

Push-Pull with Device Sampling

Yu-Guan Hsieh, Yassine Laguel, Franck Iutzeler, Jérôme Malick

▶ **To cite this version:**

Yu-Guan Hsieh, Yassine Laguel, Franck Iutzeler, Jérôme Malick. Push-Pull with Device Sampling. 2022. hal-03691122

HAL Id: hal-03691122
https://hal.science/hal-03691122

Preprint submitted on 8 Jun 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Push–Pull with Device Sampling

Yu-Guan Hsieh, Yassine Laguel, Franck Iutzeler, Jérôme Malick .

***Abstract*—We consider decentralized optimization problems in which a number of agents collaborate to minimize the average of their local functions by exchanging over an underlying communication graph. Specifically, we place ourselves in an asynchronous model where only a random portion of nodes perform computation at each iteration, while the information exchange can be conducted between all the nodes and in an asymmetric fashion. For this setting, we propose an algorithm that combines gradient tracking and variance reduction over the entire network. This enables each node to track the average of the gradients of the objective functions. Our theoretical analysis shows that the algorithm converges linearly, when the local objective functions are strongly convex, under mild connectivity conditions on the expected mixing matrices. In particular, our result does not require the mixing matrices to be doubly stochastic. In the experiments, we investigate a broadcast mechanism that transmits information from computing nodes to their neighbors, and confirm the linear convergence of our method on both synthetic and real-world datasets.**

***Index Terms*—decentralized optimization, convex optimization, random gossip, device sampling**

I. Introduction

In this paper, we focus on solving the optimization problem

$$\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}) \coloneqq \frac{1}{M}\sum_{i=1}^{M} f_i(\mathbf{x}) \tag{P}$$

where each function $f_i\colon \mathbb{R}^d \to \mathbb{R}$ is available only locally at the i-th node of a graph. Hence, in order to reach a consensus on the minimum of (P), the M nodes have to communicate using the graph's edges.

Such decentralized optimization problems have been widely studied in the literature at least since the pioneering works of Bertsekas and Tsitsiklis [1], [2]. In terms of applications, decentralized optimization methods are popular for regression or classification problems when the communication possibilities between the nodes are scarce and cannot be handled by a central entity (e.g., for wireless sensor networks, IoT-enabled edge devices, etc.); see the recent surveys [3], [4], [5], [6]. In these applications, the workload and communications between the nodes are of primary importance.

The computation at the node level mainly depends on which optimization method serves as a basis. If the nodes are able to solve optimization sub-problems, the Alternating Direction Method of Multipliers (ADMM) and other dual methods can be extended to distributed setting [7], [8]. At the other end of the spectrum, stochastic gradients methods are very popular since they require minimal computation at each node [9]. Gradient-based methods offer a good comprise between these two extremes and currently know a rebirth, especially for machine learning applications; see e.g., the recent [10].

In terms of exchanges, all communications between nodes have to go through the edges of the graph. If the graph is undirected (i.e., the edges are all bidirectional), the nodes can gossip to average their values. Mathematically, this corresponds to multiplying the agents' states by a doubly-stochastic matrix; see [6, Sec. II] for details. However, if the graph is directed, such direct gossiping is no longer possible since maintaining both a consensus among the nodes and the average of their values is not possible at the same time [11]. To overcome this problem, two main type of methods have been developed. First, Push–Sum methods (or ratio consensus) consist in exchanging an additional "weighting"; these methods can reach an average consensus for the ratio of the two values [12], [13], [14]. However, the analysis of Push–Sum gradient methods is often quite involved and the algorithm can become numerically unstable due to division by very small values, see e.g., the simulations of [15] as well as references therein. Second, Push–Pull methods rely on two communications steps with different mixings to maintain convergence, offering strong theoretical guarantees as well as good practical performance [16], [15].

Finally, a desirable feature in decentralized methods is the possibility to allow the nodes to randomly awaken, compute, and send/receive information; which is generally called randomized gossiping [17] or asynchronous decentralized methods [5], [18]. In terms of analysis, this consists in replacing the fixed communication matrices with random ones having the support corresponding to the active links; this was actively studied for decentralized gradient methods, including Push–Pull gradient [15].

A. Contributions and outline

In this paper, we focus on gradient-based methods for decentralized asynchronous optimization on directed graphs. We propose and analyze an asynchronous Push–Pull gradient algorithm where only a fraction of the nodes are actively computing a local gradient at each iteration. This feature is inspired from the device sampling (or client selection) procedure in federated learning [19], [20]. This popular mechanism enables to take into account the fact that all the nodes may not be available at all time and furthermore that querying all gradients at each iteration may be a waste of computational power if the nodes' values only change by a little amount.

In terms of algorithm, device sampling calls for a variance reduction mechanism in order to mitigate the noise induced

Yu-Guan Hsieh and Franck Iutzeler are with Université Grenoble Alpes, Grenoble, France (email: yu-guan.hsieh@univ-grenoble.alpes; franck.iutzeler@univ-grenoble.alpes).

Yassine Laguel was with Université Grenoble Alpes, Grenoble, France. He is now with Rutgers University, NJ 07102 USA (email: laguel.yassine@gmail.com).

Jérôme Malick is with CNRS and Université Grenoble Alpes, Grenoble, France (email: jerome.malick@cnrs.fr)

by the sampling of the nodes. To achieve this, we introduce a SAGA-like [21] update at the network level; see Example 4. This additional step thus calls for an original analysis.

The remainder of the paper is structured as follows. The introduction is completed by an overview on related literature. In Section II, we present our general algorithmic template (Push–Pull with Device Sampling), together with some specific cases of interest, connecting our method with existing methods. In Section III, we provide linear convergence results under classical convexity/smoothness assumptions on the objective functions and weak assumptions on communications. Section IV and Section V are dedicated to the detailed convergence analysis and illustrative numerical simulations. Finally, proofs of a couple of technical intermediate results are given in Appendix.

B. Related works

Direct extensions of the gradient method to the decentralized setting rely on decreasing stepsizes to converge and are thus limited to sublinear convergence rates, even if the minimized functions are smooth and strongly convex. To overcome this situation, the gradient tracking technique was introduced; it consists in dynamically tracking the average value of the gradient and using this value instead of the local gradient. This technique enables the use of a fixed stepsize and exhibits much better rates in theory and in practice [22], [23], [24], [25]; see also the recent [26]. Gradient tracking can be intuitively seen as a variance reduction at the network level. The method presented in this paper extends this idea of variance reduction to device sampling. Note also that in the case where the nodes' functions are themselves a finite sum, this sum can be sampled, and variance reduction can be additionally applied at the node level [27], [28], however this specific form is out of the scope of the present paper.

AB/Push–Pull gradient methods naturally involve gradient tracking; see e.g., [15, Rem. 1] and more generally [29], [30], [31], [15]. These algorithms share common ingredients and mainly differ in their communications models. The works that are the most closely related to the asynchronous directed setup considered in this paper are [15] and [30]. These two papers study an asynchronous version of AB/Push–Pull which share similarities, in the update and the communication scheme, with our proposed method (more precisely, with the special setup of Example 3). However, in contrast to our method, these methods require every node that is involved in the communication step to perform a local update. Moreover, the analysis of [15] only works for the more restrictive case where the non-diagonal coefficients of the mixing matrices are sufficiently small. Note finally that [30] does not consider a random network model, but instead performs an analysis in terms of the worst-case dependence on the delays. This analysis is thus complementary to our work.

C. Basic notation and definitions

Throughout the paper, we use bold lowercase letters to denote vectors and capital letters to denote matrices. I_k and $\mathbf{1}_k$ respectively represent the identity matrix of size $k \times k$ and the $k-$dimensional vector containing all ones. The subscript is omitted when the dimension is clear from the context. We also define $J = \mathbf{1}_M \mathbf{1}_M^\top / M$ as the projection matrix onto the consensus space, and denote by $\rho(P)$ the spectral radius of a matrix P.

The interaction topology between the nodes is modeled by a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the set of vertices (nodes) and $\mathcal{E} \in \mathcal{V} \times \mathcal{V}$ is the set of edges, such that node i can send information to j only if $(i,j) \in \mathcal{E}$. The out-neighbors and in-neighbors of a node i are respectively defined by

$$\mathcal{N}_i^{out} = \{j \in \mathcal{V} : (i,j) \in \mathcal{E}\},\ \mathcal{N}_i^{in} = \{j \in \mathcal{V} : (j,i) \in \mathcal{E}\}.$$

When the graph is undirected, the two sets coincide and we simply write $\mathcal{N}_i$. We say that the matrix $W = (w_{ij}) \in \mathbb{R}^{M \times M}$ is compatible with the underlying communication topology if $w_{ij} = 0$ whenever $(j,i) \notin \mathcal{E}$.

Finally we introduce the aggregate objective function, $F(X) = \sum_{i=1}^M f_i(\mathbf{x}_i)$, as a function of the variable $X = [\mathbf{x}_1, \dots, \mathbf{x}_M]^\top \in \mathbb{R}^{M \times d}$. When F is differentiable, we have

$$\nabla F(X) = [\nabla f_1(\mathbf{x}_1), \dots, \nabla f_M(\mathbf{x}_M)]^\top.$$

II. Algorithms: existing, new, and examples

In this section, we present our asynchronous Push-Pull gradient algorithm with device sampling. Prior to that, we recall the existing AB/Push–Pull method [29], [15] which inspires our algorithm. After detailing our general template, we instantiate it in several situations of interest, revealing its versatility.

A. The AB/Push–Pull method

If all the nodes are active at each iteration, the communication setup reduces to that of synchronous decentralized optimization. In this situation and assuming that the functions f_i are differentiable, the AB/Push–Pull algorithm [29], [15] is described as follows. In addition to the decision variables $\mathbf{x}_i^t$ that should minimize f, a variable $\mathbf{y}_i^t$ is introduced to track the gradient of f. Then, provided $\eta > 0$ a constant stepsize and two mixing matrices $A = (a_{ij}) \in \mathbb{R}^{M \times M}$ and $B = (b_{ij}) \in \mathbb{R}^{M \times M}$, the update of the algorithm at iteration t writes

$$\begin{aligned} \mathbf{x}_i^{t+1} &= \sum_{j \in \mathcal{V}} a_{ij} \mathbf{x}_j^t - \eta\, \mathbf{y}_i^t, \\ \mathbf{y}_i^{t+1} &= \sum_{j \in \mathcal{V}} b_{ij} \mathbf{y}_j^t + \nabla f_i(\mathbf{x}_i^{t+1}) - \nabla f_i(\mathbf{x}_i^t). \end{aligned}$$

It is required that A and B have non-negative weights and be respectively row-stochastic ($A\mathbf{1} = \mathbf{1}$) and column-stochastic ($\mathbf{1}^\top B = \mathbf{1}^\top$). With the notation $Y_t = [\mathbf{y}_1^t, \dots, \mathbf{y}_M^t]^\top$, the update can also be written, in a matrix form, as

$$\begin{aligned} X_{t+1} &= A_t X_t - \eta Y_t, \\ Y_{t+1} &= B_t Y_t + \nabla F(X_{t+1}) - \nabla F(X_t). \end{aligned} \tag{PP}$$

Intuitively, the use of row-stochastic matrices drive $\mathbf{x}_i^t$ to consensus, while the use of column-stochastic matrices preserves the total mass, i.e., $\mathbf{1}^\top B \omega = \mathbf{1}^\top \omega$ for any $\omega \in \mathbb{R}^M$. Moreover, if the difference $\nabla f_i(\mathbf{x}_i^{t+1}) - \nabla f_i(\mathbf{x}_i^t)$ tends to

zero, $\mathbf{y}_i^t$ converges to a multiple of $\sum_{i=1}^M \nabla f_i(\mathbf{x}_i^t)$. In fact, from the Perron-Frobenius theorem, we know that if B is primitive[1] then $\lim_{t\to+\infty} B^t = \pi_B \mathbf{1}^\top$ where π_B is the right eigenvector of B associated with the eigenvalue 1 such that $\mathbf{1}^\top \pi_B = 1$. Therefore, asymptotically every $\mathbf{x}_i^t$ descends in the direction opposite to the gradient of f. Mathematically, it can be proven that under standard convexity assumptions AB/Push–Pull converges linearly with sufficiently small constant step-size η [33, Th. 1].

B. Proposed Push–Pull with Device Sampling

Our algorithm can be viewed as a generalization of AB/Push–Pull to handle the device sampling mechanism. First, in order to allow for asynchronicity, let $(A_t)_{t\in\mathbb{N}}$ and $(B_t)_{t\in\mathbb{N}}$ be two sequences of mixing matrices that are compatible with the underlying communication topology. Now, to handle device sampling, we denote by $\mathcal{V}_t$ the set of nodes that are active at time t. This means that node i computes a local gradient at round t if and only if $i \in \mathcal{V}_t$.

With the notations $A_t = (a_{ij}^t)$, $B_t = (b_{ij}^t)$, and $D_t = \operatorname{diag}(\mathbb{1}_{i\in\mathcal{V}_t})$, i.e., D_t is the diagonal matrix in $\mathbb{R}^{M\times M}$ whose i-th diagonal element is 1 if $i \in \mathcal{V}_t$ and 0 otherwise, each iteration of our proposed Push–Pull with Device Sampling (PPDS) can be stated in the compact form

$$\begin{aligned} Y_{t+\frac{1}{2}} &= Y_t + D_t(\nabla F(X_t) - \nabla F(Z_t)),\\ X_{t+\frac{1}{2}} &= X_t - \eta D_t Y_{t+\frac{1}{2}},\\ Z_{t+1} &= D_t X_t + (I - D_t) Z_t,\\ Y_{t+1} &= B_t Y_{t+\frac{1}{2}}, \quad X_{t+1} = A_t X_{t+\frac{1}{2}}. \end{aligned} \tag{PPDS}$$

Several remarks are in order. First, we introduce auxiliary local variable $\mathbf{z}_i^t$ for each node and write $Z_t = [\mathbf{z}_1^t, \dots, \mathbf{z}_M^t]^\top$. The presence of these variables indicate the nodes store their last computed gradient. This is necessary because $\mathbf{x}_i^t$ can be modified by network communication between two successive activations of node i. In fact, while only the active nodes perform local updates at each iteration, the inactive nodes can be involved in the communication process. This flexibility allows us to take into account a wider class of algorithms, as illustrated in the forthcoming examples. Finally, as in AB/Push–Pull, we only require the matrices $(A_t)_{t\in\mathbb{N}}$ and $(B_t)_{t\in\mathbb{N}}$ to be respectively row- and column-stochastic. This means that we allow for one-way communication and in particular inactive nodes may passively receive information without sending back their local states.

In terms of implementation, our method (PPDS) gives Algorithm 1 for asynchronous optimization on directed graphs. In the next section, we discuss special cases and show that we recover existing algorithms.

C. Special cases

1) AB/Push–Pull: We first demonstrate that the original AB/Push-Pull algorithm [29], [15] indeed falls within the PPDS framework. For this, we fix $A_t \equiv A$, $B_t \equiv B$, and $\mathcal{V}_t = \mathcal{V}$. Then, after rearranging, the (PPDS) update can be written as

$$\begin{aligned} X_{t+1} &= A(X_t - \eta Y_{t+\frac{1}{2}}),\\ Y_{t+\frac{3}{2}} &= B Y_{t+\frac{1}{2}} + \nabla F(X_{t+1}) - \nabla F(X_t). \end{aligned}$$

This is exactly the adapt-then-combine variant of AB/Push–Pull as presented in [15].

