\mathbf{x}^t)-\hat{\mathbf{g}}^t)$ imply

$$\begin{aligned}\frac{2\alpha}{L}\left\|\nabla f(\mathbf{x}^t)-\nabla f(\mathbf{x}^*)\right\|^2 \leq{}& -2\|\mathbf{x}^{t+1}-\mathbf{x}^*\|^2_{\mathbf{I}+\mathbf{Z}-2\tilde{\mathbf{Z}}} + \|\mathbf{u}^t-\mathbf{u}^*\|^2_{\mathbf{G}} - \|\mathbf{u}^{t+1}-\mathbf{u}^*\|^2_{\mathbf{G}} \\ &- \|\mathbf{u}^{t+1}-\mathbf{u}^t\|^2_{\mathbf{G}} + \alpha\eta\|\mathbf{x}^t-\mathbf{x}^{t+1}\|^2 + \frac{\alpha}{\eta}\|\nabla f(\mathbf{x}^t)-\nabla f(\mathbf{x}^*)\|^2 \\ &+ 2\alpha(\mathbf{x}^t-\mathbf{x}^*)^T(\nabla f(\mathbf{x}^t)-\hat{\mathbf{g}}^t) + 2\alpha(\mathbf{x}^{t+1}-\mathbf{x}^t)^T(\nabla f(\mathbf{x}^t)-\hat{\mathbf{g}}^t).\end{aligned} \tag{74}$$

Considering that $\mathbf{x}^t-\mathbf{x}^*$ is deterministic given observations until step t and observing the relation $\mathbb{E}\left[\hat{\mathbf{g}}^t \mid \mathcal{F}^t\right]=\nabla f(\mathbf{x}^t)$, we obtain that $\mathbb{E}\left[(\mathbf{x}^t-\mathbf{x}^*)^T(\nabla f(\mathbf{x}^t)-\hat{\mathbf{g}}^t) \mid \mathcal{F}^t\right]=0$. Therefore, by computing the expected value of both sides of (74) given the observations until step t and regrouping the terms we obtain

$$\begin{aligned}\|\mathbf{u}^t-\mathbf{u}^*\|^2_{\mathbf{G}} - \mathbb{E}\left[\|\mathbf{u}^{t+1}-\mathbf{u}^*\|^2_{\mathbf{G}} \mid \mathcal{F}^t\right] \geq{}& \alpha\left(\frac{2}{L}-\frac{1}{\eta}\right)\left\|\nabla f(\mathbf{x}^t)-\nabla f(\mathbf{x}^*)\right\|^2 + \mathbb{E}\left[\|\mathbf{u}^{t+1}-\mathbf{u}^t\|^2_{\mathbf{G}} \mid \mathcal{F}^t\right] \\ &+ 2\mathbb{E}\left[\|\mathbf{x}^{t+1}-\mathbf{x}^*\|^2_{\mathbf{I}+\mathbf{Z}-2\tilde{\mathbf{Z}}} \mid \mathcal{F}^t\right] - \alpha\eta\mathbb{E}\left[\|\mathbf{x}^t-\mathbf{x}^{t+1}\|^2 \mid \mathcal{F}^t\right] \\ &- \mathbb{E}\left[2\alpha(\mathbf{x}^{t+1}-\mathbf{x}^t)^T(\nabla f(\mathbf{x}^t)-\hat{\mathbf{g}}^t) \mid \mathcal{F}^t\right].\end{aligned} \tag{75}$$

By applying inequality $2\mathbf{a}^T\mathbf{b} \leq \eta\|\mathbf{a}\|^2 + \eta^{-1}\|\mathbf{b}\|^2$ for the vectors $\mathbf{a} = \mathbf{x}^{t+1} - \mathbf{x}^t$ and $\mathbf{b} = \nabla f(\mathbf{x}^t) - \hat{\mathbf{g}}^t$, we obtain that the inner product $2(\mathbf{x}^{t+1} - \mathbf{x}^t)^T(\nabla f(\mathbf{x}^t) - \hat{\mathbf{g}}^t)$ is bounded above by $\eta\|\mathbf{x}^{t+1} - \mathbf{x}^t\|^2 + (1/\eta)\|\nabla f(\mathbf{x}^t) - \hat{\mathbf{g}}^t\|^2$. Replacing $2(\mathbf{x}^{t+1} - \mathbf{x}^t)^T(\nabla f(\mathbf{x}^t) - \hat{\mathbf{g}}^t)$ in (75) by its upper bound $\eta\|\mathbf{x}^{t+1} - \mathbf{x}^t\|^2 + (1/\eta)\|\nabla f(\mathbf{x}^t) - \hat{\mathbf{g}}^t\|^2$ yields

$$\begin{aligned}\|\mathbf{u}^t - \mathbf{u}^*\|_{\mathbf{G}}^2 - \mathbb{E}\left[\|\mathbf{u}^{t+1} - \mathbf{u}^*\|_{\mathbf{G}}^2 \mid \mathcal{F}^t\right] \geq \alpha\left(\frac{2}{L} - \frac{1}{\eta}\right)\left\|\nabla f(\mathbf{x}^t) - \nabla f(\mathbf{x}^*)\right\|^2 + \mathbb{E}\left[\|\mathbf{u}^{t+1} - \mathbf{u}^t\|_{\mathbf{G}}^2 \mid \mathcal{F}^t\right] \\ + 2\mathbb{E}\left[\|\mathbf{x}^{t+1} - \mathbf{x}^*\|_{\mathbf{I}+\mathbf{Z}-2\tilde{\mathbf{Z}}}^2 \mid \mathcal{F}^t\right] - 2\alpha\eta\mathbb{E}\left[\|\mathbf{x}^t - \mathbf{x}^{t+1}\|^2 \mid \mathcal{F}^t\right] \\ - \frac{\alpha}{\eta}\mathbb{E}\left[\|\nabla f(\mathbf{x}^t) - \hat{\mathbf{g}}^t\|^2 \mid \mathcal{F}^t\right]. \qquad (76)\end{aligned}$$

Observe that the squared norm $\|\mathbf{u}^{t+1} - \mathbf{u}^t\|_{\mathbf{G}}^2$ can be expanded as $\|\mathbf{x}^{t+1} - \mathbf{x}^t\|_{\tilde{\mathbf{Z}}}^2 + \|\mathbf{v}^{t+1} - \mathbf{v}^t\|^2$. Using this simplification for $\|\mathbf{u}^{t+1} - \mathbf{u}^t\|_{\mathbf{G}}^2$ and regrouping the terms in (76) lead to

$$\begin{aligned}\|\mathbf{u}^t - \mathbf{u}^*\|_{\mathbf{G}}^2 - \mathbb{E}\left[\|\mathbf{u}^{t+1} - \mathbf{u}^*\|_{\mathbf{G}}^2 \mid \mathcal{F}^t\right] \geq \alpha\left(\frac{2}{L} - \frac{1}{\eta}\right)\left\|\nabla f(\mathbf{x}^t) - \nabla f(\mathbf{x}^*)\right\|^2 \qquad (77) \\ + \mathbb{E}\left[\|\mathbf{x}^{t+1} - \mathbf{x}^t\|_{\tilde{\mathbf{Z}}-2\alpha\eta\mathbf{I}}^2 \mid \mathcal{F}^t\right] + \mathbb{E}\left[\|\mathbf{v}^{t+1} - \mathbf{v}^t\|^2 \mid \mathcal{F}^t\right] \\ + 2\mathbb{E}\left[\|\mathbf{x}^{t+1} - \mathbf{x}^*\|_{\mathbf{I}+\mathbf{Z}-2\tilde{\mathbf{Z}}}^2 \,|\, \mathcal{F}^t\right] - \frac{\alpha}{\eta}\mathbb{E}\left[\|\nabla f(\mathbf{x}^t) - \hat{\mathbf{g}}^t\|^2 \,|\, \mathcal{F}^t\right].\end{aligned}$$

We proceed by simplifying the expectation $\mathbb{E}\left[\|\nabla f(\mathbf{x}^t) - \hat{\mathbf{g}}^t\|^2 \,|\, \mathcal{F}^t\right]$ in (77). Note that by adding and subtracting $\nabla f(\mathbf{x}^*)$, the expectation can be written as $\mathbb{E}\left[\|\nabla f(\mathbf{x}^t) - \nabla f(\mathbf{x}^*) + \nabla f(\mathbf{x}^*) - \hat{\mathbf{g}}^t\|^2 \mid \mathcal{F}^t\right]$ and by expanding the squared norm and simplifying the terms we obtain

$$\mathbb{E}\left[\left\|\nabla f(\mathbf{x}^t) - \hat{\mathbf{g}}^t\right\|^2 \mid \mathcal{F}^t\right] = \mathbb{E}\left[\left\|\hat{\mathbf{g}}^t - \nabla f(\mathbf{x}^*)\right\|^2 \mid \mathcal{F}^t\right] - \mathbb{E}\left[\left\|\nabla f(\mathbf{x}^t) - \nabla f(\mathbf{x}^*)\right\|^2 \mid \mathcal{F}^t\right]. \qquad (78)$$

Substituting the simplification in (78) into (77) yields

$$\begin{aligned}\|\mathbf{u}^t - \mathbf{u}^*\|_{\mathbf{G}}^2 - \mathbb{E}\left[\|\mathbf{u}^{t+1} - \mathbf{u}^*\|_{\mathbf{G}}^2 \mid \mathcal{F}^t\right] \geq \frac{2\alpha}{L}\left\|\nabla f(\mathbf{x}^t) - \nabla f(\mathbf{x}^*)\right\|^2 \qquad (79) \\ + \mathbb{E}\left[\|\mathbf{x}^{t+1} - \mathbf{x}^t\|_{\tilde{\mathbf{Z}}-2\alpha\eta\mathbf{I}}^2 \mid \mathcal{F}^t\right] + \mathbb{E}\left[\|\mathbf{v}^{t+1} - \mathbf{v}^t\|^2 \mid \mathcal{F}^t\right] \\ + 2\mathbb{E}\left[\|\mathbf{x}^{t+1} - \mathbf{x}^*\|_{\mathbf{I}+\mathbf{Z}-2\tilde{\mathbf{Z}}}^2 \,|\, \mathcal{F}^t\right] - \frac{\alpha}{\eta}\mathbb{E}\left[\|\hat{\mathbf{g}}^t - \nabla f(\mathbf{x}^*)\|^2 \mid \mathcal{F}^t\right].\end{aligned}$$

Considering the strong convexity of the global objective function f with constant μ we can write $\|\nabla f(\mathbf{x}^t) - \nabla f(\mathbf{x}^*)\|^2 \geq 2\mu\left(f(\mathbf{x}^t) - f(\mathbf{x}^*) - \nabla f(\mathbf{x}^*)^T(\mathbf{x}^t - \mathbf{x}^*)\right)$. Substituting the squared norm $\|\nabla f(\mathbf{x}^t) - \nabla f(\mathbf{x}^*)\|^2$ by this lower bound in (79) follows

$$\begin{aligned}\|\mathbf{u}^t - \mathbf{u}^*\|_{\mathbf{G}}^2 - \mathbb{E}\left[\|\mathbf{u}^{t+1} - \mathbf{u}^*\|_{\mathbf{G}}^2 \mid \mathcal{F}^t\right] \geq \frac{4\alpha\mu}{L}\left(f(\mathbf{x}^t) - f(\mathbf{x}^*) - \nabla f(\mathbf{x}^*)^T(\mathbf{x}^t - \mathbf{x}^*)\right) \qquad (80) \\ + \mathbb{E}\left[\|\mathbf{x}^{t+1} - \mathbf{x}^t\|_{\tilde{\mathbf{Z}}-2\alpha\eta\mathbf{I}}^2 \mid \mathcal{F}^t\right] + \mathbb{E}\left[\|\mathbf{v}^{t+1} - \mathbf{v}^t\|^2 \mid \mathcal{F}^t\right] \\ + 2\mathbb{E}\left[\|\mathbf{x}^{t+1} - \mathbf{x}^*\|_{\mathbf{I}+\mathbf{Z}-2\tilde{\mathbf{Z}}}^2 \,|\, \mathcal{F}^t\right] - \frac{\alpha}{\eta}\mathbb{E}\left[\|\hat{\mathbf{g}}^t - \nabla f(\mathbf{x}^*)\|^2 \mid \mathcal{F}^t\right].\end{aligned}$$

Substituting the upper bound for the expectation $\mathbb{E}\left[\|\hat{\mathbf{g}}^t - \nabla f(\mathbf{x}^*)\|^2 \mid \mathcal{F}^t\right]$ in (34) into (80) and regrouping the terms show the validity of the claim in (36).