[1] A square non-negative matrix W is called primitive if there exists a power $k \geq 1$ such that $W^k > 0$; see [32, Th. 8.5.2]

Algorithm 1 PPDS (at each node i)

Initialize: $\mathbf{y}_i^1 = \nabla f_i(\mathbf{x}_i^1)$; $\mathbf{z}_i^1 = \mathbf{x}_i^1$
for $t = 1, 2, \dots$ **do**
  `Local update`
  if $i \in \mathcal{V}_t$ **then**
    $\mathbf{y}_i^{t+\frac{1}{2}} \leftarrow \mathbf{y}_i^t + \nabla f_i(\mathbf{x}_i^t) - \nabla f_i(\mathbf{z}_i^t)$
    $\mathbf{x}_i^{t+\frac{1}{2}} \leftarrow \mathbf{x}_i^t - \eta \mathbf{y}_i^{t+\frac{1}{2}}$
    Set $\mathbf{z}_i^{t+1} \leftarrow \mathbf{x}_i^t$ and store $\nabla f_i(\mathbf{z}_i^{t+1})$
  else
    $\mathbf{y}_i^{t+\frac{1}{2}} \leftarrow \mathbf{y}_i^t$; $\mathbf{x}_i^{t+\frac{1}{2}} \leftarrow \mathbf{x}_i^t$; $\mathbf{z}_i^{t+1} \leftarrow \mathbf{z}_i^t$
  end if
  `Communication`
  $\mathbf{x}_i^{t+1} = \sum_{j\in\mathcal{V}} a_{ij}^t \mathbf{x}_j^{t+\frac{1}{2}}$ ▷ A_t is row-stochastic
  $\mathbf{y}_i^{t+1} = \sum_{j\in\mathcal{V}} b_{ij}^t \mathbf{y}_j^{t+\frac{1}{2}}$ ▷ B_t is column-stochastic
end for

2) Communication between active nodes: We can imagine a situation where only active agents participate in the communication. Then, these active agents may communicate with each other using mixing matrices $A(\mathcal{V}_t), B(\mathcal{V}_t)$ that are compatible with the induced subgraph $\mathcal{G}[\mathcal{V}_t]$, defined by the vertex set $\mathcal{V}_t$ and the edges of $\mathcal{E}$ that connect two vertices of $\mathcal{V}_t$. For example, if the graph is symmetric and $A(\mathcal{V}_t) = B(\mathcal{V}_t)$ is the Metropolis matrix of $\mathcal{G}[\mathcal{V}_t]$, we have $A_t = B_t$ and

$$a_{ij}^t = \begin{cases} \frac{1}{\max(\deg_t(i),\deg_t(j))} & \text{if } i \neq j \text{ and } \{i,j\} \in \mathcal{E} \cap 2^{\mathcal{V}_t},\\ 1 - \sum_{k=1}^M a_{ik}^t & \text{if } i = j,\\ 0 & \text{otherwise,} \end{cases}$$

where $\deg_t(i) = \operatorname{card}(\mathcal{N}_i \cap \mathcal{V}_t)$ is the degree of $i \in \mathcal{V}_t$ in the induced graph $\mathcal{G}[\mathcal{V}_t]$.

3) Broadcast-type update: As mentioned previously, our algorithm allows inactive nodes to passively receive information from active nodes. Therefore, the active nodes can simply broadcast their local variables to their neighbors, no matter whether these neighbors are active or not. To ensure the row-stochasticity of A_t, the received $\mathbf{x}_i^t$'s are averaged out. On the other hand, to guarantee the column-stochasticity of B_t, an active node divides its $\mathbf{y}_i^t$ by the number of nodes it sends the information to, as usually done in a push-sum scheme.

For concreteness, let us denote by $\mathcal{N}_{j,t}^{out}$ the set of neighbors that active worker $j \in \mathcal{V}_t$ transmit information to (including itself) in round t and by $\mathcal{N}_{i,t}^{in} = \{j \in \mathcal{V}_t : i \in \mathcal{N}_{j,t}^{out}\}$ the set of active workers that send information to i in this same

round. The mixing matrices A_t and B_t are then defined by

$$a_{ij}^t = \begin{cases} \frac{1}{\operatorname{card}(\mathcal{N}_{i,t}^{in} \cup \{i\})} & \text{if } j \in \mathcal{N}_{i,t}^{in} \cup \{i\}, \\ 0 & \text{otherwise;} \end{cases}$$

$$b_{ij}^t = \begin{cases} \frac{1}{\operatorname{card}(\mathcal{N}_{j,t}^{out})} & \text{if } j \in \mathcal{N}_{i,t}^{in}, \\ 1 & \text{if } j = i \text{ and } j \notin \mathcal{V}_t, \\ 0 & \text{otherwise.} \end{cases}$$

In this example, we see that our method offers an additional degree of freedom compared to G-Push–Pull [15] since in that algorithm $a_{ij}^t > 0$ only if $i, j \in \mathcal{V}_t$; this is not necessarily the case in our approach.

4) SAGA: SAGA [21] is a well-known (centralized) variance reduction methods that replaces the stochastic gradient $\nabla f_i(\mathbf{x}_t)$ with an unbiased gradient estimator with diminishing variance. For this, we store a table of gradients $(\nabla f_i(\mathbf{z}_i^t))_{i=1}^M$, where, similar to PPDS, $\mathbf{z}_i^t$ is the iterate at which ∇f_i was last evaluated. Let i_t be sampled from the index set $\{1, \dots, M\}$. The update of SAGA is then given by:

$$\boldsymbol{g}_t = \nabla f_{i_t}(\mathbf{x}_t) - \nabla f_{i_t}(\mathbf{z}_t) + \frac{1}{M} \sum_{i=1}^{M} \nabla f_i(\mathbf{z}_t), \quad (1)$$

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta \boldsymbol{g}_t.$$

To recover SAGA from PPDS, we set $A_t = B_t \equiv J$. This ensures $\mathbf{y}_i^t = (1/M) \sum_{i=1}^M \nabla f_i(\mathbf{z}_t)$ and thus $\mathbf{y}_i^{t+\frac{1}{2}}$ is exactly updated as in (1) when i is active. Specifically, if exactly one node is sampled at each iteration, (PPDS) with step-size η and $A_t = B_t \equiv J$ is equivalent to SAGA with stepsize η/M. If multiple workers are active at a same time slot, we get a mini-batch version of SAGA.

III. Linear convergence of PPDS

In this section, we present convergence guarantees of PPDS for strongly convex functions over a random network model. Concretely, we make the following standard convexity/smoothness assumption on the objective functions:

Assumption 1. All the individual f_i's are L-smooth and convex; the global function f is μ-strongly convex.

Thanks to the strong convexity of f, we know there exists a unique solution of (P) which we will denote by $\mathbf{x}_\star$. Moreover, we model $(\mathcal{V}_t)_{t\in\mathbb{N}}$, $(A_t)_{t\in\mathbb{N}}$, and $(B_t)_{t\in\mathbb{N}}$ as random variables satisfying that:

Assumption 2. The random variables $((\mathcal{V}_t, A_t, B_t))_{t\in\mathbb{N}}$ are temporally independent and identically distributed (i.i.d.).

Assumption 2 is actually only needed to provide the contractions of Lemmas 3 and 6. Hence, it could be weakened accordingly. We chose to keep it as such for ease of reading and for consistency with the literature.

A. The general case

First, we present our linear convergence result under rather weak assumptions on communications (essentially that the information can flow all over the network) and device sampling (each node is sampled with positive probability).

Assumption 3. The mixing matrices $(A_t)_{t\in\mathbb{N}}$ and $(B_t)_{t\in\mathbb{N}}$ have the following properties:

a) For all $t \in \mathbb{N}$, A_t is row-stochastic and B_t is column-stochastic.
b) Both $A \coloneqq \mathbb{E}[A_1]$ and $B \coloneqq \mathbb{E}[B_1]$ are primitive.
c) There exists $\nu > 0$ such that $a_{ii}^t \geq \nu$ and $b_{ii}^t \geq \nu$ for all $i \in \mathcal{V}, t \in \mathbb{N}$.

Assumption 4. Every node is sampled with positive probability, i.e., $p_i \coloneqq \mathbb{P}(i \in \mathcal{V}_1) > 0$ for all $i \in \mathcal{V}$.

It is straightforward to verify that Assumptions 2–4 are fulfilled in all the aforementioned examples. In particular, in Example 3, the primitivity of matrices A and B are ensured by the strong connectivity of the underlying graph $\mathcal{G}$ since $A_{ij} > 0$ if and only if $(j, i) \in \mathcal{E}$ if and only if $B_{ij} > 0$. On the other hand, Assumption 3c posits that at each iteration each nodes maintains a fraction of its previous states. This rules out the counterexamples in which the states of the active nodes are always overwritten by those of the inactive states. Under these fairly weak assumptions, we manage to prove the convergence of PPDS as stated in the following theorem.

Theorem 1. *Let Assumptions 1–4 hold. If* (PPDS) *is run with a sufficiently small step-size $\eta > 0$, then*

a) $\mathbf{x}_i^t$ converges almost surely to the solution $\mathbf{x}_\star$.
b) The expected squared distance between the iterate and the solution $\mathbb{E}[\|\mathbf{x}_i^t - \mathbf{x}_\star\|^2]$ vanishes geometrically.

Theorem 1 shows that the nice properties of gradient tracking and variance reduced methods are also preserved by our algorithm: it converges with constant step-size and enjoys a linear convergence rate as centralized gradient descent. Therefore, our method effectively reduces the variances of the noises induced by both sampling and communication.

We note that the assumptions for this result are quite similar to the ones for G-Push–Pull in [15], except that i) we do not put additional restrictions on the coefficients of the gossip matrix (unlike Eq. (24a) in [15]); and ii) we allow for a device sampling strategy that can be independent or correlated with the gossiping step.

B. The case of doubly stochastic matrices.

Due to the generality of the result, Theorem 1 only describes the qualitative behavior of the algorithm. To derive an explicit convergence rate, we focus on the specific situation where the mixing matrices are doubly stochastic and the active devices are sampled uniformly at random. Formally, we make the following assumptions.

Assumption 3′. For all $t \in \mathbb{N}$, both A_t and B_t are doubly stochastic. Moreover, we have the inequality

$$\lambda \coloneqq \max(\rho(\mathbb{E}[A_1^\top (I - J) A_1]), \rho(\mathbb{E}[B_1^\top (I - J) B_1])) < 1.$$

Assumption 4′. $\mathcal{V}_t$ is of fixed size S and is sampled uniformly from all the subsets of this size.

As detailed later in Section IV-A, the bistochasticity of the matrices and the condition $\lambda < 1$ allows us to derive a per-step contraction of the variance of the nodes' variables (see

Lemma 3). Using uniform sampling further facilitates the analysis and makes the final expression much more concise. The next theorem states the step-size condition and the convergence rate of PPDS when these assumptions are fulfilled.

Theorem 2. *Let Assumptions 1, 2, 3′ and 4′ hold. If* (PPDS) *is run with step-size*

$$\eta \leq \min\left(\frac{(1-\lambda)^2}{14L}\sqrt{\frac{M}{S}}, \frac{(1-\lambda)^2}{2304L}\left(\frac{M}{S}\right)^{\frac{3}{2}}, \frac{1}{576L}\sqrt{\frac{M}{S}}\right), \quad (2)$$

then the expected squared distance $\mathbb{E}[\|\mathbf{x}_i^t - \mathbf{x}_\star\|^2]$ *vanishes geometrically in* $\mathcal{O}(\gamma^t)$ *with* $\gamma = \max\left(1 - \frac{\eta\mu S}{2M}, 1 - \frac{S}{4M}\right)$. *In particular, it takes*

$$\mathcal{O}\left(\left(\frac{L}{\mu}\sqrt{\frac{M}{S}}\frac{1}{(1-\lambda)^2} + \frac{M}{S}\right)\log\left(\frac{1}{\varepsilon}\right)\right) \quad (3)$$

iterations to achieve ε *accuracy when* η *is suitably tuned.*

Theorem 2 indicates that a larger step-size can (and should) be taken for smaller sample size, if all the other parameters are fixed. Intuitively, this is because at each iteration fewer gradients enter the network, and thus these gradients can be used with a larger weight.

Although the complexity in terms of iterations is degraded by $\sqrt{M/S}$ compared to asynchronous Push–Pull without device sampling (i.e., $S = M$), the complexity in number of computed gradients is actually improved. To see this, we multiply (3) by S and and verify that the resulting quantity indeed decreases when S gets smaller.

Nonetheless, device sampling may also affect the connectivity of the network and thus λ if the communication matrices are chosen according to the sampled devices $\mathcal{V}_t$ (for instance, in Example 3). Therefore, unlike in the centralized case, sampling with variance reduction is not always guaranteed to converge faster here. Rather, there is a communication-computation trade-off in which the sampling size S plays a critical role.

IV. Convergence analysis

In this section we outline the proofs of Theorems 1 and 2. To begin, let us define

$$\boldsymbol{g}_i^t = \mathbf{y}_i^t + \nabla f_i(\mathbf{x}_i^t) - \nabla f_i(\mathbf{z}_i^t).$$

as the gradient estimator of node i at iteration t so that $\mathbf{y}_i^{t+\frac{1}{2}} = \boldsymbol{g}_i^t$ if and only if $i \in \mathcal{V}_t$, and $\mathbf{y}_i^{t+\frac{1}{2}} = \mathbf{y}_i^t$ otherwise. With the mass preservation property of column-stochastic matrices and the definition of $\mathbf{z}_i^t$, we have immediately the following lemma.

Lemma 1. *Suppose that the matrices* $(B_t)_{t\in\mathbb{N}}$ *are column-stochastic. It holds that*

$$\sum_{i=1}^M \mathbf{y}_i^t = \sum_{i=1}^M \nabla f_i(\mathbf{z}_i^t), \quad \sum_{i=1}^M \boldsymbol{g}_i^t = \sum_{i=1}^M \nabla f_i(\mathbf{x}_i^t).$$

Therefore, if the iterates move in the direction $-\sum_{i=1}^M \boldsymbol{g}_i^t$, we can expect convergence of the algorithm. This idea is crucial for our proof.

Another important step in the analysis is to establish that the nodes' decisions variables converge to a consensus. For this, let us write $\bar{\mathbf{x}}_t = \mathbf{1}^\top X_t/M$ for the average of these variables. Similarly, we also use the notation $\bar{\mathbf{y}}_t = \mathbf{1}^\top Y_t/M$.

Finally, we would like to highlight that the expectation $\mathbb{E}$ is taken over the randomness induced by both sampling and communication. We define $(\mathcal{F}_t)_{t\in\mathbb{N}}$ as the natural filtration associated to the sequence $(X_t)_{t\in\mathbb{N}}$ so that $((\mathcal{V}_s, A_s, B_s))_{1\leq s\leq t-1}$ is $\mathcal{F}_t$ measurable while $(\mathcal{V}_t, A_t, B_t)$ is not. For simplicity, we write $\mathbb{E}_t$ for the expectation conditioned on the history up to time t, i.e., $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot \mid \mathcal{F}_t] = \mathbb{E}[\cdot \mid ((\mathcal{V}_s, A_s, B_s))_{1\leq s\leq t-1}]$.