Appendix C. Proof of Lemma 6

Given the information until time t, each auxiliary vector $\mathbf{y}_{n,i}^{t+1}$ is a random variable that takes values $\mathbf{y}_{n,i}^t$ and $\mathbf{x}_n^t$ with associated probabilities $1 - 1/q_n$ and $1/q_n$, respectively. This observation holds

since with probability $1/q_n$ node n may choose index i to update at time $t+1$ and with probability $1-(1/q_n)$ choose other indices. Therefore, we can write

$$\mathbb{E}\left[\frac{1}{q_n}\sum_{i=1}^{q_n}\left(\nabla f_{n,i}(\tilde{\mathbf{x}}^*)^T(\mathbf{y}_{n,i}^{t+1}-\tilde{\mathbf{x}}^*)\right)\mid\mathcal{F}^t\right]=\left[1-\frac{1}{q_n}\right]\frac{1}{q_n}\sum_{i=1}^{q_n}\nabla f_{n,i}(\tilde{\mathbf{x}}^*)^T(\mathbf{y}_{n,i}^t-\tilde{\mathbf{x}}^*)+\frac{1}{q_n}\nabla f_n(\tilde{\mathbf{x}}^*)^T(\mathbf{x}_n^t-\tilde{\mathbf{x}}^*). \tag{81}$$

Likewise, the distribution of random function $f_{n,i}(\mathbf{y}_{n,i}^{t+1})$ given observation until time t has two possibilities $f_{n,i}(\mathbf{y}_{n,i}^t)$ and $f_{n,i}(\mathbf{x}_n^t)$ with associated probabilities $1-1/q_n$ and $1/q_n$, respectively. Hence, we can write $\mathbb{E}\left[f_{n,i}(\mathbf{y}_{n,i}^{t+1})\mid\mathcal{F}^t\right]=(1-1/q_n)f_{n,i}(\mathbf{y}_{n,i}^t)+(1/q_n)f_{n,i}(\mathbf{x}_n^t)$. By summing this relation for all $i\in 1,\ldots,q_n$ and divining both sides of the resulted expression by q_n we obtain

$$\mathbb{E}\left[\frac{1}{q_n}\sum_{i=1}^{q_n}f_{n,i}(\mathbf{y}_{n,i}^{t+1})\mid\mathcal{F}^t\right]=\left[1-\frac{1}{q_n}\right]\frac{1}{q_n}\sum_{i=1}^{q_n}f_{n,i}(\mathbf{y}_{n,i}^t)+\frac{1}{q_n}f_n(\mathbf{x}_n^t). \tag{82}$$

To simplicity equations let us define the sequence p_n^t as

$$p_n^t:=\frac{1}{q_n}\sum_{i=1}^{q_n}f_{n,i}(\mathbf{y}_{n,i}^t)-f_n(\tilde{\mathbf{x}}^*)-\frac{1}{q_n}\sum_{i=1}^{q_n}\nabla f_{n,i}(\tilde{\mathbf{x}}^*)^T(\mathbf{y}_{n,i}^t-\tilde{\mathbf{x}}^*). \tag{83}$$

Subtracting (81) from (82) and adding $-f_n(\tilde{\mathbf{x}}^*)$ to the both sides of equality in association with the definition of the sequence p_n^t in (83) yield

$$\mathbb{E}\left[p_n^{t+1}\mid\mathcal{F}^t\right]=\left[1-\frac{1}{q_n}\right]p_n^t+\frac{1}{q_n}\left[f_n(\mathbf{x}_n^t)-f_n(\tilde{\mathbf{x}}^*)-\nabla f_n(\tilde{\mathbf{x}}^*)^T(\mathbf{x}_n^t-\tilde{\mathbf{x}}^*)\right]. \tag{84}$$

We proceed to find and upper bound for the terms in the right hand side of (84). First note that according to the strong convexity of the local instantaneous functions $f_{n,i}$ and local functions f_n both terms in the right hand side of (84) are non-negative. Observing that the number of instantaneous functions at each node q_n satisfies the condition $q_{\min}\leq q_n\leq q_{\max}$, we obtain

$$1-\frac{1}{q_n}\leq 1-\frac{1}{q_{\max}}\ ,\qquad \frac{1}{q_n}\leq\frac{1}{q_{\min}}. \tag{85}$$

Substituting the upper bounds in (85) into (84), summing both sides of the implied inequality over $n\in\{1,\ldots,N\}$, and considering the definitions of the optimal argument $\mathbf{x}^*=[\tilde{\mathbf{x}}^*;\ldots;\tilde{\mathbf{x}}^*]$ and the aggregate function $f(\mathbf{x})=\sum_{n=1}^N f_n(\mathbf{x}_n)$ lead to

$$\sum_{n=1}^N\mathbb{E}\left[p_n^{t+1}\mid\mathcal{F}^t\right]\leq\left[1-\frac{1}{q_{\max}}\right]\sum_{n=1}^N p_n^t+\frac{1}{q_{\min}}\left[f(\mathbf{x}^t)-f(\mathbf{x}^*)-\nabla f(\mathbf{x}^*)^T(\mathbf{x}^t-\mathbf{x}^*)\right]. \tag{86}$$

Now observe that according to the definitions of the sequences p^t and p_n^t in (33) and (83), respectively, p^t is the sum of p_n^t for all n, i.e. $p^t=\sum_{n=1}^N p_n^t$. Hence, we can rewrite (86) as

$$\mathbb{E}\left[p^{t+1}\mid\mathcal{F}^t\right]\leq\left[1-\frac{1}{q_{\max}}\right]p^t+\frac{1}{q_{\min}}\left[f(\mathbf{x}^t)-f(\mathbf{x}^*)-\nabla f(\mathbf{x}^*)^T(\mathbf{x}^t-\mathbf{x}^*)\right]. \tag{87}$$

Therefore, the claim in (37) is valid.

Appendix D. Proof of Theorem 7

To prove the result in Theorem 7 first we prove the following Lemma to establish an upper bound for the squared error norm $\|\mathbf{v}^t - \mathbf{v}^*\|^2$.

Lemma 11 *Consider the DSA algorithm as defined in* (6)-(9). *Further, recall* γ' *as the smallest non-zero eigenvalue and* Γ' *as the largest eigenvalue of the matrix* $\tilde{\mathbf{Z}} - \mathbf{Z}$. *If Assumptions 1-3 hold, then the squared norm of the difference* $\|\mathbf{v}^t - \mathbf{v}^*\|^2$ *is bounded above as*

$$\|\mathbf{v}^t - \mathbf{v}^*\|^2 \leq \frac{8}{\gamma'}\mathbb{E}\left[\left\|\mathbf{x}^{t+1}-\mathbf{x}^*\right\|^2_{(\mathbf{I}+\mathbf{Z}-2\tilde{\mathbf{Z}})^2}|\,\mathcal{F}^t\right]+\frac{8}{\gamma'}\mathbb{E}\left[\left\|\mathbf{x}^t-\mathbf{x}^{t+1}\right\|^2_{\tilde{\mathbf{Z}}^2}|\,\mathcal{F}^t\right]+\frac{16\alpha^2 L}{\gamma'}\;p^t$$
$$+\frac{2\Gamma'}{\gamma'}\mathbb{E}\left[\|\mathbf{v}^t-\mathbf{v}^{t+1}\|^2\mid\mathcal{F}^t\right]+\frac{8\alpha^2\,(2L-\mu)}{\gamma'}\left[f(\mathbf{x}^t)-f(\mathbf{x}^*)-\nabla f(\mathbf{x}^*)^T(\mathbf{x}^t-\mathbf{x}^*)\right]. \quad (88)$$

Proof Consider the inequality $\|\mathbf{a}+\mathbf{b}\|^2 \leq 2\|\mathbf{a}\|^2+2\|\mathbf{b}\|^2$ for the case that $\mathbf{a} = \mathbf{U}(\mathbf{v}^{t+1}-\mathbf{v}^*)$, $\mathbf{b} = \mathbf{U}(\mathbf{v}^t-\mathbf{v}^{t+1})$ which can be written as

$$\|\mathbf{U}(\mathbf{v}^t-\mathbf{v}^*)\|^2 \leq 2\|\mathbf{U}(\mathbf{v}^{t+1}-\mathbf{v}^*)\|^2+2\|\mathbf{U}(\mathbf{v}^t-\mathbf{v}^{t+1})\|^2. \quad (89)$$

We proceed by finding an upper bound for $2\|\mathbf{U}(\mathbf{v}^{t+1}-\mathbf{v}^*)\|^2$. Based on the result of Lemma 3 in (30), the term $\mathbf{U}(\mathbf{v}^{t+1}-\mathbf{v}^*)$ is equal to the sum of vectors $\mathbf{a}+\mathbf{b}$ where $\mathbf{a} = (\mathbf{I}+\mathbf{Z}-2\tilde{\mathbf{Z}})(\mathbf{x}^{t+1}-\mathbf{x}^*)-\tilde{\mathbf{Z}}(\mathbf{x}^t-\mathbf{x}^{t+1})$ and $\mathbf{b} = -\alpha\hat{\mathbf{g}}^t-\nabla f(\mathbf{x}^*)$. Therefore, using the inequality $\|\mathbf{a}+\mathbf{b}\|^2 \leq 2\|\mathbf{a}\|^2+2\|\mathbf{b}\|^2$ we can write

$$\left\|\mathbf{U}(\mathbf{v}^{t+1}-\mathbf{v}^*)\right\|^2 \leq 2\left\|(\mathbf{I}+\mathbf{Z}-2\tilde{\mathbf{Z}})(\mathbf{x}^{t+1}-\mathbf{x}^*)-\tilde{\mathbf{Z}}(\mathbf{x}^t-\mathbf{x}^{t+1})\right\|^2+2\alpha^2\left\|\hat{\mathbf{g}}^t-\nabla f(\mathbf{x}^*)\right\|^2. \quad (90)$$

By using the inequality $\|\mathbf{a}+\mathbf{b}\|^2 \leq 2\|\mathbf{a}\|^2+2\|\mathbf{b}\|^2$ one more time for vectors $\mathbf{a} = (\mathbf{I}+\mathbf{Z}-2\tilde{\mathbf{Z}})(\mathbf{x}^{t+1}-\mathbf{x}^*)$ and $\mathbf{b} = -\tilde{\mathbf{Z}}(\mathbf{x}^t-\mathbf{x}^{t+1})$, we obtain an upper bound for the term $\|(\mathbf{I}+\mathbf{Z}-2\tilde{\mathbf{Z}})(\mathbf{x}^{t+1}-\mathbf{x}^*)-\tilde{\mathbf{Z}}(\mathbf{x}^t-\mathbf{x}^{t+1})\|^2$. Substituting this upper bound into (90) yields

$$\left\|\mathbf{U}(\mathbf{v}^{t+1}-\mathbf{v}^*)\right\|^2 \leq 4\left\|\mathbf{x}^{t+1}-\mathbf{x}^*\right\|^2_{(\mathbf{I}+\mathbf{Z}-2\tilde{\mathbf{Z}})^2}+4\left\|\mathbf{x}^t-\mathbf{x}^{t+1}\right\|^2_{\tilde{\mathbf{Z}}^2}+2\alpha^2\left\|\hat{\mathbf{g}}^t-\nabla f(\mathbf{x}^*)\right\|^2. \quad (91)$$