A. Analysis with doubly stochastic matrices

As a warm-up, we first establish the convergence of the algorithm in the simpler case where both Assumption 3′ and Assumption 4′ hold. This allows us to highlight our proof strategy without having to deal with the additional difficulties caused by the fact of having asymmetric communications. Following previous works that analyze gradient tracking and variance reduced methods, the essential idea of our proof is to derive a system of inequalities for the following quantities

$$\begin{aligned} d_t &= \mathbb{E}[\|\bar{\mathbf{x}}_t - \mathbf{x}_\star\|^2], \quad e_t = \mathbb{E}[f(\bar{\mathbf{x}}_t) - f(\mathbf{x}_\star)], \\ \rho_t &= \mathbb{E}[\|X_t - \mathbf{1}\bar{\mathbf{x}}_t^\top\|^2], \quad \zeta_t = \mathbb{E}[\|Y_t - \mathbf{1}\bar{\mathbf{y}}_t^\top\|^2], \\ \psi_t &= \sum_{i=1}^M \mathbb{E}[\|\nabla f_i(\mathbf{z}_i^t) - \nabla f_i(\mathbf{x}_\star)\|^2]. \end{aligned} \quad (4)$$

Here, d_t and e_t measure the performance of the averaged iterate; ρ_t and ζ_t measure the variances of the two variables of the agents; and ψ_t measures the quality of the control variates and is standard in the analysis of variance reduced algorithms [34], [35]. The following proposition bounds these quantities by a linear combination of their previous values.

Proposition 1. *Let* $\mathbf{r}_t = [d_t, \rho_t, \zeta_t, \psi_t]^\top$. *Under Assumptions 1, 2, 3′ and 4′, we have*

$$\mathbf{r}_{t+1} \leq Q\mathbf{r}_t + e_t\mathbf{h} \quad (5)$$

where the entries of Q *and* $\mathbf{h}$ *are given by*

$$Q = \begin{bmatrix} 1 - \frac{\eta\mu S}{2M} & \frac{\eta LS}{M^2} + \frac{10\eta^2L^2S^2}{M^3} & \frac{2\eta^2S^2}{M^3} & \frac{4\eta^2S^2}{M^3} \\ 0 & \frac{1+\lambda}{2} + \frac{20\eta^2L^2S}{M(1-\lambda)} & \frac{4\eta^2S}{M(1-\lambda)} & \frac{8\eta^2S}{M(1-\lambda)} \\ 0 & \frac{8L^2S}{M(1-\lambda)} & \frac{1+\lambda}{2} & \frac{4S}{M(1-\lambda)} \\ 0 & \frac{2L^2S}{M} & 0 & 1 - \frac{S}{M} \end{bmatrix},$$

$$\mathbf{h} = \left[-\frac{\eta S}{M} + \frac{20\eta^2LS^2}{M^2}, \frac{40\eta^2LS}{1-\lambda}, \frac{16LS}{1-\lambda}, 4LS\right]^\top.$$

To prove Proposition 1, we start by presenting a series of technical lemmas that are useful for this purpose. First, in order to deal with device sampling, we observe that $G_t = [\boldsymbol{g}_1^t, \ldots, \boldsymbol{g}_M^t]^\top$ plays an important role since $D_tY_{t+\frac{1}{2}} = D_tG_t$. With the uniform sampling of Assumption 4′, we obtain the following lemma.

Lemma 2. *Let Assumptions 2 and 4′ hold. Then*

a) $\mathbb{E}_t[\mathbf{1}^\top D_t Y_{t+\frac{1}{2}}] = \frac{S}{M}\sum_{i=1}^M \boldsymbol{g}_i^t$.

b) $\mathbb{E}_t[\|\mathbf{1}^\top D_t Y_{t+\frac{1}{2}}\|^2] \le \frac{S^2}{M}\sum_{i=1}^M \|\boldsymbol{g}_i^t\|^2$.

Proof. a) Note that $\mathbf{1}^\top D_t Y_{t+\frac{1}{2}} = \sum_{i\in\mathcal{V}_t} \mathbf{y}_i^t = \sum_{i\in\mathcal{V}_t} \boldsymbol{g}_i^t$. Therefore,

$$\mathbb{E}_t[\mathbf{1}^\top D_t Y_{t+\frac{1}{2}}] = \mathbb{E}_t\left[\sum_{i\in\mathcal{V}_t} \boldsymbol{g}_i^t\right] = \mathbb{E}_t\left[\sum_{i=1}^M \mathbb{1}_{i\in\mathcal{V}_t}\boldsymbol{g}_i^t\right] = \sum_{i=1}^M \boldsymbol{g}_i^t\,\mathbb{E}_t[\mathbb{1}_{i\in\mathcal{V}_t}] = \frac{S}{M}\sum_{i=1}^M \boldsymbol{g}_i^t.$$

We can put g_i^t outside the expectation since it is $\mathcal{F}_t$-measurable.

b) Similarly, we have

$$\mathbb{E}_t[\|\mathbf{1}^\top D_t Y_{t+\frac{1}{2}}\|^2] = \mathbb{E}_t\left[\left\|\sum_{i\in\mathcal{V}_t}\boldsymbol{g}_i^t\right\|^2\right] \le \mathbb{E}_t\left[S\sum_{i\in\mathcal{V}_t}\|\boldsymbol{g}_i^t\|^2\right] = S\sum_{i=1}^M \mathbb{E}_t[\mathbb{1}_{i\in\mathcal{V}_t}\|\boldsymbol{g}_i^t\|^2] = \frac{S^2}{M}\sum_{i=1}^M\|\boldsymbol{g}_i^t\|^2.$$

□

To control the distance to consensus, we use the lemma below that shows a contraction property of the mixing matrices.

Lemma 3. *Let Assumptions 2 and 3′ hold. Then*

a) $\mathbb{E}_t[\|A_tX_t - \mathbf{1}\bar{\mathbf{x}}_t^\top\|^2] \le \lambda\|X_t - \mathbf{1}\bar{\mathbf{x}}_t^\top\|^2$.

b) $\mathbb{E}_t[\|B_tY_t - \mathbf{1}\bar{\mathbf{y}}_t^\top\|^2] \le \lambda\|Y_t - \mathbf{1}\bar{\mathbf{y}}_t^\top\|^2$.

Proof. Since A_t is doubly stochastic, we can write

$$\begin{aligned}
&\mathbb{E}_t[\|A_tX_t - \mathbf{1}\bar{\mathbf{x}}_t^\top\|^2]\\
&= \mathbb{E}_t[\|(I-J)A_t(X_t - \mathbf{1}\bar{\mathbf{x}}_t^\top)\|^2]\\
&= \mathbb{E}_t[\mathrm{tr}[(X_t - \mathbf{1}\bar{\mathbf{x}}_t^\top)^\top A_t^\top (I-J)^2 A_t(X_t - \mathbf{1}\bar{\mathbf{x}}_t^\top)]]\\
&= \mathrm{tr}[(X_t - \mathbf{1}\bar{\mathbf{x}}_t^\top)^\top \mathbb{E}_t[A_t^\top(I-J)A_t](X_t - \mathbf{1}\bar{\mathbf{x}}_t^\top)]\\
&\le \rho(\mathbb{E}_t[A_t^\top(I-J)A_t])\|X_t - \mathbf{1}\bar{\mathbf{x}}_t^\top\|^2.
\end{aligned}$$

Under Assumption 2 we have $\mathbb{E}_t[A_t^\top(I-J)A_t] = \mathbb{E}[A_1^\top(I-J)A_1]$ and *a)* follows immediately given that $\rho(\mathbb{E}[A_1^\top(I-J)A_1]) \le \lambda$. Property *b)* is proved in the same way. □

Finally, we can use the smoothness of the objective functions and the optimality conditions to bound the expected squared norm of G_t and gradients differences by the quantities introduced in (4).

Lemma 4. *Let Assumption 1 hold and $(B_t)_{t\in\mathbb{N}}$ be column-stochastic. We have*

a) $\mathbb{E}[\|\nabla F(X_t) - \nabla F(\mathbf{1}^\top\mathbf{x}_\star)\|^2] \le 2L^2\rho_t + 4MLe_t$.

b) $\mathbb{E}[\|\nabla F(X_t) - \nabla F(Z_t)\|^2] \le 4L^2\rho_t + 8MLe_t + 2\psi_t$.

c) $\mathbb{E}[\|G_t\|^2] \le 10L^2\rho_t + 20MLe_t + 4\psi_t + 2\zeta_t$.

Proof. See Appendix A. □

We are now ready to prove Proposition 1 by leveraging the above lemmas.

Proof of Proposition 1. Below we bound the four quantities in question respectively.

Bounding d_{t+1}. We develop

$$\begin{aligned}
\|\bar{\mathbf{x}}_{t+1} - \mathbf{x}_\star\|^2 &= \|\bar{\mathbf{x}}_t - \frac{\eta}{M}\mathbf{1}^\top D_t Y_{t+\frac{1}{2}} - \mathbf{x}_\star\|^2\\
&= \|\bar{\mathbf{x}}_t - \mathbf{x}_\star\|^2 - \frac{2\eta}{M}\langle\bar{\mathbf{x}}_t - \mathbf{x}_\star, \mathbf{1}^\top D_t Y_{t+\frac{1}{2}}\rangle\\
&\quad + \frac{\eta^2}{M^2}\|\mathbf{1}^\top D_t Y_{t+\frac{1}{2}}\|^2. \qquad (6)
\end{aligned}$$

Using Lemmas 1 and 2 and Assumption 1, we get

$$\begin{aligned}
&\mathbb{E}_t[\langle\bar{\mathbf{x}}_t - \mathbf{x}_\star, \mathbf{1}^\top D_t Y_{t+\frac{1}{2}}\rangle]\\
&= \langle\bar{\mathbf{x}}_t - \mathbf{x}_\star, \frac{S}{M}\sum_{i=1}^M \nabla f_i(\mathbf{x}_i^t)\rangle\\
&= \frac{S}{M}\sum_{i=1}^M(\langle\bar{\mathbf{x}}_t - \mathbf{x}_i^t, \nabla f_i(\mathbf{x}_i^t)\rangle + \langle\mathbf{x}_i^t - \mathbf{x}_\star, \nabla f_i(\mathbf{x}_i^t)\rangle)\\
&\ge \frac{S}{M}\sum_{i=1}^M(f_i(\bar{\mathbf{x}}_t) - f_i(\mathbf{x}_i^t) - \frac{L}{2}\|\mathbf{x}_i^t - \bar{\mathbf{x}}_t\|^2 + f_i(\mathbf{x}_i^t) - f_i(\mathbf{x}_\star))\\
&= S(f(\bar{\mathbf{x}}_t) - f(\mathbf{x}_\star)) - \frac{LS}{2M}\|X_t - \mathbf{1}\bar{\mathbf{x}}_t^\top\|^2\\
&\ge \frac{S}{2}(f(\bar{\mathbf{x}}_t) - f(\mathbf{x}_\star)) + \frac{\mu S}{4}\|\bar{\mathbf{x}}_t - \mathbf{x}_\star\|^2 - \frac{LS}{2M}\|X_t - \mathbf{1}\bar{\mathbf{x}}_t^\top\|^2. \qquad (7)
\end{aligned}$$

In the last line we have used the fact that $f(\mathbf{x}) - f(\mathbf{x}_\star) \ge (\mu/2)\|\mathbf{x} - \mathbf{x}_\star\|^2$ for every $\mathbf{x}\in\mathbb{R}^d$ since f is strongly convex. As for the last term of (6), we resort to Lemma 2b and Lemma 4c. This gives

$$\mathbb{E}[\|\mathbf{1}^\top D_t Y_{t+\frac{1}{2}}\|^2] \le \frac{S^2}{M}(10L^2\rho_t + 20MLe_t + 4\psi_t + 2\zeta_t).$$

Combining the above inequalities we get

$$\begin{aligned}
d_{t+1} \le \left(1 - \frac{\eta\mu S}{2M}\right)d_t + \left(\frac{\eta LS}{M^2} + \frac{10\eta^2L^2S^2}{M^3}\right)\rho_t\\
+ \frac{2\eta^2S^2}{M^3}\zeta_t + \frac{4\eta^2S^2}{M^3}\psi_t - \left(\frac{\eta S}{M} - \frac{20\eta^2LS^2}{M^2}\right)e_t.
\end{aligned}$$

Bounding ρ_{t+1}. In the inequality $\|a+b\|^2 \le (1+\delta)\|a\|^2 + (1+1/\delta)\|b\|^2$, choosing $\delta = (1-\lambda)/2\lambda$ gives[2]

$$\|a+b\|^2 \le \frac{1+\lambda}{2\lambda}\|a\|^2 + \frac{1+\lambda}{1-\lambda}\|b\|^2. \qquad (8)$$

Since A_t is doubly stochastic and hence column-stochastic, it holds $JA_t = J$. We then have,

$$\begin{aligned}
&\mathbb{E}_t[\|X_{t+1} - \mathbf{1}\bar{\mathbf{x}}_{t+1}^\top\|^2]\\
&= \mathbb{E}_t[\|A_tX_t - \eta A_tD_tY_{t+\frac{1}{2}} - (\mathbf{1}\bar{\mathbf{x}}_t^\top - \eta JD_tY_{t+\frac{1}{2}})\|^2]\\
&\le \frac{1+\lambda}{2\lambda}\mathbb{E}_t[\|A_tX_t - \mathbf{1}\bar{\mathbf{x}}_t^\top\|^2]\\
&\quad + \frac{1+\lambda}{1-\lambda}\eta^2\,\mathbb{E}_t[\|A_tD_tY_{t+\frac{1}{2}} - JD_tY_{t+\frac{1}{2}}\|^2]. \qquad (9)
\end{aligned}$$

[2]Without loss of generality we assume $\lambda > 0$. Otherwise the first term in the inequalities are always 0 and we can simply take $\delta = 1$. The same remark applies to the analysis in Section IV-B.

Using Lemma 3a the first term can be bounded by $(1+\lambda)\|X_t - \mathbf{1}\bar{\mathbf{x}}_t^\top\|^2/2$. The same does not apply to the second term as A_t and D_t are not independent. Nonetheless, with the bistochasticity of A_t, we can still write

$$\mathbb{E}_t[\|A_tD_tY_{t+\frac{1}{2}} - JD_tY_{t+\frac{1}{2}}\|^2] \le \mathbb{E}_t[\|D_tY_{t+\frac{1}{2}}\|^2] = \frac{S}{M}\sum_{i=1}^M \|\boldsymbol{g}_i^t\|^2.$$

With Lemma 4c, taking total expectation in (9) then gives

$$\begin{aligned}\rho_{t+1} &\le \frac{1+\lambda}{2}\rho_t + \frac{1+\lambda}{1-\lambda}\frac{\eta^2 S}{M}(10L^2\rho_t + 20MLe_t + 4\psi_t + 2\zeta_t)\\ &\le \left(\frac{1+\lambda}{2} + \frac{20\eta^2L^2S}{M(1-\lambda)}\right)\rho_t + \frac{4\eta^2S}{M(1-\lambda)}\zeta_t\\ &\quad + \frac{8\eta^2S}{M(1-\lambda)}\psi_t + \frac{40\eta^2LS}{1-\lambda}e_t.\end{aligned}$$

Bounding ζ_{t+1}. Similar to the above, using (8) and the bistochasticity of B_t, we obtain

$$\begin{aligned}\|Y_{t+1} - \mathbf{1}\bar{\mathbf{y}}_{t+1}^\top\|^2 &= \|B_tY_t - B_tD_t(\nabla F(X_t) - \nabla F(Z_t))\\ &\quad - (\mathbf{1}\bar{\mathbf{y}}_t^\top - JD_t(\nabla F(X_t) - \nabla F(Z_t)))\|^2\\ &\le \frac{1+\lambda}{2\lambda}\|B_tY_t - \mathbf{1}\bar{\mathbf{y}}_t^\top\|^2\\ &\quad + \frac{1+\lambda}{1-\lambda}\|D_t(\nabla F(X_t) - \nabla F(Z_t))\|^2. \quad (10)\end{aligned}$$

The uniform sampling assumption implies that

$$\begin{aligned}&\mathbb{E}_t[\|D_t(\nabla F(X_t) - \nabla F(Z_t))\|^2]\\ &= \mathbb{E}_t\left[\sum_{i\in\mathcal{V}_t}\|\nabla f_i(\mathbf{x}_i^t) - \nabla f_i(\mathbf{z}_i^t)\|^2\right]\\ &= \frac{S}{M}\sum_{i=1}^M\|\nabla f_i(\mathbf{x}_i^t) - \nabla f_i(\mathbf{z}_i^t)\|^2.\end{aligned}$$

Taking expectation in (10) and applying Lemma 3b and Lemma 4b then yields

$$\begin{aligned}\zeta_{t+1} &\le \frac{1+\lambda}{2}\zeta_t + \frac{1+\lambda}{1-\lambda}\frac{S}{M}(4L^2\rho_t + 8MLe_t + 2\psi_t)\\ &\le \frac{1+\lambda}{2}\zeta_t + \frac{8L^2S}{M(1-\lambda)}\rho_t + \frac{4S}{M(1-\lambda)}\psi_t + \frac{16LS}{1-\lambda}e_t.\end{aligned}$$

Bounding ψ_{t+1}. Note that by the update rule of $\mathbf{z}_i^t$, we have

$$\begin{aligned}&\mathbb{E}_t[\|\nabla f_i(\mathbf{z}_i^{t+1}) - \nabla f_i(\mathbf{x}_\star)\|^2]\\ &= \left(1 - \frac{S}{M}\right)\|\nabla f_i(\mathbf{z}_i^t) - \nabla f_i(\mathbf{x}_\star)\|^2\\ &\quad + \frac{S}{M}\|\nabla f_i(\mathbf{x}_i^t) - \nabla f_i(\mathbf{x}_\star)\|^2.\end{aligned}$$

Summing from $i=1$ to M, applying Lemma 4a and taking total expectation, we get

$$\begin{aligned}\psi_{t+1} &\le \left(1 - \frac{S}{M}\right)\psi_t + \frac{S}{M}(2L^2\rho_t + 4MLe_t)\\ &= \left(1 - \frac{S}{M}\right)\psi_t + \frac{2L^2S}{M}\rho_t + 4LSe_t.\end{aligned}$$

Conclude. Putting all together we get exactly (5). □

From the linear system of inequalities (5) there are multiple ways to derive the linear convergence of the algorithm. To obtain the explicit convergence rate and step-size condition presented in Theorem 2, we construct a suitable Lyapunov function which is a linear combination of d_t, ρ_t, ζ_t, and ψ_t with positive coefficients, and prove that this function decreases geometrically at each iteration.

Proof of Theorem 2. Let us consider the vector

$$\omega = \begin{bmatrix}1 & \frac{\sqrt{S}(1-\lambda)}{M^{\frac{3}{2}}} & \frac{\eta(1-\lambda)}{96ML} & \frac{\eta}{12ML}\end{bmatrix}^\top,$$

and γ as defined in Theorem 2, it can be verified that Proposition 1 implies

$$\omega^\top\mathbf{r}_{t+1} \le \gamma\,\omega^\top\mathbf{r}_t \quad (11)$$

whenever step-size condition (2) is satisfied. This means $\omega^\top\mathbf{r}_t$ converges geometrically in $\mathcal{O}(\gamma^t)$. To conclude, we use the inequality

$$\mathbb{E}[\|\mathbf{x}_i^t - \mathbf{x}_\star\|^2] \le \mathbb{E}[2\|\mathbf{x}_i^t - \bar{\mathbf{x}}_t\|^2 + 2\|\bar{\mathbf{x}}_t - \mathbf{x}_\star\|^2] \le 2\rho_t + 2d_t.$$

Detailed computations for proving (11) are provided in Appendix B. □

B. Analysis for the general case

Under our weakest set of assumptions (Assumptions 3 and 4), the mixing matrices do not provide a contraction towards a consensus at each iteration. Nevertheless, the primitivity of the mixing matrices in expectation enables us to show that after a certain number of gossip steps l (implicitly defined), some sort of contraction happens for both matrices sequences but with respect to a time-varying weighted average instead of a uniform one.

This has direct consequences on our proof technique since the linear system of equations developed previously has to be modified and in particular extended to track l successive iterations. With this augmentation, the proof techniques developed before do not hold anymore and we resort to analyzing the spectral radius of the recurrence matrix by perturbation theory arguments when the stepsize is small.

These two points significantly complicate the convergence proof of the method and constitute the main technical contributions of the paper.

1) Multi-step contraction: To establish the multi-step contraction brought by the mixing matrices, we first leverage the primitivity assumption on $A = \mathbb{E}[A_1]$ and $B = \mathbb{E}[B_1]$ to show that inequalities similar to the one in Assumption 3′ hold when we consider the product of successive matrices, which we abbreviate as[3]

$$A_{t:s} = A_tA_{t-1}\dots A_s, \quad B_{t:s} = B_tB_{t-1}\dots B_s.$$

The following lemma generalizes Assumption 3′ and is useful for deriving inequalities in the form of Lemma 3.

[3]If $t < s$ we use the notation $A_{t:s} = B_{t:s} = I$.

Lemma 5. *Let Assumptions 2 and 3 hold. Then, there exists an integer l such that*

$$\rho(\mathbb{E}[A_{1+l:1}^\top (I-J)A_{1+l:1}]) < 1,$$
$$\rho(\mathbb{E}[(I-J)^\top B_{1+l:1}^\top B_{1+l:1}(I-J)]) < 1.$$

Proof. We will write $\|W\|$ and $\|W\|_{\mathrm{F}}$ respectively for the spectral norm and the Frobenius norm of a matrix W. Lemma 5 is an immediate result of [36, Prop. 2], which states that $\mathbb{E}[\|(I-J)A_{1+l:1}\|_{\mathrm{F}}^2]$ converges to 0 at a geometric rate. We can thus set l sufficiently large so that $\mathbb{E}[\|(I-J)A_{1+l:1}\|_{\mathrm{F}}^2] < 1$, and the first inequality then follows from that

$$\begin{aligned}\rho(\mathbb{E}[A_{1+l:1}^\top (I-J)A_{1+l:1}]) &\le \mathbb{E}[\rho(A_{1+l:1}^\top (I-J)A_{1+l:1})] \\ &= \mathbb{E}[\|(I-J)A_{1+l:1}\|^2] \\ &\le \mathbb{E}[\|(I-J)A_{1+l:1}\|_{\mathrm{F}}^2],\end{aligned}$$

where we have used the convexity of the spectral radius function ρ and the fact that the spectral norm of a matrix is bounded from above by its Frobenius norm.

For the second inequality, we observe that the matrices $(B_t^\top)_{t\in\mathbb{N}}$ have exactly the same assumptions as $(A_t)_{t\in\mathbb{N}}$. Moreover,

$$\begin{aligned}&\rho(\mathbb{E}[(I-J)^\top B_{1+l:1}^\top B_{1+l:1}(I-J)]) \\ &\le \mathbb{E}[\rho((I-J)^\top B_{1+l:1}^\top B_{1+l:1}(I-J))] \\ &= \mathbb{E}[\|B_{1+l:1}(I-J)\|^2] \\ &= \mathbb{E}[\|(I-J)B_1^\top \dots B_{1+l}^\top\|^2].\end{aligned}$$

Hence the same argument applies. □

Another important challenge towards proving a result in the spirit of Lemma 3 is that the matrices $(A_t)_{t\in\mathbb{N}}$ (resp. $(B_t)_{t\in\mathbb{N}}$) do not have a fixed left (resp. right) Perron vector, and as a consequence there are not predetermined values that the variables should converge to after the mixing matrices are applied. To overcome this difficulty, we instead introduce two sequence of random vectors $(\mathbf{v}_t)_{t\in\mathbb{N}}$ and $(\mathbf{u}_t)_{1\le t\le T}$. Here T is a positive integer fixed in advance. Let π_A be the left Perron vector of A such that $\mathbf{1}^\top \pi_A = 1$. These sequences are defined recursively by

$$\mathbf{v}_1 = \frac{1}{M}\mathbf{1},\ \mathbf{v}_{t+1} = B_t\mathbf{v}_t;\quad \mathbf{u}_T = \pi_A,\ \mathbf{u}_{t+1}^\top A_t = \mathbf{u}_t.$$

The sequence $(\mathbf{u}_t)_{1\le t\le T}$ is defined in a time-reversed manner and mimics the absolute probability sequence [37], [38] that can be defined for $(A_t)_{t\in\mathbb{N}}$. However, the above construction gives an explicit expression of $\mathbf{u}_t$ which turns out to be useful for our proof. Also notice that the value of $\mathbf{u}_t$ is dependent on the choice of T though this is implicit from the notation.

Since the $(B_t)_{t\in\mathbb{N}}$ are column-stochastic and the $(A_t)_{t\in\mathbb{N}}$ are row-stochastic, one deduces immediately that both $(\mathbf{v}_t)_{t\in\mathbb{N}}$ and $(\mathbf{u}_t)_{1\le t\le T}$ are sequences of probability vectors. Moreover, under Assumption 2 we have $\mathbb{E}[\mathbf{u}_t] = \mathbb{E}[\mathbf{u}_{t+1}^\top]\,\mathbb{E}[A_t] = \mathbb{E}[\mathbf{u}_{t+1}^\top]A$. By induction we then get

$$\mathbb{E}[\mathbf{u}_t] = \pi_A, \quad \forall t \in \{1,\dots,T\}. \tag{12}$$

In the remainder of the section, we will take $l \ge 0$ such that the inequalities of Lemma 5 are satisfied and define

$$\begin{aligned}\lambda = \max(&\rho(\mathbb{E}[A_{1+l:1}^\top (I-J)A_{1+l:1}]), \\ &\rho(\mathbb{E}[(I-J)^\top B_{1+l:1}^\top B_{1+l:1}(I-J)]))\end{aligned} \tag{13}$$

so that $\lambda < 1$. The multi-step contraction property is stated as follows.

Lemma 6. *Let Assumptions 2 and 3 hold. Take l as in Lemma 5 and λ from (13). Then,*

a) $\mathbb{E}_t[\|(I-J)A_{t+l:t}X_t\|^2] \le \lambda\|X_t - \mathbf{1}\bar{\mathbf{x}}_t^\top\|^2$.
b) $\mathbb{E}_t[\|B_{t+l:t}(I-\mathbf{v}_t\mathbf{1}^\top)Y_t\|^2] \le \lambda\|(I-\mathbf{v}_t\mathbf{1}^\top)Y_t\|^2$.

Proof. The lemma is proved exactly in the same way as Lemma 3. Just notice that

$$\begin{aligned}(I-J)A_{t+l:t}X_t &= (I-J)A_{t+l:t}(I-J)X_t \\ &= (I-J)A_{t+l:t}(X_t - \mathbf{1}\bar{\mathbf{x}}_t^\top)\end{aligned}$$

since $A_{t+l:t}$ is row-stochastic. On the other hand,

$$B_{t+l:t}(I-\mathbf{v}_t\mathbf{1}^\top)Y_t = B_{t+l:t}(I-J)(I-\mathbf{v}_t\mathbf{1}^\top)Y_t.$$

since $\mathbf{v}_t$ is a probability vector. □

2) Linear system of inequalities: As in Section IV-A, the proof for Theorem 1 also relies on the derivation of a linear system of inequalities. Nevertheless, since there is a contraction only every $l+1$ steps, we need to take into account the values of relevant quantities for $l+1$ consecutive iterations and the system becomes $l+1$ times larger. Given that the mixing matrices are no longer doubly stochastic, the variables that come into play also need to be modified accordingly. We consider the following quantities

$$\begin{aligned}&d_t' = \mathbb{E}[\|\mathbf{u}_t^\top X_t - \mathbf{x}_\star\|^2],\quad e_t = \mathbb{E}[f(\bar{\mathbf{x}}_t) - f(\mathbf{x}_\star)], \\ &\rho_t = \mathbb{E}[\|X_t - \mathbf{1}\bar{\mathbf{x}}_t^\top\|^2],\quad \zeta_t' = \mathbb{E}[\|Y_t - \mathbf{v}_t\mathbf{1}^\top Y_t\|], \\ &\psi_t' = \mathbb{E}[\|\nabla F(X_t) - \nabla F(Z_t)\|^2].\end{aligned}$$

Compared to (4), we define d_t' because we no longer have $\mathbb{1}^\top A_t X_t = \mathbb{1}^\top X_t$ while it holds $\mathbf{u}_{t+1}A_tX_t = \mathbf{u}_tX_t$. The definition of ζ_t' is consistent with Lemma 6b. Finally, we also replace ψ_t by ψ_t' for technical reasons. Note that the value of d_t' depends on T since its definition involves $\mathbf{u}_t$.

The following two lemmas collects several inequalities that will be useful for our proof.

Lemma 7. *It holds that*

a) $\|D_t\| \le 1$.
b) $\|\mathbf{u}_t^\top X_t - \bar{\mathbf{x}}_t\|^2 \le \|X_t - \mathbf{1}\bar{\mathbf{x}}_t^\top\|^2$.
c) The spectral norm of a row- or column-stochastic matrix of size $M\times M$ is not larger than $\sqrt{M}$.

Proof. a) is trivial and *c)* can be proven by using the fact that the spectral norm of a matrix is bounded by its Frobenius norm. As for *b)*, since $\mathbf{u}_t$ is a probability vector,

$$\begin{aligned}\|\mathbf{u}_t^\top X_t - \bar{\mathbf{x}}_t\|^2 = \left\|\sum_{i=1}^M u_i^t\mathbf{x}_i^t - \bar{\mathbf{x}}_t\right\|^2 &\le \sum_{i=1}^M u_i^t\|\mathbf{x}_i^t - \bar{\mathbf{x}}_t\|^2 \\ &\le \sum_{i=1}^M \|\mathbf{x}_i^t - \bar{\mathbf{x}}_t\|^2 = \|X_t - \mathbf{1}\bar{\mathbf{x}}_t^\top\|^2.\end{aligned}$$

In the above we have used the notation $\mathbf{u}_t = (u_i^t)_{i\in\mathcal{V}}$. □

Lemma 8. *Let Assumption 1 hold and $(B_t)_{t\in\mathbb{N}}$ be column-stochastic. We have*

a) $\mathbb{E}[\|G_t - \mathbf{v}_t\mathbf{1}^\top G_t\|^2] \le 2\zeta'_t + (4M+4)\psi'_t$.

b) $\mathbb{E}[\|G_t\|^2] \le 4ML^2\rho_t + (8M+8)\psi'_t + 4\zeta'_t + 8M^2Le_t$.