Inequality (91) shows an upper bound for $2\|\mathbf{U}(\mathbf{v}^{t+1}-\mathbf{v}^*)\|^2$ in (89). Moreover, we know that the second term $\|\mathbf{U}(\mathbf{v}^t-\mathbf{v}^{t+1})\|^2$ is also bounded above by $\Gamma'\|\mathbf{v}^t-\mathbf{v}^{t+1}\|^2$ where Γ' is the largest eigenvalue of matrix $\tilde{\mathbf{Z}}-\mathbf{Z} = \mathbf{U}^2$. Substituting these upper bounds into (89) and computing the expected value of both sides given the information until step t yield

$$\|\mathbf{U}(\mathbf{v}^t-\mathbf{v}^*)\|^2 \leq\; 8\mathbb{E}\left[\left\|\mathbf{x}^{t+1}-\mathbf{x}^*\right\|^2_{(\mathbf{I}+\mathbf{Z}-2\tilde{\mathbf{Z}})^2}\mid\mathcal{F}^t\right]+8\mathbb{E}\left[\left\|\mathbf{x}^t-\mathbf{x}^{t+1}\right\|^2_{\tilde{\mathbf{Z}}^2}\mid\mathcal{F}^t\right]$$
$$+4\alpha^2\mathbb{E}\left[\left\|\hat{\mathbf{g}}^t-\nabla f(\mathbf{x}^*)\right\|^2\mid\mathcal{F}^t\right]+2\Gamma'\mathbb{E}\left[\|\mathbf{v}^t-\mathbf{v}^{t+1}\|^2\mid\mathcal{F}^t\right]. \quad (92)$$

Note the vectors $\mathbf{v}^t$ and $\mathbf{v}^*$ lie in the column space of the matrix $\mathbf{U}$. Thus, we obtain that $\|\mathbf{U}(\mathbf{v}^t-\mathbf{v}^*)\|^2 \geq \gamma'\|\mathbf{v}^t-\mathbf{v}^*\|^2$. Substituting this lower bound for $\|\mathbf{U}(\mathbf{v}^t-\mathbf{v}^*)\|^2$ in (92) and deviding both sides of the imposed inequality by γ' yield

$$\|\mathbf{v}^t-\mathbf{v}^*\|^2 \leq \frac{8}{\gamma'}\mathbb{E}\left[\left\|\mathbf{x}^{t+1}-\mathbf{x}^*\right\|^2_{(\mathbf{I}+\mathbf{Z}-2\tilde{\mathbf{Z}})^2}\mid\mathcal{F}^t\right]+\frac{8}{\gamma'}\mathbb{E}\left[\left\|\mathbf{x}^t-\mathbf{x}^{t+1}\right\|^2_{\tilde{\mathbf{Z}}^2}\mid\mathcal{F}^t\right]$$
$$+\frac{4\alpha^2}{\gamma'}\mathbb{E}\left[\left\|\hat{\mathbf{g}}^t-\nabla f(\mathbf{x}^*)\right\|^2\mid\mathcal{F}^t\right]+\frac{2\Gamma'}{\gamma'}\mathbb{E}\left[\|\mathbf{v}^t-\mathbf{v}^{t+1}\|^2\mid\mathcal{F}^t\right]. \quad (93)$$

By substituting the expectation $\mathbb{E}\left[\|\hat{\mathbf{g}}^t-\nabla f(\mathbf{x}^*)\|^2\mid\mathcal{F}^t\right]$ in the right hand side of (93) with its upper bound in (34), the claim in (88) follows. ■

Using the result in Lemma 11 we show that the sequence $\|\mathbf{u}^t - \mathbf{u}^*\|^2_{\mathbf{G}} + c\ p^t$ converges linearly to zero.

Proof of Theorem 7: Proving the linear convergence claim in (40) is equivalent to showing that

$$\delta\|\mathbf{u}^t - \mathbf{u}^*\|^2_{\mathbf{G}} + \delta c\ p^t \leq \|\mathbf{u}^t - \mathbf{u}^*\|^2_{\mathbf{G}} - \mathbb{E}\left[\|\mathbf{u}^{t+1} - \mathbf{u}^*\|^2_{\mathbf{G}} \mid \mathcal{F}^t\right] + c\ (p^t - \mathbb{E}\left[p^{t+1} \mid \mathcal{F}^t\right]). \tag{94}$$

Substituting the terms $\mathbb{E}\left[\|\mathbf{u}^{t+1} - \mathbf{u}^*\|^2_{\mathbf{G}} \mid \mathcal{F}^t\right]$ and $\mathbb{E}\left[p^{t+1} \mid \mathcal{F}^t\right]$ by their upper bounds as introduced in Lemma 5 and Lemma 6, respectively, yields a sufficient condition for the claim in (94) as

$$\begin{aligned}\delta\|\mathbf{u}^t - \mathbf{u}^*\|^2_{\mathbf{G}} + \delta c\ p^t \leq\ & \mathbb{E}\left[\|\mathbf{x}^{t+1} - \mathbf{x}^t\|^2_{\tilde{\mathbf{Z}}-2\alpha\eta\mathbf{I}} \mid \mathcal{F}^t\right] + \mathbb{E}\left[\|\mathbf{v}^{t+1} - \mathbf{v}^t\|^2 \mid \mathcal{F}^t\right] \\ & + 2\mathbb{E}\left[\|\mathbf{x}^{t+1} - \mathbf{x}^*\|^2_{\mathbf{I}+\mathbf{Z}-2\tilde{\mathbf{Z}}} \mid \mathcal{F}^t\right] + \left(\frac{c}{q_{\max}} - \frac{4\alpha L}{\eta}\right) p^t \\ & + \left[\frac{4\alpha\mu}{L} - \frac{2\alpha(2L-\mu)}{\eta} - \frac{c}{q_{\min}}\right]\left[f(\mathbf{x}^t) - f(\mathbf{x}^*) - \nabla f(\mathbf{x}^*)^T(\mathbf{x}^t - \mathbf{x}^*)\right]. \end{aligned} \tag{95}$$