Proof. See Appendix C. □

Since the sampling is not uniform, Lemma 2 does not hold anymore and we need to approximate G_t by $\mathbf{v}_t\mathbf{1}^\top G_t$ when deriving the descent inequality. Given the definition of d'_t and the fact that the nodes are sampled, we say that the *effective step-size* at time t is $\eta\alpha_t$ with

$$\alpha_t = \mathbf{u}_t^\top D_t \mathbf{v}_t.$$

The following lemma controls $\mathbb{E}[\alpha_t\chi_t]$ for any real-valued non-negative random variable χ_t that is $\mathcal{F}_t$-measurable.

Lemma 9. *Let Assumptions 2–4 hold. We define $\underline{p} = \min_{i\in\mathcal{V}} p_i$, $\underline{\pi_A} = \min_{i\in\mathcal{V}}[\pi_A]_i$, and $\underline{\alpha} = \underline{\pi_A}\nu\underline{p}$. Then, $\underline{\alpha} > 0$ and for any $\mathcal{F}_t$-measurable real-valued non-negative random variable χ_t, we have*

$$\underline{\alpha}\,\mathbb{E}[\chi_t] \le \mathbb{E}[\alpha_t\chi_t] \le \mathbb{E}[\chi_t]. \tag{14}$$

Proof. See Appendix D. □

We are now ready to state and prove the linear system of inequalities in question. We denote by $P\otimes Q$ the Kronecker product of two matrices P and Q, and write E^k_{ij} for the matrix of size $k\times k$ that has a single non-zero entry with value 1 at position (i,j).

Proposition 2. *For $t\in\{1,...,T-l\}$, let $\underline{\alpha}$ be defined as in Lemma 9 and $\mathbf{r}'_t \in \mathbb{R}^{4(l+1)}$ be defined by*

$$\mathbf{r}'_t = [d'_{t+l}\ \dots\ d'_t\ \rho_{t+l}\ \dots\ \rho_t\ \psi'_{t+l}\ \dots\ \psi'_t\ \zeta'_{t+l}\ \dots\ \zeta'_t]^\top.$$

We also define $W_1, W_2 \in \mathbb{R}^{(l+1)\times(l+1)}$ as

$$W_1 = \begin{bmatrix} 0 & \cdots & \cdots & \cdots & 0 \\ 1 & \ddots & & & \vdots \\ 0 & \ddots & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & 1 & 0 \end{bmatrix}, \quad W_2 = \begin{bmatrix} 1 & \cdots & \cdots & \cdots & 1 \\ 0 & \cdots & \cdots & \cdots & 0 \\ \vdots & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & \cdots & 0 \end{bmatrix}.$$

Then, under Assumptions 1–4, if PPDS is run with $\eta \le \underline{\alpha}/(16ML)$, we have

$$\mathbf{r}'_{t+1} \le (Q_0 + \eta Q_e)\mathbf{r}'_t \tag{15}$$

where

$$Q_0 = I_4\otimes W_1 + c_{13}E^4_{4,3}\otimes W_2 + \frac{1+\lambda}{2}(E^4_{1,1}+E^4_{3,3})\otimes E^{l+1}_{1,l+1} + \left(E^4_{1,1} + \left(1-\frac{\underline{p}}{2}\right)E^4_{3,3} + c_{10}E^4_{3,2}\right)\otimes E^{l+1}_{1,1}.$$

and

$$Q_e = \begin{bmatrix} -c_1 & c_2 & c_3 & c_4 \\ 0 & 0 & 0 & 0 \\ c_9 & 0 & c_{11} & c_{12} \\ 0 & 0 & 0 & 0 \end{bmatrix}\otimes E^{l+1}_{1,1} + \begin{bmatrix} 0 & 0 & 0 & 0 \\ c_5 & c_6 & c_7 & c_8 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}\otimes W_2$$

are defined with positive constants $(c_k)_{1\le k\le 13}$ that are entirely determined by $\mu, L, M, \lambda, l, \underline{p}$, and $\underline{\alpha}$.

Proof. We will make use of the inequality

$$e_t \le L^2(d'_t + \rho_t). \tag{16}$$

This comes from the simple fact that

$$\begin{aligned} f(\bar{\mathbf{x}}_t) - f(\mathbf{x}_\star) &\le \frac{L^2}{2}\|\bar{\mathbf{x}}_t - \mathbf{x}_\star\|^2 \\ &\le L^2(\|\mathbf{u}_t^\top X_t - \bar{\mathbf{x}}_t\|^2 + \|\mathbf{u}_t^\top X_t - \mathbf{x}_\star\|^2) \\ &\le L^2(\|X_t - \mathbf{1}\bar{\mathbf{x}}_t^\top\|^2 + \|\mathbf{u}_t^\top X_t - \mathbf{x}_\star\|^2). \end{aligned}$$

Also notice that $\|D_tY_{t+\frac12}\|$ can be bounded as

$$\|D_tY_{t+\frac12}\| = \|D_tG_t\| \le \|D_t\|\|G_t\| \le \|G_t\|. \tag{17}$$

Now, let us fix $t\in\{l+1,...,T-1\}$. We bound ρ_{t+1}, ζ'_{t+1}, ψ'_{t+1}, and d'_{t+1} in terms of the previous values of these same variables.

Bounding d'_{t+1}. We decompose

$$\begin{aligned} &\|\mathbf{u}_{t+1}^\top X_{t+1} - \mathbf{x}_\star\|^2 \\ &= \|\mathbf{u}_{t+1}^\top A_tX_t - \eta\mathbf{u}_{t+1}^\top A_tD_tY_{t+\frac12} - \mathbf{x}_\star\|^2 \\ &= \|\mathbf{u}_t^\top X_t - \mathbf{x}_\star\|^2 + \eta^2\|\mathbf{u}_t^\top D_tY_{t+\frac12}\|^2 \\ &\quad - 2\eta\langle\mathbf{u}_t^\top X_t - \mathbf{x}_\star, \mathbf{u}_t^\top D_tY_{t+\frac12}\rangle. \end{aligned}$$

With (17), the second term can be easily bounded using

$$\|\mathbf{u}_t^\top D_tY_{t+\frac12}\| \le \|\mathbf{u}_t\|\|D_tY_{t+\frac12}\| \le \|G_t\|.$$

As for the third term, it can be further decomposed as

$$\begin{aligned} &\langle\mathbf{u}_t^\top X_t - \mathbf{x}_\star, \mathbf{u}_t^\top D_tY_{t+\frac12}\rangle \\ &= \langle\mathbf{u}_t^\top X_t - \bar{\mathbf{x}}_t, \mathbf{u}_t^\top D_tG_t\rangle \\ &\quad + \langle\bar{\mathbf{x}}_t - \mathbf{x}_\star, \mathbf{u}_t^\top D_t(G_t - \mathbf{v}_t\mathbf{1}^\top G_t)\rangle \\ &\quad + \langle\bar{\mathbf{x}}_t - \mathbf{x}_\star, \mathbf{u}_t^\top D_t\mathbf{v}_t\mathbf{1}^\top G_t\rangle, \end{aligned} \tag{18}$$

where we used again $D_tY_{t+\frac12} = D_tG_t$. Let us bound the three terms separately. Using Lemma 8b, for any $\delta_1 > 0$, we have

$$\begin{aligned} &\mathbb{E}[-2\eta\langle\mathbf{u}_t^\top X_t - \bar{\mathbf{x}}_t, \mathbf{u}_t^\top D_tG_t\rangle] \\ &\le \mathbb{E}[\eta\delta_1\|\mathbf{u}_t^\top X_t - \bar{\mathbf{x}}_t\|^2 + \frac{\eta}{\delta_1}\|\mathbf{u}_t^\top D_tG_t\|^2] \\ &\le \eta\delta_1\,\mathbb{E}[\|X_t - \mathbf{1}\bar{\mathbf{x}}_t^\top\|^2] + \frac{\eta}{\delta_1}\,\mathbb{E}[\|G_t\|^2]] \\ &\le \eta\delta_1\rho_t + \frac{\eta}{\delta_1}(4ML^2\rho_t + (8M+8)\psi'_t + 4\zeta'_t + 8M^2Le_t). \end{aligned}$$

With Lemma 7b and Lemma 8a, we can bound the second term of (18) for any $\delta_2 > 0$ as

$$\begin{aligned} &\mathbb{E}[-2\eta\langle\bar{\mathbf{x}}_t - \mathbf{x}_\star, \mathbf{u}_t^\top D_t(G_t - \mathbf{v}_t\mathbf{1}^\top G_t)\rangle] \\ &\le \mathbb{E}[\eta\delta_2\|\bar{\mathbf{x}}_t - \mathbf{x}_\star\|^2 + \frac{\eta}{\delta_2}\|\mathbf{u}_t^\top D_t(G_t - \mathbf{v}_t\mathbf{1}^\top G_t)\|^2] \\ &\le 2\eta\delta_2\,\mathbb{E}[\|\bar{\mathbf{x}}_t - \mathbf{u}_t^\top X_t\|^2] + 2\eta\delta_2\,\mathbb{E}[\|\mathbf{u}_t^\top X_t - \mathbf{x}_\star\|^2] \\ &\quad + \frac{\eta}{\delta_2}\,\mathbb{E}[\|G_t - \mathbf{v}_t\mathbf{1}^\top G_t\|^2] \\ &\le 2\eta\delta_2\rho_t + 2\eta\delta_2 d'_t + \frac{\eta}{\delta_2}(2\zeta'_t + (4M+4)\psi'_t). \end{aligned} \tag{19}$$

To bound the last term of (18), we use $\mathbf{1}^\top G_t = \sum_{i=1}^M \nabla f_i(\mathbf{x}_i^t)$. Following (7), we then get

$$\begin{aligned}
&\langle \bar{\mathbf{x}}_t - \mathbf{x}_\star, \mathbf{1}^\top G_t\rangle \\
&\geq \frac{M}{2}(f(\bar{\mathbf{x}}_t) - f(\mathbf{x}_\star)) + \frac{\mu M}{4}\|\bar{\mathbf{x}}_t - \mathbf{x}_\star\|^2 - \frac{L}{2}\|X_t - \mathbf{1}\bar{\mathbf{x}}_t^\top\|^2 \\
&\geq \frac{M}{2}(f(\bar{\mathbf{x}}_t) - f(\mathbf{x}_\star)) + \frac{\mu M}{8}\|\mathbf{u}_t^\top X_t - \mathbf{x}_\star\|^2 \\
&\quad - \frac{\mu M}{4}\|\bar{\mathbf{x}}_t - \mathbf{u}_t^\top X_t\|^2 - \frac{L}{2}\|X_t - \mathbf{1}\bar{\mathbf{x}}_t^\top\|^2.
\end{aligned}$$

Applying Lemma 9 and Lemma 7b gives

$$\begin{aligned}
\mathbb{E}[-2\eta\alpha_t\langle \bar{\mathbf{x}}_t - \mathbf{x}_\star, \mathbf{1}^\top G_t\rangle] &\leq -\eta\underline{\alpha} M e_t - \frac{\eta\underline{\alpha}\mu M}{4} d_t' \\
&\quad + \eta\left(L + \frac{\mu M}{2}\right)\rho_t.
\end{aligned}$$

We recall that $\alpha_t = \mathbf{v}_t D_t \mathbf{u}_t$. Putting all together and choosing $\delta_1 = 16ML/\underline{\alpha}$ and $\delta_2 = \underline{\alpha}\mu M/16$, we obtain

$$\begin{aligned}
d_{t+1}' &\leq \left(1 - \frac{\eta\underline{\alpha}\mu M}{4}\right) d_t' - \eta\underline{\alpha} M e_t \\
&\quad + \eta\left(L + \frac{\mu M}{2}\right)\rho_t + \frac{16\eta ML}{\underline{\alpha}}\rho_t \\
&\quad + \left(\frac{\eta\underline{\alpha}}{16ML} + \eta^2\right) \\
&\qquad \cdot (4ML^2\rho_t + (8M+8)\psi_t' + 4\zeta_t' + 8M^2 L e_t) \\
&\quad + \frac{\eta\underline{\alpha}\mu M}{8}\rho_t + \frac{\eta\underline{\alpha}\mu M}{8} d_t' + \frac{16\eta}{\underline{\alpha}\mu M}(2\zeta_t' + (4M+4)\psi_t').
\end{aligned}$$

The coefficient of e_t is $-\eta M\left(\frac{\underline{\alpha}}{2} - 8\eta ML\right)$. Since $\eta \leq \underline{\alpha}/(16ML)$, this is non-positive and we have indeed

$$d_{t+1}' \leq (1 - c_1\eta) d_t' + c_2\eta\rho_t + c_3\eta\psi_t' + c_4\eta\zeta_t'$$

for some positive constants $(c_k)_{1\leq k\leq 4}$.

Bounding ρ_{t+1}**.** Let $s \in \{1, \dots, t\}$. As the matrices $(A_t)_{t\in\mathbb{N}}$ are row-stochastic, it holds

$$\begin{aligned}
(I-J)A_{t:s+1}(I-J)A_s &= (I-J)A_{t:s+1}A_s \\
&= (I-J)A_{t:s+1}A_s(I-J).
\end{aligned}$$

Hence, for any $\delta > 0$, we can write

$$\begin{aligned}
&\|(I-J)A_{t:s+1}(I-J)X_{s+1}\|^2 \\
&= \|(I-J)A_{t:s+1}(I-J)A_s(X_s - \eta D_s Y_{s+\frac{1}{2}})\|^2 \\
&\leq (1+\delta)\|(I-J)A_{t:s}(I-J)X_s\|^2 \\
&\quad + \left(1 + \frac{1}{\delta}\right)\eta^2\|I-J\|^2\|A_{t:s}\|^2\|D_s Y_{s+\frac{1}{2}}\|^2 \\
&\leq (1+\delta)\|(I-J)A_{t:s}(I-J)X_s\|^2 + \left(1+\frac{1}{\delta}\right)\eta^2 M\|G_s\|^2.
\end{aligned} \tag{20}$$

In the last line we have used the fact that $A_{t:s}$ is row-stochastic so that $\|A_{t:s}\| \leq \sqrt{M}$ and the inequality $\|D_s Y_{s+\frac{1}{2}}\| \leq \|G_s\|$.

Since $X_{t+1} - \mathbf{1}\bar{\mathbf{x}}_{t+1}^\top = (I-J)A_{t:t+1}(I-J)X_{t+1}$, applying (20) repeatedly then gives

$$\begin{aligned}
\|X_{t+1} - \mathbf{1}\bar{\mathbf{x}}_{t+1}^\top\|^2 &\leq (1+\delta)^{l+1}\|(I-J)A_{t:t-l}(I-J)X_s\|^2 \\
&\quad + \eta^2 M\left(1+\frac{1}{\delta}\right)\sum_{s=0}^{l}(1+\delta)^s\|G_{t-s}\|^2.
\end{aligned}$$

Let $\delta = \frac{1}{l+1}\log\frac{1+\lambda}{2\lambda} > 0$ so that for all $0 \leq s \leq l+1$, we have $(1+\delta)^s \leq \frac{1+\lambda}{2\lambda} < \frac{1}{\lambda}$. Taking expectation in the above inequality and invoking Lemma 6a and Lemma 8b leads to

$$\rho_{t+1} \leq \frac{1+\lambda}{2}\rho_{t-l} + \frac{\eta^2 M}{\lambda}\left(1 + \frac{1}{\delta}\right)\sum_{s=0}^{l}\Delta_{t-s},$$

where $\Delta_{t-s} = 8M^2 L e_{t-s} + 4ML^2\rho_{t-s} + (8M+8)\psi_{t-s}' + 4\zeta_{t-s}'$. With (16) and $\eta \leq \underline{\alpha}/(16ML)$ we thus see there exist positive constants $(c_k)_{5\leq k\leq 8}$ such that

$$\rho_{t+1} \leq \frac{1+\lambda_1}{2}\rho_t + \eta\sum_{s=0}^{l}(c_5 d_{t-s}' + c_6\rho_{t-s} + c_7\psi_{t-s}' + c_8\zeta_{t-s}').$$

Bounding ψ_{t+1}'**.** By Young's inequality,

$$\begin{aligned}
\|\nabla f_i(\mathbf{z}_i^{t+1}) - \nabla f_i(\mathbf{x}_i^{t+1})\|^2 &\leq \frac{2-p_i}{2-2p_i}\|\nabla f_i(\mathbf{z}_i^{t+1}) - \nabla f_i(\mathbf{x}_i^t)\|^2 \\
&\quad + \frac{2-p_i}{p_i}\|\nabla f_i(\mathbf{x}_i^t) - \nabla f_i(\mathbf{x}_i^{t+1})\|^2.
\end{aligned}$$

The update rule of $\mathbf{z}_i^t$ implies that

$$\begin{aligned}
&\mathbb{E}_t[\|\nabla f_i(\mathbf{z}_i^{t+1}) - \nabla f_i(\mathbf{x}_i^t)\|^2] \\
&= (1-p_i)\|\nabla f_i(\mathbf{z}_i^t) - \nabla f_i(\mathbf{x}_i^t)\|^2 \\
&\quad + p_i\|\nabla f_i(\mathbf{x}_i^t) - \nabla f_i(\mathbf{x}_i^t)\|^2 \\
&= (1-p_i)\|\nabla f_i(\mathbf{z}_i^t) - \nabla f_i(\mathbf{x}_i^t)\|^2.
\end{aligned}$$

With $\underline{p} = \min_{i\in\mathcal{V}} p_i$ as defined in Lemma 9, we then have

$$\begin{aligned}
&\mathbb{E}_t[\|\nabla f_i(\mathbf{z}_i^{t+1}) - \nabla f_i(\mathbf{x}_i^{t+1})\|^2] \\
&\leq \left(1 - \frac{p_i}{2}\right)\|\nabla f_i(\mathbf{z}_i^t) - \nabla f_i(\mathbf{x}_i^t)\|^2 \\
&\quad + \frac{2}{p_i}\,\mathbb{E}_t[\|\nabla f_i(\mathbf{x}_i^t) - \nabla f_i(\mathbf{x}_i^{t+1})\|^2] \\
&\leq \left(1 - \frac{\underline{p}}{2}\right)\|\nabla f_i(\mathbf{z}_i^t) - \nabla f_i(\mathbf{x}_i^t)\|^2 \\
&\quad + \frac{2}{\underline{p}}\,\mathbb{E}_t[\|\nabla f_i(\mathbf{x}_i^t) - \nabla f_i(\mathbf{x}_i^{t+1})\|^2].
\end{aligned}$$

Taking total expectation and summing from $i = 1$ to M gives

$$\psi_{t+1}' \leq \left(1 - \frac{\underline{p}}{2}\right)\psi_t' + \frac{2}{\underline{p}}\,\mathbb{E}[\|\nabla F(X_t) - \nabla F(X_{t+1})\|^2].$$

By Lipschitz-continuity of the gradients, it holds $\|\nabla F(X_t) - \nabla F(X_{t+1})\| \leq L\|X_t - X_{t+1}\|$. We then develop

$$\begin{aligned}
X_{t+1} - X_t &= A_t(X_t - \eta D_t Y_{t+\frac{1}{2}}) - X_t \\
&= (A_t - I)(I-J)X_t - \eta A_t D_t Y_{t+\frac{1}{2}}.
\end{aligned}$$

With $\|A_t - I\|^2 \leq 2\|A_t\|^2 + 2\|I\|^2 \leq 2M + 2$, we obtain that

$$\begin{aligned}
&\|\nabla F(X_t) - \nabla F(X_{t+1})\|^2 \\
&\leq L^2((4M+4)\|X_t - \mathbf{1}\bar{\mathbf{x}}_t^\top\|^2 + 2\eta^2 M\|G_t\|^2).
\end{aligned}$$

Combining the above and applying Lemma 8b leads to

$$\begin{aligned}
\psi_{t+1}' &\leq \left(1 - \frac{\underline{p}}{2}\right)\psi_t' + \frac{2L^2}{\underline{p}}((4M+4)\rho_t \\
&\quad + 2\eta^2 M(4ML^2\rho_t + (8M+8)\psi_t' + 4\zeta_t' + 8M^2 L e_t)).
\end{aligned}$$

Using (16) and $\eta \leq \underline{\alpha}/(16ML)$ we deduce the existence of positive constants $(c_k)_{9\leq k\leq 12}$ such that

$$\psi_{t+1}' \leq c_9\eta d_t' + c_{10}\rho_t + \left(1 - \frac{\underline{p}}{2} + c_{11}\eta\right) + c_{12}\eta\zeta_t'.$$

Bounding ζ'_{t+1}. Let $s \in \{1,\dots,t\}$. Using the column-stochasticity of B_s and the definition $\mathbf{v}_{s+1} = B_s\mathbf{v}_s$, we get

$$(I - \mathbf{v}_{s+1}\mathbf{1}^\top)B_s = B_s - \mathbf{v}_{s+1}\mathbf{1}^\top = B_s - B_s\mathbf{v}_s\mathbf{1}^\top.$$

Hence, for any $\delta > 0$, it holds that

$$\begin{aligned}
&\|B_{t:s+1}(I - \mathbf{v}_{s+1}\mathbf{1}^\top)Y_{s+1}\|^2 \\
&= \|B_{t:s+1}(I - \mathbf{v}_{s+1}\mathbf{1}^\top)B_s(Y_s + D_s(\nabla F(X_s) - \nabla F(Z_s)))\|^2 \\
&\le (1+\delta)\|B_{t:s}(I - \mathbf{v}_s\mathbf{1}^\top)Y_s\|^2 \\
&\quad + \left(1 + \frac{1}{\delta}\right)\|B_{t:s} - B_{t:s}\mathbf{v}_s\mathbf{1}^\top\|^2\|D_s\|^2\|\nabla F(X_s) - \nabla F(Z_s)\|^2 \\
&\le (1+\delta)\|B_{t:s}(I - \mathbf{v}_s\mathbf{1}^\top)Y_s\|^2 \\
&\quad + 4M\left(1 + \frac{1}{\delta}\right)\|\nabla F(X_s) - \nabla F(Z_s)\|^2. \qquad (21)
\end{aligned}$$

In the last inequality we have used

$$\|B_{t:s} - B_{t:s}\mathbf{v}_s\mathbf{1}^\top\| \le \|B_{t:s}\| + \|B_{t:s}\mathbf{v}_s\mathbf{1}^\top\| \le 2\sqrt{M}$$

which is true because both $B_{t:s}$ and $B_{t:s}\mathbf{v}_s\mathbf{1}^\top$ are column-stochastic.

Since $Y_{t+1} - \mathbf{v}_{t+1}\mathbf{1}^\top Y_{t+1} = B_{t:t+1}(I - \mathbf{v}_{t+1}\mathbf{1}^\top)Y_{t+1}$, applying (21) repeatedly then gives

$$\begin{aligned}
&\|Y_{t+1} - \mathbf{v}_{t+1}\mathbf{1}^\top Y_{t+1}\|^2 \\
&\le (1+\delta)^{l+1}\|B_{t:t-l}(I - \mathbf{v}_{t-l}\mathbf{1}^\top)Y_{t-l}\|^2 \\
&\quad + 4M\left(1 + \frac{1}{\delta}\right)\sum_{s=0}^{l}(1+\delta)^s\|\nabla F(X_{t-s}) - \nabla F(Z_{t-s})\|^2.
\end{aligned}$$

Let us take $\delta = \frac{1}{l+1}\log\frac{1+\lambda}{2\lambda} > 0$ as before. Taking total expectation in the above inequality and invoking Lemma 6b leads to

$$\zeta'_{t+1} \le \frac{1+\lambda_2}{2}\zeta'_{t-l} + \frac{4M}{\lambda}\left(1 + \frac{1}{\delta}\right)\sum_{s=0}^{l}\psi'_{t-s}.$$

We set $c_{13} = \frac{4M}{\lambda}\left(1 + \frac{1}{\delta}\right)$.

Conclude. Putting all together we get exactly (15). □

3) Geometric convergence of PPDS: From Proposition 2, we are now in position to prove the geometric convergence of PPDS by showing that the spectral radius of $Q_0 + \eta Q_e$ is smaller than 1 for $\eta > 0$ sufficiently small.

Proof of Theorem 1. In the following we analyze the eigenvalues of $Q_0 + \eta Q_e$ with help of matrix perturbation theory. We first notice that Q_0 is a block-triangular matrix. Its characteristic polynomial can be easily computed and is given by

$$P_{Q_0}(\nu) = \nu^{2l}(\nu - 1)\left(\nu - \left(1 - \frac{p}{2}\right)\right)\left(\nu^{l+1} - \frac{1+\lambda}{2}\right)^2.$$

This shows that the spectral radius of Q_0 is 1 and 1 is also the unique eigenvalue of largest modulus of the matrix.

Let us denote by $\theta_1 = 1, \theta_2, \dots, \theta_{4(l+1)}$ the eigenvalues of Q_0 so that $|\theta_k| < 1$ for all $k \in \{2, ..., 4(l+1)\}$. By continuity of the eigenvalues, for any $\varepsilon > 0$ there exists $\delta > 0$ such that if $\eta < \delta$, for any θ_k of multiplicity m the matrix $Q_0 + \eta Q_e$ has exactly m eigenvalues (counting multiplicity) in $\mathbb{B}(\theta_k, \varepsilon)$, the open disk centered at θ_k with radius ε; see [39, Chap. 5.1]. Let us take ε small enough such that all the eigenvalues of $Q_0 + \eta Q_e$ are smaller than $1 - \varepsilon$ in modulus except the greatest one. For $\eta < \delta$, we can then define $\theta_1(\eta)$ as the unique eigenvalue of $Q_0 + \eta Q_e$ that is in $\mathbb{B}(1, \varepsilon)$. We will now show that $|\theta_1(\eta)| < 1$ for η sufficiently small. For this, let

$$\mathbf{u} = [1\ 0\ \dots\ 0]^\top, \quad \mathbf{v} = [\underbrace{1\ \dots\ 1}_{l+1\text{ times}}\ 0\ \dots\ 0]^\top$$

be respectively the left and the right eigenvector of Q_0 associated with the eigenvalue 1. By [40, Th. 6.3.12] (see also [41, Th. 1]), we have

$$\theta'_1(0) = \frac{\mathbf{u}^\top Q_e \mathbf{v}}{\mathbf{u}^\top \mathbf{v}} = -c_1 < 0.$$

As a consequence, $|\theta_1(\eta)| < 1$ for η sufficiently small and subsequently $\rho(Q_0 + \eta Q_e) < 1$.

In order to conclude, we need to get rid of the dependence on T which plays a role in the definition of the vectors $(\mathbf{u}_t)_{1\le t\le T}$ and the quantities $(d'_t)_{1\le t\le T}$. We recall that $d_t = \mathbb{E}[\|\bar{\mathbf{x}}_t - \mathbf{x}_\star\|^2]$. As in (19), we have both $d_t \le 2d'_t + 2\rho_t$ and $d'_t \le 2d_t + 2\rho_t$. Let us define $\mathbf{r}''_t$ by replacing $(d'_s)_{t\le s\le t+l}$ by $(d_s)_{t\le s\le t+l}$ in $\mathbf{r}'_t$. The above inequalities can then be translated into $\mathbf{r}''_t \le W\mathbf{r}'_t$ and $\mathbf{r}'_t \le W\mathbf{r}''_t$ for a non-negative matrix W properly defined. Note that neither W nor $Q_0 + \eta Q_e$ depend on t or T. Therefore, the inequality

$$\mathbf{r}''_t \le W(Q_0 + \eta Q_e)^{t-1}W\mathbf{r}''_1$$

which holds for all $t \in \mathbb{N}$ guarantees the geometric convergence of $\mathbf{r}''_t$ and subsequently of all the relevant quantities when η is small enough.

Finally, from the geometric convergence of $\mathbb{E}[\|\mathbf{x}_i^t - \mathbf{x}_\star\|^2]$, we deduce that $\mathbf{x}_i^t$ converges to $\mathbf{x}_\star$ almost surely by using Markov's inequality and the Borel–Cantelli lemma. □

V. Simulations

In this section, we illustrate the interest of PPDS for asynchronous decentralized optimization on a) a synthetic ridge regression problem; and b) a logistic regression problem on a real dataset. The code to reproduce the experiments can be found at https://github.com/yassine-laguel/ppds.

TABLE I
Datasets and graph description

	Synthetic	EMNIST
Number of features d	10	784
Number of examples	10000	2500
Devices M	100	50
Local size n_{local}	100	50
RGG radius r	0.2	0.3
Sampled nodes / round	20	10
Sampled neighbors / communication	1	1

A. Dataset, tasks and models

For both problems, we minimize an objective of the form

$$f(\mathbf{x}) = \frac{1}{M}\sum_{i=1}^{M}\underbrace{\left(\sum_{j=1}^{n_{\text{local}}} f_{i,j}(\mathbf{x}) + \lambda\|\mathbf{x}\|_2^2\right)}_{f_i(\mathbf{x})}$$

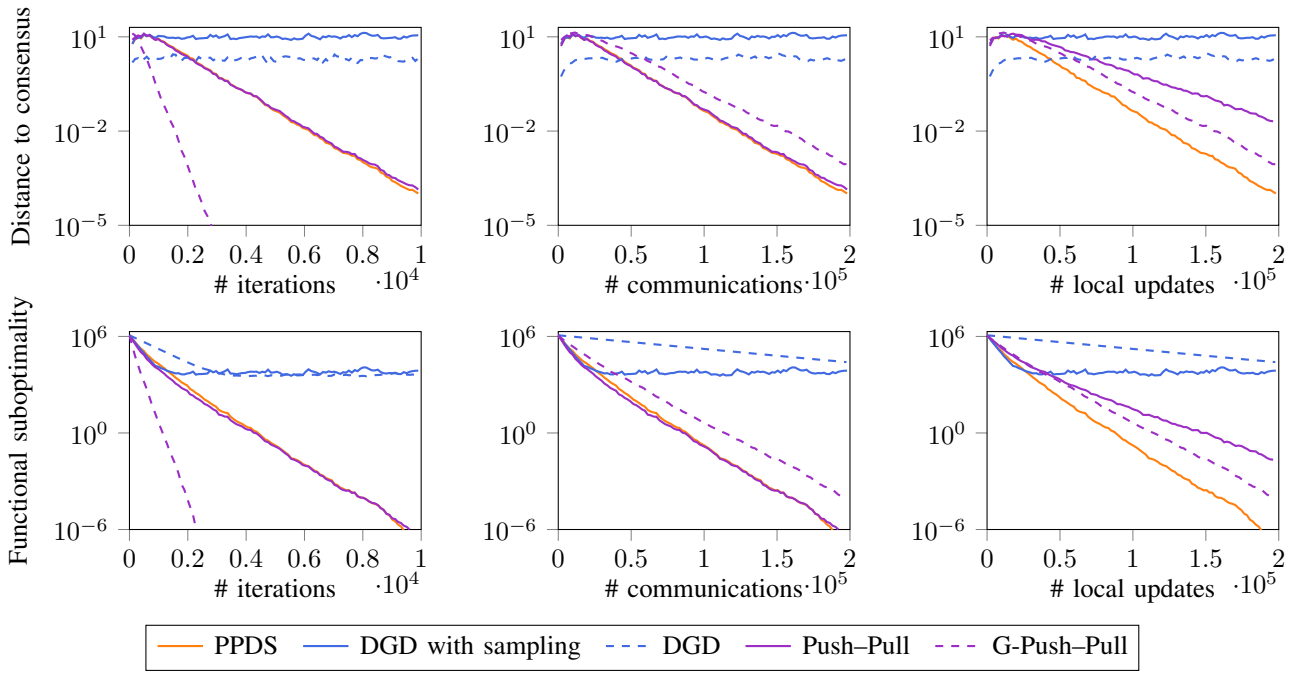


Fig. 1. Numerical illustrations for Ridge regression on synthetic dataset.

that is, each worker i has a local dataset of n_{local} examples and a Tikhonov regularization term with parameter $\lambda = 1/n_{\text{local}}$. Since the local objectives are convex, this regularization make the problem strongly convex.