We emphasize that if the inequality in (95) holds, then the inequalities in (94) and (40) are valid. Note that the weighted norm $\|\mathbf{u}^t - \mathbf{u}^*\|^2_{\mathbf{G}}$ in the left hand side of (95) can be simplified as $\|\mathbf{x}^t - \mathbf{x}^*\|^2_{\tilde{\mathbf{Z}}} + \|\mathbf{v}^t - \mathbf{v}^*\|^2$. Considering the definition of Γ as the maximum eigenvalue of the matrix $\tilde{\mathbf{Z}}$, we can conclude that $\|\mathbf{x}^t - \mathbf{x}^*\|^2_{\tilde{\mathbf{Z}}}$ is bounded above by $\Gamma\|\mathbf{x}^t - \mathbf{x}^*\|^2$. Considering this relation and observing the upper bound for $\|\mathbf{v}^t - \mathbf{v}^*\|^2$ in (88), we obtain that $\|\mathbf{u}^t - \mathbf{u}^*\|^2_{\mathbf{G}} = \|\mathbf{x}^t - \mathbf{x}^*\|^2_{\tilde{\mathbf{Z}}} + \|\mathbf{v}^t - \mathbf{v}^*\|^2$ is bounded above by

$$\begin{aligned}\|\mathbf{u}^t - \mathbf{u}^*\|^2_{\mathbf{G}}\ \leq\ & \frac{8}{\gamma'}\mathbb{E}\left[\left\|\mathbf{x}^{t+1} - \mathbf{x}^*\right\|^2_{(\mathbf{I}+\mathbf{Z}-2\tilde{\mathbf{Z}})^2} \mid \mathcal{F}^t\right] + \frac{8}{\gamma'}\mathbb{E}\left[\left\|\mathbf{x}^t - \mathbf{x}^{t+1}\right\|^2_{\tilde{\mathbf{Z}}^2} \mid \mathcal{F}^t\right] + \frac{16\alpha^2 L}{\gamma'}\ p^t \\ & + \frac{2\Gamma'}{\gamma'}\mathbb{E}\left[\|\mathbf{v}^t - \mathbf{v}^{t+1}\|^2 \mid \mathcal{F}^t\right] + \Gamma\|\mathbf{x}^t - \mathbf{x}^*\|^2 \\ & + \frac{8\alpha^2(2L-\mu)}{\gamma'}\left[f(\mathbf{x}^t) - f(\mathbf{x}^*) - \nabla f(\mathbf{x}^*)^T(\mathbf{x}^t - \mathbf{x}^*)\right]. \end{aligned} \tag{96}$$

Further, the strong convexity of the global objective function f implies that the squared norm $\|\mathbf{x}^t - \mathbf{x}^*\|^2$ is upper bound by $(2/\mu)(f(\mathbf{x}^t) - f(\mathbf{x}^*) - \nabla f(\mathbf{x}^*)^T(\mathbf{x}^t - \mathbf{x}^*))$. Replacing the the squared norm $\|\mathbf{x}^t - \mathbf{x}^*\|^2$ in (96) by its upper bound leads to

$$\begin{aligned}\|\mathbf{u}^t - \mathbf{u}^*\|^2_{\mathbf{G}} \leq\ & \frac{8}{\gamma'}\mathbb{E}\left[\left\|\mathbf{x}^{t+1} - \mathbf{x}^*\right\|^2_{(\mathbf{I}+\mathbf{Z}-2\tilde{\mathbf{Z}})^2} \mid \mathcal{F}^t\right] + \frac{8}{\gamma'}\mathbb{E}\left[\left\|\mathbf{x}^t - \mathbf{x}^{t+1}\right\|^2_{\tilde{\mathbf{Z}}^2} \mid \mathcal{F}^t\right] + \frac{16\alpha^2 L}{\gamma'}\ p^t \\ & + \frac{2\Gamma'}{\gamma'}\mathbb{E}\left[\|\mathbf{v}^t - \mathbf{v}^{t+1}\|^2 \mid \mathcal{F}^t\right] \\ & + \left(\frac{8\alpha^2(2L-\mu)}{\gamma'} + \frac{2\Gamma}{\mu}\right)\left[f(\mathbf{x}^t) - f(\mathbf{x}^*) - \nabla f(\mathbf{x}^*)^T(\mathbf{x}^t - \mathbf{x}^*)\right]. \end{aligned} \tag{97}$$

Replacing $\|\mathbf{u}^t - \mathbf{u}^*\|^2_{\mathbf{G}}$ in (95) by the upper bound in (97) and regrouping the terms lead to

$$\begin{aligned}0 \leq\ & \mathbb{E}\left[\|\mathbf{x}^{t+1} - \mathbf{x}^t\|^2_{\tilde{\mathbf{Z}}-\alpha(\eta+\eta)\mathbf{I}-\frac{8\delta}{\gamma'}\tilde{\mathbf{Z}}^2} \mid \mathcal{F}^t\right] + \mathbb{E}\left[\|\mathbf{x}^{t+1} - \mathbf{x}^*\|^2_{(\mathbf{I}+\mathbf{Z}-2\tilde{\mathbf{Z}})^{\frac{1}{2}}\left[2\mathbf{I}-\frac{8\delta}{\gamma'}(\mathbf{I}+\mathbf{Z}-2\tilde{\mathbf{Z}})\right](\mathbf{I}+\mathbf{Z}-2\tilde{\mathbf{Z}})^{\frac{1}{2}}} \mid \mathcal{F}^t\right] \\ & + \mathbb{E}\left[\|\mathbf{v}^{t+1} - \mathbf{v}^t\|^2_{(1-\frac{2\delta\Gamma'}{\gamma'})\mathbf{I}} \mid \mathcal{F}^t\right] + \left[\frac{c}{q_{\max}} - \frac{4\alpha L}{\eta} - \delta c - \frac{16\delta\alpha^2 L}{\gamma'}\right] p^t \\ & + \left[\frac{4\alpha\mu}{L} - \frac{2\alpha(2L-\mu)}{\eta} - \frac{c}{q_{\min}} - \frac{8\delta\alpha^2(2L-\mu)}{\gamma'} - \frac{2\delta\Gamma}{\mu}\right](f(\mathbf{x}^t) - f(\mathbf{x}^*) - \nabla f(\mathbf{x}^*)^T(\mathbf{x}^t - \mathbf{x}^*)). \end{aligned} \tag{98}$$

Notice that if the inequality in (98) holds true, then the relation in (95) is valid and as we mentioned before the claim in (94) holds. To verify the sum in the right hand side of (98) is always positive and the inequality is valid, we enforce each summands in the right hand side of (98) to be non-negative. Therefore, the following conditions should be satisfied

$$\gamma - \alpha(\eta+\eta) - \frac{8\delta}{\gamma'}\Gamma^2 \geq 0, \quad 2 - \frac{8\delta}{\gamma'}\lambda_{max}(\mathbf{I}+\mathbf{Z}-2\tilde{\mathbf{Z}}) \geq 0, \quad 1 - \frac{2\delta\Gamma'}{\gamma'} \geq 0,$$
$$\frac{c}{q_{\max}} - \frac{4\alpha L}{\eta} - \delta c - \frac{16\delta\alpha^2 L}{\gamma'} \geq 0, \quad \frac{4\alpha\mu}{L} - \frac{2\alpha(2L-\mu)}{\eta} - \frac{c}{q_{\min}} - \frac{8\delta\alpha^2\,(2L-\mu)}{\gamma'} - \frac{2\delta\Gamma}{\mu} \geq 0. \quad (99)$$

Recall that γ is the smallest eigenvalue of the positive definite matrix $\mathbf{Z}$. All the inequalities in (99) are satisfied, if δ is chosen as

$$\delta = \min\Bigg\{ \frac{(\gamma-2\alpha\eta)\gamma'}{8\Gamma^2}, \frac{\gamma'}{4\lambda_{max}(\mathbf{I}+\mathbf{Z}-2\tilde{\mathbf{Z}})}, \frac{\gamma'}{2\Gamma'}, \frac{\gamma'(c\eta - 4\alpha L q_{\max})}{\eta q_{\max}(c\gamma' + 16\alpha^2 L)}, \left[\frac{4\alpha\mu}{L} - \frac{2\alpha(2L-\mu)}{\eta} - \frac{c}{q_{\min}}\right]\left[\frac{8\alpha^2\,(2L-\mu)}{\gamma'} + \frac{2\Gamma}{\mu}\right]^{-1} \Bigg\}. \quad (100)$$

where η, c and α are selected from the intervals

$$\eta \in \left(\frac{L^2 q_{\max}}{\mu q_{\min}} + \frac{L^2}{\mu} - \frac{L}{2}\ ,\ \infty\right),\ \alpha \in \left(0\ ,\ \frac{\gamma}{2\eta}\right),\ c \in \left(\frac{4\alpha L q_{\max}}{\eta}\ ,\ \frac{4\alpha\mu q_{\min}}{L} - \frac{2\alpha q_{\min}(2L-\mu)}{\eta}\right). \quad (101)$$

Notice that considering the conditions for the variables η, α and c in (101), the constant δ in (100) is strictly positive $\delta > 0$. Moreover, according to the definition in (100) the constant δ is smaller than $\gamma'/2\Gamma'$ which leads to the conclusion that $\delta \leq 1/2 < 1$. Therefore, we obtain that $0 < \delta < 1$ and the claim in (40) is valid.

Appendix E. Proof of Theorem 9

The proof uses the relationship in the statement (40) of Theorem 7 to build a supermartingale sequence. To do this define the stochastic processes ζ^t and β^t as

$$\zeta^t := \|\mathbf{u}^t - \mathbf{u}^*\|^2_{\mathbf{G}} + c\,p^t, \quad \beta^t := \delta\left(\|\mathbf{u}^t - \mathbf{u}^*\|^2_{\mathbf{G}} + c\,p^t\right). \quad (102)$$

The stochastic processes ζ^t and β^t are alway non-negative. Let now $\mathcal{F}_t$ be a sigma-algebra measuring ζ^t, β^t, and $\mathbf{u}^t$. Considering the definitions of ζ^t and β^t and the relation in (40) we can write

$$\mathbb{E}\left[\zeta^{t+1} \mid \mathcal{F}^t\right] \leq \zeta^t - \beta^t. \quad (103)$$

Since the sequences α^t and β^t are nonnegative it follows from (103) that they satisfy the conditions of the supermartingale convergence theorem – see e.g. theorem E7.4 Solo and Kong (1995) . Therefore, we obtain that: (i) The sequence ζ^t converges almost surely. (ii) The sum $\sum_{t=0}^{\infty}\beta^t < \infty$ is almost surely finite. The definition of β^t in (102) implies that

$$\sum_{t=0}^{\infty}\delta\left(\|\mathbf{u}^t - \mathbf{u}^*\|^2_{\mathbf{G}} + c\,p^t\right) < \infty, \qquad \text{a.s.} \quad (104)$$

Since $\|\mathbf{x}^t - \mathbf{x}^*\|^2_{\tilde{\mathbf{Z}}} \leq \|\mathbf{u}^t - \mathbf{u}^*\|^2_{\mathbf{G}} + c\,p^t$ and the eigenvalues of $\tilde{\mathbf{Z}}$ are lower bounded by γ we can write $\gamma\|\mathbf{x}^t - \mathbf{x}^*\|^2 \leq \|\mathbf{u}^t - \mathbf{u}^*\|^2_{\mathbf{G}} + c\,p^t$. This inequality in association with the fact that the sum in (104) is finite leads to

$$\sum_{t=0}^{\infty}\delta\,\gamma\,\|\mathbf{x}^t - \mathbf{x}^*\|^2 < \infty, \qquad \text{a.s.} \quad (105)$$

Observing the fact that δ and γ are positive constants, we can conclude from (105) that the sequence $\|\mathbf{x}^t - \mathbf{x}^*\|^2$ is almost surely summable and it converges with probability 1 to 0.