We form the communication networks by generating Random Geometric Graphs (RGG) using the library `networkx` [42] with different number of nodes M and radius r for each experiment. We consider the *broadcast* setting illustrated in the third example of Section II-C. Precisely, all activated nodes broadcast their models to one (randomly chosen) neighbor during a communication step. We illustrate the effect of device sampling by comparing algorithms with *full-device participation* and with *random device sampling* (20 nodes for the synthetic dataset and 10 nodes for EMNIST). The relevant parameters are reported in Table I.

a) Ridge regression on synthetic dataset: For this problem, the local losses are defined as

$$f_{i,j}(\mathbf{x}) = (b_{i,j} - \mathbf{x}^\top a_{i,j})^2$$

where $(a_{i,j}, b_{i,j}) \in \mathbb{R}^d \times \mathbb{R}$ are data points generated using the procedure `make_regression` from `scikit-learn` [43]. Different seeds are used for different nodes, yielding statistically heterogeneous distributions between the nodes.

b) Logistic regression on EMNIST: The EMNIST dataset [44] is comprised of images of handwritten digits and letters from several authors. We consider the problem of finding which character is written from its image. For this, we consider local losses of the form

$$f_{i,j}(\mathbf{x}) = -b_{i,j} \log\left(\text{softmax}(\mathbf{x}^\top a_{i,j})\right)$$

where the $(a_{i,j}, b_{i,j})$ are respectively $d = 28 \times 28$ gray-scale images of handwritten digits character and their associated one-hot label $b_{i,j} \in (0-9, a-z, A-Z)$ totaling 62 classes. Each worker's local dataset comes from images from the same author.

B. Algorithms, hyperparameters and evaluation metrics

We compare the proposed algorithm PPDS with several baselines: Decentralized Gradient Descent (DGD) with and without sampling, Push–Pull and G-Push–Pull. The same broadcast communication scheme is applied to all the methods, and the same uniform sampling strategy is adopted whenever device sampling is involved.

For each method, the stepsize is taken fixed, tuned with a coarse-to-fine strategy: we first select the stepsize η_1 within $\{10^{-k}, -2 \leq k \leq 5\}$ yielding the best global training loss; then, a second search is performed over $\{\eta_1 2^k, -2 \leq k \leq 2\}$.

We report two evaluation metrics: (*i*) the distance to consensus $(1/M)\sum_{i=1}^M \|\mathbf{x}_i^t - \bar{\mathbf{x}}_t\|^2$; and (*ii*) the functional suboptimality $(1/M)\sum_{i=1}^M f(\mathbf{x}_i^t) - f_\star$. For the synthetic dataset, the optimal solution is computed by inversion of a linear system. For the real dataset, it is set as the best solution in terms of final training losses found by the implemented algorithms.

For each experiment, we report these metrics in terms of three different measures: (*i*) number of iterations; (*ii*) communication cost, i.e., the cumulative number of activated communication links; and (*iii*) number of local updates. These cover different aspects that influence the efficiency of distributed optimization.

C. Numerical results

First, we observe on Figs. 1 and 2 that the proposed method PPDS converges linearly, as expected from Theorem 1. Furthermore, looking at the right-hand plots, we see that PPDS outperforms all the other methods when it comes to measuring

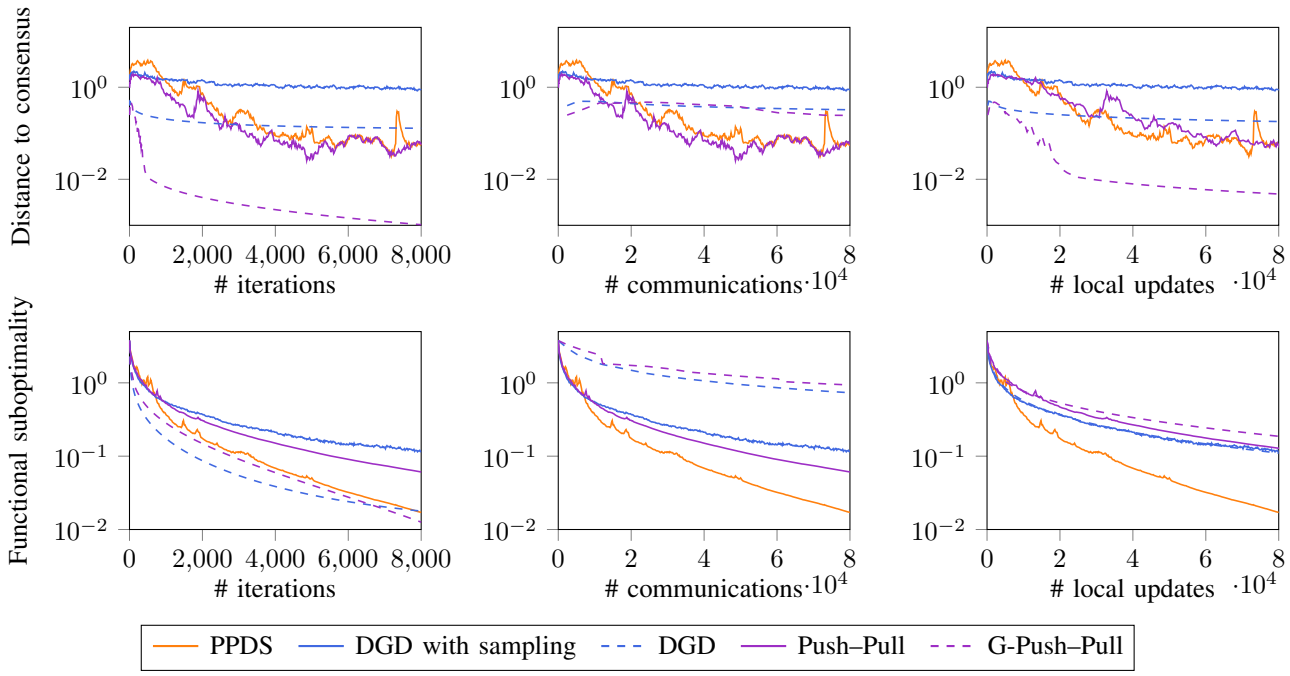


Fig. 2. Numerical illustrations for logistic regression on EMNIST.

the functional optimality with respect to the number of local updates. This illustrates that PPDS indeed saves computational resources, by an efficient interplay between computation and communications. Concerning the communication complexity (middle plots), PPDS is at least as competitive as G-Push–Pull, which was shown in [15] to beat other baselines. Finally, we observe that Push–Pull, as a synchronous gradient-tracking method, naturally achieves the best performance when measuring in terms of the number of iterations (left-hand plots), but tends to be less efficient when we consider the actual communication and computational costs.

VI. Conclusions

In this paper, we showed how device sampling can be incorporated in asynchronous decentralized gradient descent, by extending the Push–Pull method. We proved linear convergence of the method on strongly convex functions and validate our approach on problems with synthetic and real data. This work also opens towards several research directions, including the combination of device sampling with local gradients computation or with gradient compression techniques.

Appendix

A. Proof of Lemma 4

Proof. a) Since f_i is L-smooth, it holds for all $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^d$ that

$$\|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{x}')\|^2 \leq 2L(f_i(\mathbf{x}) - f_i(\mathbf{x}') - \langle \mathbf{x} - \mathbf{x}', \nabla f_i(\mathbf{x}')\rangle).$$

Subsequently,

$$\begin{aligned}
&\sum_{i=1}^M \|\nabla f_i(\mathbf{x}_i^t) - \nabla f_i(\mathbf{x}_\star)\|^2 \\
&\leq 2\sum_{i=1}^M \|\nabla f_i(\mathbf{x}_i^t) - \nabla f_i(\bar{\mathbf{x}}_t)\|^2 + 2\sum_{i=1}^M \|\nabla f_i(\bar{\mathbf{x}}_t) - \nabla f_i(\mathbf{x}_\star)\|^2 \\
&\leq 2L^2 \sum_{i=1}^M \|\mathbf{x}_i^t - \bar{\mathbf{x}}_t\|^2 \\
&\quad + 4L\sum_{i=1}^M (f_i(\bar{\mathbf{x}}_t) - f_i(\mathbf{x}_\star) - \langle \bar{\mathbf{x}}_t - \mathbf{x}_\star, \nabla f_i(\mathbf{x}_\star)\rangle) \\
&= 2L^2 \|X_t - \mathbf{1}\bar{\mathbf{x}}_t^\top\|^2 \\
&\quad + 4ML(f(\bar{\mathbf{x}}_t) - f(\mathbf{x}_\star) - \langle \bar{\mathbf{x}}_t - \mathbf{x}_\star, \nabla f(\mathbf{x}_\star)\rangle) \\
&= 2L^2 \|X_t - \mathbf{1}\bar{\mathbf{x}}_t^\top\|^2 + 4ML(f(\bar{\mathbf{x}}_t) - f(\mathbf{x}_\star)).
\end{aligned}$$

In the last line we used that $\nabla f(\mathbf{x}_\star) = 0$. Taking expectation gives the desired inequality.

b) The inequality is straightforward from *a)* and the following decomposition

$$\begin{aligned}
&\|\nabla F(X_t) - \nabla F(Z_t)\|^2 \\
&\leq 2\|\nabla F(X_t) - \nabla F(\mathbf{1}^\top \mathbf{x}_\star)\|^2 + 2\|\nabla F(\mathbf{1}^\top \mathbf{x}_\star) - \nabla F(Z_t)\|^2.
\end{aligned}$$

c) We first decompose

$$\|G_t\|^2 = \|(I-J)G_t + JG_t\|^2 = \|(I-J)G_t\|^2 + \|JG_t\|^2.$$

For the first term we simply use the definition of G_t to obtain

$$\begin{aligned}
&\|(I-J)G_t\|^2 \\
&\leq 2\|(I-J)Y\|^2 + 2\|(I-J)(\nabla F(X_t) - \nabla F(Z_t))\|^2 \\
&\leq 2\|Y_t - \mathbf{1}\bar{\mathbf{y}}_t^\top\|^2 + 2\|\nabla F(X_t) - \nabla F(Z_t)\|^2.
\end{aligned}$$

From Lemma 1b, we know that $\sum_{i=1}^M \boldsymbol{g}_i^t = \sum_{i=1}^M \nabla f_i(\mathbf{x}_i^t)$ or equivalently $JG_t = J\nabla F(X_t)$. Thus for the second term we use again $\nabla f(\mathbf{x}_\star) = 0$ to get

$$\begin{aligned}\|JG_t\|^2 &= M\Big\|\frac{1}{M}\sum_{i=1}^M \nabla f_i(\mathbf{x}_i^t))\Big\|^2 \\ &= M\Big\|\frac{1}{M}\sum_{i=1}^M \nabla f_i(\mathbf{x}_i^t) - \frac{1}{M}\sum_{i=1}^M \nabla f_i(\mathbf{x}_\star)\Big\|^2 \\ &\leq \sum_{i=1}^M \|\nabla f_i(\mathbf{x}_i^t) - \nabla f_i(\mathbf{x}_\star)\|^2.\end{aligned}$$

Combining the above, taking expectation, and using *a)* and *b)* gives the desired result. □

B. Proof of Eq. (11)

Proof. Let Q_k denotes the k-th column of Q and $\mathbf{e}_k$ denote the k-th canonical vector of $\mathbb{R}^4$. First, $\omega^\top Q_1 = 1 - \frac{\eta\mu S}{2M} = (1 - \frac{\eta\mu S}{2M})\ \omega^\top \mathbf{e}_1$.
Since $\eta \leq \frac{(1-\lambda)^2}{14L}\sqrt{\frac{M}{S}}$, it holds

$$\begin{aligned}\omega^\top Q_2 &= \frac{\sqrt{S}(1-\lambda)}{M^{\frac{3}{2}}}\left(\frac{\eta L}{1-\lambda}\sqrt{\frac{S}{M}} + \frac{10\eta^2L^2}{1-\lambda}\left(\frac{S}{M}\right)^{\frac{3}{2}}\right. \\ &\qquad \left. + \frac{1+\lambda}{2} + \frac{20\eta^2L^2S}{M(1-\lambda)} + \frac{\eta L}{4(1-\lambda)}\sqrt{\frac{S}{M}}\right) \\ &\leq \frac{\sqrt{S}(1-\lambda)}{M^{\frac{3}{2}}}\left(\frac{5\eta L}{4(1-\lambda)}\sqrt{\frac{S}{M}} + \frac{30\eta^2L^2S}{M(1-\lambda)} + \frac{1+\lambda}{2}\right) \\ &\leq \frac{\sqrt{S}(1-\lambda)}{M^{\frac{3}{2}}}\frac{3+\lambda}{4} = \frac{3+\lambda}{4}\ \omega^\top \mathbf{e}_2.\end{aligned}$$

With $\eta \leq \frac{(1-\lambda)^2}{2304L}\left(\frac{M}{S}\right)^{\frac{3}{2}}$, we have

$$\begin{aligned}\omega^\top Q_3 &= \frac{\eta(1-\lambda)}{96ML}\left(\frac{192\eta L}{1-\lambda}\left(\frac{S}{M}\right)^2\right. \\ &\qquad \left. + \frac{384\eta L}{1-\lambda}\left(\frac{S}{M}\right)^{\frac{3}{2}} + \frac{1+\lambda}{2}\right) \\ &\leq \frac{\eta(1-\lambda)}{96ML}\left(\frac{576\eta L}{1-\lambda}\left(\frac{S}{M}\right)^{\frac{3}{2}} + \frac{1+\lambda}{2}\right) \\ &\leq \frac{\eta(1-\lambda)}{96ML}\frac{3+\lambda}{4} = \frac{3+\lambda}{4}\ \omega^\top \mathbf{e}_3.\end{aligned}$$

Similarly, using $\eta \leq \frac{1}{576L}\sqrt{\frac{M}{S}}$, we get

$$\begin{aligned}\omega^\top Q_4 &= \frac{\eta}{12ML}\left(48\eta L\left(\frac{S}{M}\right)^2 + 96\eta L\left(\frac{S}{M}\right)^{\frac{3}{2}}\right. \\ &\qquad \left. + \frac{S}{2M} + 1 - \frac{S}{M}\right) \\ &\leq \frac{\eta}{12ML}\left(1 - \frac{S}{2M} + 144\eta L\left(\frac{S}{M}\right)^{\frac{3}{2}}\right) \\ &\leq \frac{\eta}{12ML}\left(1 - \frac{S}{4M}\right) = \left(1 - \frac{S}{4M}\right)\ \omega^\top \mathbf{e}_4.\end{aligned}$$

As for e_t, we note that $\eta \leq \frac{(1-\lambda)^2}{2304L}\left(\frac{M}{S}\right)^{\frac{3}{2}} < \frac{1}{120L}\left(\frac{M}{S}\right)^{\frac{3}{2}}$ and thus

$$\begin{aligned}\omega^\top \mathbf{h} &= -\frac{\eta S}{M} + \frac{20\eta^2LS^2}{M^2} + 40\eta^2L\left(\frac{S}{M}\right)^{\frac{3}{2}} + \frac{\eta S}{6M} + \frac{\eta S}{3M} \\ &\leq -\frac{\eta S}{2M} + 60\eta^2L\left(\frac{S}{M}\right)^{\frac{3}{2}} \leq 0.\end{aligned}$$

As $\eta \leq \frac{(1-\lambda)^2}{14L}\sqrt{\frac{M}{S}}$ implies that $1 - \frac{\eta\mu S}{2M} \geq \frac{3+\lambda}{4}$, we have

$$\omega^\top Q \leq \gamma\ \omega^\top I$$

where the inequality is elementwise and $\gamma = \max\left(1 - \frac{\eta\mu S}{2M}, 1 - \frac{S}{4M}\right)$. Since all the involved terms are non-negative, combining with the above inequalities gives

$$\omega^\top \mathbf{r}_{t+1} = \omega^\top Q\ \mathbf{r}_t + e_t\ \omega^\top \mathbf{h} \leq \gamma\ \omega^\top \mathbf{r}_t$$

which concludes the proof. □

C. Proof of Lemma 8

Proof. *a)* From the definition of $\boldsymbol{g}_t$ we can write

$$\begin{aligned}&\|G_t - \mathbf{v}_t\mathbf{1}^\top G_t\|^2 \\ &= \|Y_t - \mathbf{v}_t\mathbf{1}^\top Y_t + (I - \mathbf{v}_t\mathbf{1}^\top)(\nabla F(X_t) - \nabla F(Z_t))\|^2 \\ &\leq 2\|Y_t - \mathbf{v}_t\mathbf{1}^\top Y_t\|^2 + 2\|I - \mathbf{v}_t\mathbf{1}^\top\|^2\|\nabla F(X_t) - \nabla F(Z_t)\|^2.\end{aligned}$$

We conclude by using

$$\|I - \mathbf{v}_t\mathbf{1}^\top\|^2 \leq 2\|I\|^2 + 2\|\mathbf{v}_t\mathbf{1}^\top\|^2 \leq 2M + 2.$$

and taking expectation over the above inequalities.

b) By Young's inequality,

$$\|G_t\|^2 \leq 2\|G_t - \mathbf{v}_t\mathbf{1}^\top G_t\|^2 + 2\|\mathbf{v}_t\mathbf{1}^\top G_t\|^2.$$

Using $\mathbf{1}^\top G_t = \mathbf{1}^\top \nabla F(X)$, $\nabla f(\mathbf{x}_\star) = 0$, and the fact that $\mathbf{v}_t$ is a probability vector, we deduce that

$$\begin{aligned}\|\mathbf{v}_t\mathbf{1}^\top G_t\|^2 &= \sum_{i=1}^M \left\|\mathbf{v}_i^t \sum_{j=1}^M \nabla f_j(\mathbf{x}_j^t)\right\|^2 \\ &\leq \left\|\sum_{j=1}^M \nabla f_j(\mathbf{x}_j^t)\right\|^2 \\ &= \left\|\sum_{i=1}^M \nabla f_i(\mathbf{x}_i^t) - \sum_{i=1}^M \nabla f_i(\mathbf{x}_\star)\right\|^2 \\ &\leq M\sum_{i=1}^M \left\|\nabla f_i(\mathbf{x}_i^t) - \nabla f_i(\mathbf{x}_\star)\right\|^2.\end{aligned}$$

Combining the above two inequalities with *a)* and Lemma 4a we get the desired result. □

D. Proof of Lemma 9

Proof. Using the independence assumption, we can write

$$\mathbb{E}[\alpha_t\chi_t] = \mathbb{E}[\mathbf{u}_{t+1}^\top A_t D_t \mathbf{v}_t \chi_t] = \mathbb{E}[\mathbf{u}_{t+1}^\top]\,\mathbb{E}[A_t D_t]\,\mathbb{E}[\mathbf{v}_t\chi_t].$$

From Assumption 3c we deduce that

$$\mathbb{E}[A_t D_t] \geq \mathbb{E}[\mathrm{diag}(\nu\mathbf{1})D_t] = \nu\,\mathrm{diag}(\mathbf{p}),$$

where $\mathbf{p} = (p_i)_{i\in\mathcal{V}}$. We have shown in (12) that $\mathbb{E}[\mathbf{u}_{t+1}] = \pi_A$. Notice that all the elements of π_A are positive according to the Perron-Frobenius theorem. Thus, $\underline{\pi_A} > 0$ and we have

$$\mathbb{E}[\mathbf{u}_{t+1}^\top]\,\mathbb{E}[A_t D_t] \geq \pi_A^\top \nu\,\mathrm{diag}(\mathbf{p}) \geq \underline{\alpha}\mathbf{1}^\top$$

where Assumption 4 ensures that $\underline{p} > 0$ and thus $\underline{\alpha} > 0$.

On the other hand, $\mathbf{u}_t$ being a probability vector we can always upper bound $\mathbf{u}_t^\top D_t$ by $\mathbf{1}^\top$. Using the non-negativity of $\mathbf{v}_t$ and χ_t, we then obtain

$$\underline{\alpha}\,\mathbb{E}[\mathbf{1}^\top \mathbf{v}_t\chi_t] \leq \mathbb{E}[\alpha_t\chi_t] \leq \mathbb{E}[\mathbf{1}^\top \mathbf{v}_t\chi_t].$$

This is exactly (14) since $\mathbf{v}_t$ is a probability vector. □

ACKNOWLEDGMENT

This research was partially supported by MIAI@Grenoble Alpes (ANR-19-P3IA-0003).

REFERENCES

[1] J. Tsitsiklis, D. Bertsekas, and M. Athans, "Distributed asynchronous deterministic and stochastic gradient optimization algorithms," *IEEE transactions on automatic control*, vol. 31, no. 9, pp. 803–812, 1986.

[2] D. Bertsekas and J. Tsitsiklis, *Parallel and distributed computation: numerical methods*. Athena Scientific, 2015.

[3] A. Nedić, A. Olshevsky, and M. G. Rabbat, "Network topology and communication-computation tradeoffs in decentralized optimization," *Proceedings of the IEEE*, vol. 106, no. 5, pp. 953–976, 2018.

[4] T. Yang, X. Yi, J. Wu, Y. Yuan, D. Wu, Z. Meng, Y. Hong, H. Wang, Z. Lin, and K. H. Johansson, "A survey of distributed optimization," *Annual Reviews in Control*, vol. 47, pp. 278–305, 2019.

[5] M. Assran, A. Aytekin, H. R. Feyzmahdavian, M. Johansson, and M. G. Rabbat, "Advances in asynchronous parallel and distributed optimization," *Proceedings of the IEEE*, vol. 108, no. 11, pp. 2013–2031, 2020.

[6] A. Nedic, "Distributed gradient methods for convex machine learning problems in networks: Distributed optimization," *IEEE Signal Processing Magazine*, vol. 37, no. 3, pp. 92–101, 2020.

[7] J. C. Duchi, A. Agarwal, and M. J. Wainwright, "Dual averaging for distributed optimization: Convergence analysis and network scaling," *IEEE Transactions on Automatic control*, vol. 57, no. 3, pp. 592–606, 2011.

[8] S. Boyd, N. Parikh, and E. Chu, *Distributed optimization and statistical learning via the alternating direction method of multipliers*. Now Publishers Inc, 2011.

[9] A. Nedic and A. Ozdaglar, "Distributed subgradient methods for multi-agent optimization," *IEEE Transactions on Automatic Control*, vol. 54, no. 1, pp. 48–61, 2009.

[10] D. Kovalev, A. Salim, and P. Richtárik, "Optimal and practical algorithms for smooth and strongly convex decentralized optimization," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[11] J. M. Hendrickx and J. N. Tsitsiklis, "Fundamental limitations for anonymous distributed systems with broadcast communications," in *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, 2015, pp. 9–16.

[12] D. Kempe, A. Dobra, and J. Gehrke, "Gossip-based computation of aggregate information," in *44th Annual IEEE Symposium on Foundations of Computer Science, 2003. Proceedings.* IEEE, 2003, pp. 482–491.

[13] F. Iutzeler, P. Ciblat, and W. Hachem, "Analysis of sum-weight-like algorithms for averaging in wireless sensor networks," *IEEE Transactions on Signal Processing*, vol. 61, no. 11, pp. 2802–2814, 2013.

[14] A. Nedić and A. Olshevsky, "Stochastic gradient-push for strongly convex functions on time-varying directed graphs," *IEEE Transactions on Automatic Control*, vol. 61, no. 12, pp. 3936–3947, 2016.

[15] S. Pu, W. Shi, J. Xu, and A. Nedić, "Push–pull gradient methods for distributed optimization in networks," *IEEE Transactions on Automatic Control*, vol. 66, no. 1, pp. 1–16, 2020.

[16] R. Xin and U. A. Khan, "A linear algorithm for optimization over directed graphs with geometric convergence," *IEEE Control Systems Letters*, vol. 2, no. 3, pp. 315–320, 2018.

[17] S. Boyd, A. Ghosh, B. Prabhakar, and D. Shah, "Randomized gossip algorithms," *IEEE transactions on information theory*, vol. 52, no. 6, pp. 2508–2530, 2006.

[18] F. Iutzeler, P. Bianchi, P. Ciblat, and W. Hachem, "Asynchronous distributed optimization using a randomized alternating direction method of multipliers," in *52nd IEEE conference on decision and control*. IEEE, 2013, pp. 3671–3676.

[19] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics*. PMLR, 2017, pp. 1273–1282.

[20] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings *et al.*, "Advances and open problems in federated learning," *arXiv preprint arXiv:1912.04977*, 2019.

[21] A. Defazio, F. Bach, and S. Lacoste-Julien, "Saga: A fast incremental gradient method with support for non-strongly convex composite objectives," in *Advances in neural information processing systems*, 2014, pp. 1646–1654.

[22] J. Xu, S. Zhu, Y. C. Soh, and L. Xie, "Augmented distributed gradient methods for multi-agent optimization under uncoordinated constant stepsizes," in *2015 54th IEEE Conference on Decision and Control (CDC)*. IEEE, 2015, pp. 2055–2060.

[23] A. Nedic, A. Olshevsky, and W. Shi, "Achieving geometric convergence for distributed optimization over time-varying graphs," *SIAM Journal on Optimization*, vol. 27, no. 4, pp. 2597–2633, 2017.

[24] G. Qu and N. Li, "Harnessing smoothness to accelerate distributed optimization," *IEEE Transactions on Control of Network Systems*, vol. 5, no. 3, pp. 1245–1260, 2017.

[25] W. Shi, Q. Ling, G. Wu, and W. Yin, "Extra: An exact first-order algorithm for decentralized consensus optimization," *SIAM Journal on Optimization*, vol. 25, no. 2, pp. 944–966, 2015.

[26] H. Li and Z. Lin, "Revisiting extra for smooth distributed optimization," *SIAM Journal on Optimization*, vol. 30, no. 3, pp. 1795–1821, 2020.

[27] R. Xin, U. A. Khan, and S. Kar, "Variance-reduced decentralized stochastic optimization with accelerated convergence," *IEEE Transactions on Signal Processing*, vol. 68, pp. 6255–6271, 2020.

[28] H. Hendrikx, F. Bach, and L. Massoulie, "An optimal algorithm for decentralized finite sum optimization," *SIAM Journal on Optimization*, 2021.

[29] R. Xin and U. A. Khan, "A linear algorithm for optimization over directed graphs with geometric convergence," *IEEE Control Systems Letters*, vol. 2, no. 3, pp. 315–320, 2018.

[30] J. Zhang and K. You, "Fully asynchronous distributed optimization with linear convergence in directed networks," *arXiv preprint arXiv:1901.08215*, 2019.

[31] F. Saadatniaki, R. Xin, and U. A. Khan, "Decentralized optimization over time-varying directed graphs with row and column-stochastic matrices," *IEEE Transactions on Automatic Control*, vol. 65, no. 11, pp. 4769–4780, 2020.

[32] R. A. Horn and C. R. Johnson, *Matrix analysis*. Cambridge university press, 2012.

[33] S. Pu, W. Shi, J. Xu, and A. Nedić, "A push-pull gradient method for distributed optimization in networks," in *2018 IEEE Conference on Decision and Control (CDC)*. IEEE, 2018, pp. 3385–3390.

[34] R. Johnson and T. Zhang, "Accelerating stochastic gradient descent using predictive variance reduction," *Advances in neural information processing systems*, vol. 26, pp. 315–323, 2013.

[35] A. Kulunchakov and J. Mairal, "Estimate sequences for stochastic composite optimization: Variance reduction, acceleration, and robustness to noise," *arXiv preprint arXiv:1901.08788*, 2019.

[36] F. Iutzeler, P. Ciblat, and W. Hachem, "Analysis of sum-weight-like algorithms for averaging in wireless sensor networks," *IEEE Transactions on Signal Processing*, vol. 61, no. 11, pp. 2802–2814, 2013.

[37] A. Kolmogoroff, "Zur theorie der markoffschen ketten," *Mathematische Annalen*, vol. 112, no. 1, pp. 155–160, 1936.

[38] B. Touri, *Product of random stochastic matrices and distributed averaging*. Springer Science & Business Media, 2012.

[39] T. Kato, *Perturbation theory for linear operators*. Springer Science & Business Media, 2013, vol. 132.

[40] C. R. Johnson and R. A. Horn, *Matrix analysis, 2nd ed*. Cambridge university press, 2013.

[41] A. Greenbaum, R.-c. Li, and M. L. Overton, "First-order perturbation theory for eigenvalues and eigenvectors," *SIAM review*, vol. 62, no. 2, pp. 463–482, 2020.

[42] A. Hagberg, P. Swart, and D. S Chult, "Exploring network structure, dynamics, and function using networkx," Los Alamos National Lab.(LANL), Los Alamos, NM (United States), Tech. Rep., 2008.

[43] L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. VanderPlas, A. Joly, B. Holt, and G. Varoquaux, "API design for machine learning software: experiences from the scikit-learn project," in *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, 2013, pp. 108–122.

[44] G. Cohen, S. Afshar, J. Tapson, and A. Van Schaik, "Emnist: Extending mnist to handwritten letters," in *2017 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2017, pp. 2921–2926.