References

Ron Bekkerman, Mikhail Bilenko, and John Langford. *Scaling up machine learning: Parallel and distributed approaches*. Cambridge University Press, 2011.

Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011.

Francesco Bullo, Jorge Cortés, and Sonia Martinez. *Distributed control of robotic networks: a mathematical approach to motion coordination algorithms*. Princeton University Press, 2009.

Yongcan Cao, Wenwu Yu, Wei Ren, and Guanrong Chen. An overview of recent progress in the study of distributed multi-agent coordination. *Industrial Informatics, IEEE Transactions on*, 9 (1):427–438, 2013.

Volkan Cevher, Steffen Becker, and Martin Schmidt. Convex optimization for big data: Scalable, randomized, and parallel algorithms for big data analytics. *Signal Processing Magazine, IEEE*, 31(5):32–43, 2014.

Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing Systems*, pages 1646–1654, 2014.

John C Duchi, Alekh Agarwal, and Martin J Wainwright. Dual averaging for distributed optimization: convergence analysis and network scaling. *Automatic control, IEEE Transactions on*, 57(3): 592–606, 2012.

Franck Iutzeler, Pascal Bianchi, Philippe Ciblat, and Walid Hachem. Explicit convergence rate of a distributed alternating direction method of multipliers. *arXiv preprint arXiv:1312.1085*, 2013.

Dusan Jakovetic, Joao Xavier, and Jose MF Moura. Fast distributed gradient methods. *Automatic Control, IEEE Transactions on*, 59(5):1131–1146, 2014.

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems*, pages 315–323, 2013.

Usman A Khan, Soummya Kar, and JoséM F Moura. Diland: An algorithm for distributed sensor localization with noisy distance measurements. *Signal Processing, IEEE Transactions on*, 58(3): 1940–1947, 2010.

Jakub Konečnỳ and Peter Richtárik. Semi-stochastic gradient descent methods. *arXiv preprint arXiv:1312.1666*, 2013.

Qing Ling and Alejandro Ribeiro. Decentralized linearized alternating direction method of multipliers. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 5447–5451. IEEE, 2014.

Qing Ling, Wei Shi, Gang Wu, and Alejandro Ribeiro. Dlm: Decentralized linearized alternating direction method of multipliers. *Signal Processing, IEEE Transactions on*, 63(15):4051–4064, 2015.

Cassio G Lopes and Ali H Sayed. Diffusion least-mean squares over adaptive networks: Formulation and performance analysis. *Signal Processing, IEEE Transactions on*, 56(7):3122–3136, 2008.

Aryan Mokhtari, Qing Ling, and Alejandro Ribeiro. Network newton-part i: Algorithm and convergence. *arXiv preprint arXiv:1504.06017*, 2015a.

Aryan Mokhtari, Qing Ling, and Alejandro Ribeiro. Network newton-part ii: Convergence rate and implementation. *arXiv preprint arXiv:1504.06020*, 2015b.

Aryan Mokhtari, Wei Shi, Qing Ling, and Alejandro Ribeiro. Decentralized quadratically approximated alternating direction method of multipliers. In *Proc. IEEE Global Conf. on Signal and Inform. Process.*, 2015c.

Angelia Nedić and Asuman Ozdaglar. Distributed subgradient methods for multi-agent optimization. *Automatic Control, IEEE Transactions on*, 54(1):48–61, 2009.

Michael Rabbat and Robert Nowak. Distributed optimization in sensor networks. In *Proceedings of the 3rd international symposium on Information processing in sensor networks*, pages 20–27. ACM, 2004.

Alejandro Ribeiro. Ergodic stochastic optimization algorithms for wireless communication and networking. *Signal Processing, IEEE Transactions on*, 58(12):6369–6386, 2010.

Alejandro Ribeiro. Optimal resource allocation in wireless communication and networking. *EURASIP Journal on Wireless Communications and Networking*, 2012(1):1–19, 2012.

Nicolas L Roux, Mark Schmidt, and Francis R Bach. A stochastic gradient method with an exponential convergence _rate for finite training sets. In *Advances in Neural Information Processing Systems*, pages 2663–2671, 2012.

Ioannis D Schizas, Alejandro Ribeiro, and Georgios B Giannakis. Consensus in ad hoc wsns with noisy links–part i: Distributed estimation of deterministic signals. *Signal Processing, IEEE Transactions on*, 56(1):350–364, 2008.

Mark Schmidt, Nicolas Le Roux, and Francis Bach. Minimizing finite sums with the stochastic average gradient. *arXiv preprint arXiv:1309.2388*, 2013.

Shai Shalev-Shwartz and Tong Zhang. Stochastic dual coordinate ascent methods for regularized loss. *The Journal of Machine Learning Research*, 14(1):567–599, 2013.

Wei Shi, Qing Ling, Kun Yuan, Gang Wu, and Wotao Yin. On the linear convergence of the admm in decentralized consensus optimization. *Signal Processing, IEEE Transactions on*, 62(7):1750–1761, 2014.

Wei Shi, Qing Ling, Gang Wu, and Wotao Yin. Extra: An exact first-order algorithm for decentralized consensus optimization. *SIAM Journal on Optimization*, 25(2):944–966, 2015.

Victor Solo and Xuan Kong. *Adaptive Signal Processing Algorithms: Stability and Performance*. NJ: Prentice-Hall, Englewood Cliffs, 1995.

Konstantinos Tsianos, Sean Lawlor, Michael G Rabbat, et al. Consensus-based distributed optimization: Practical issues and applications in large-scale machine learning. In *Communication, Control, and Computing (Allerton), 2012 50th Annual Allerton Conference on*, pages 1543–1550. IEEE, 2012a.

Konstantinos I. Tsianos, Sean Lawlor, and Michael G Rabbat. Push-sum distributed dual averaging for convex optimization. In *Decision and Control (CDC), 2012 IEEE 51st Annual Conference on*, pages 5453–5458. IEEE, 2012b.

Kun Yuan, Qing Ling, and Wotao Yin. On the convergence of decentralized gradient descent. *arXiv preprint arXiv:1310.7063*, 2013.