

Revisiting Randomized Gossip Algorithms: General Framework, Convergence Rates and Novel Block and Accelerated Protocols

Nicolas Loizou
University of Edinburgh
n.loizou@sms.ed.ac.uk

Peter Richtárik
KAUST , MIPT
peter.richtarik@kaust.edu.sa

May 20, 2019*

Abstract

In this work we present a new framework for the analysis and design of randomized gossip algorithms for solving the average consensus problem. We show how classical randomized iterative methods for solving linear systems can be interpreted as gossip algorithms when applied to special systems encoding the underlying network and explain in detail their decentralized nature. Our general framework recovers a comprehensive array of well-known gossip algorithms as special cases, including the pairwise randomized gossip algorithm and path averaging gossip, and allows for the development of provably faster variants. The flexibility of the new approach enables the design of a number of new specific gossip methods. For instance, we propose and analyze novel *block* and the first provably *accelerated* randomized gossip protocols, and *dual* randomized gossip algorithms.

From a numerical analysis viewpoint, our work is the first that explores in depth the decentralized nature of randomized iterative methods for linear systems and proposes them as methods for solving the average consensus problem.

We evaluate the performance of the proposed gossip protocols by performing extensive experimental testing on typical wireless network topologies.

Keywords randomized gossip algorithms · average consensus · weighted average consensus · stochastic methods · linear systems · randomized Kaczmarz · randomized block Kaczmarz · randomized coordinate descent · heavy ball momentum · Nesterov’s acceleration · duality · convex optimization · wireless sensor networks

Mathematical Subject Classifications 93A14 · 68W15 · 68Q25 · 68W20 · 68W40 · 65Y20 · 90C15 · 90C20 · 90C25 · 15A06 · 15B52 · 65F10

*Part of the results was presented in [46, 49, 45]

Contents

1	Introduction	3
1.1	Main contributions	3
1.2	Structure of the paper	4
1.3	Notation	4
2	Background - Technical Preliminaries	5
2.1	Randomized iterative methods for linear systems	5
2.2	Other related work	8
3	Sketch and Project Methods as Gossip Algorithms	9
3.1	Weighted average consensus	9
3.2	Gossip algorithms through sketch and project framework	10
3.2.1	Standard form and mass preservation	10
3.2.2	ε -Averaging time	11
3.3	Randomized Kaczmarz method as gossip algorithm	11
3.3.1	AC system with incidence matrix \mathbf{Q}	12
3.3.2	AC system with Laplacian matrix \mathbf{L}	12
3.3.3	Details on complexity results	13
3.4	Block gossip algorithms	15
4	Faster and Provably Accelerated Randomized Gossip Algorithms	17
4.1	Gossip algorithms with heavy ball momentum	17
4.1.1	Sketch and project with heavy ball momentum	17
4.1.2	Randomized Kaczmarz gossip with heavy ball momentum	19
4.1.3	Connections with existing fast randomized gossip algorithms	20
4.1.4	Randomized block Kaczmarz gossip with heavy ball momentum	21
4.1.5	Mass preservation	22
4.2	Provably accelerated randomized gossip algorithms	23
4.2.1	Accelerated Kaczmarz methods using Nesterov's momentum	24
4.2.2	Theoretical guarantees of AccRK	24
4.2.3	Accelerated randomized gossip algorithms	27
5	Dual Randomized Gossip Algorithms	27
5.1	Dual problem and SDSA	28
5.2	Randomized Newton method as a dual gossip algorithm	29
6	Further Connections Between Methods for Solving Linear Systems and Gossip Algorithms	30
7	Numerical Evaluation	32
7.1	Convergence on weighted average consensus	33
7.2	Benefit of block variants	33
7.3	Accelerated gossip algorithms	34
7.3.1	Impact of momentum parameter on mRK	34
7.3.2	Comparison of mRK and shift-Register algorithm [40]	35
7.3.3	Impact of momentum parameter on mRBK	35
7.3.4	Performance of AccGossip	35
7.4	Relaxed randomized gossip without momentum	36
8	Conclusion and Future Direction of Research	37
9	Acknowledgments	37
A	Missing Proofs	42
A.1	Proof of Theorem 4	42
A.2	Proof of Theorem 5	42
B	Notation Glossary	44

1 Introduction

Average consensus is a fundamental problem in distributed computing and multi-agent systems. It comes up in many real world applications such as coordination of autonomous agents, estimation, rumour spreading in social networks, PageRank and distributed data fusion on ad-hoc networks and decentralized optimization. Due to its great importance there is much classical [81, 13] and recent [88, 87, 6] work on the design of efficient algorithms/protocols for solving it.

In the average consensus (AC) problem we are given an undirected connected network $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with node set $\mathcal{V} = \{1, 2, \dots, n\}$ and edges \mathcal{E} . Each node $i \in \mathcal{V}$ “knows” a private value $c_i \in \mathbb{R}$. The goal of AC is for every node to compute the average of these private values, $\bar{c} := \frac{1}{n} \sum_i c_i$, in a distributed fashion. That is, the exchange of information can only occur between connected nodes (neighbors).

One of the most attractive classes of protocols for solving the average consensus problem are gossip algorithms. The development and design of gossip algorithms was studied extensively in the last decade. The seminal 2006 paper of Boyd et al. [6] motivated a flurry of subsequent research and gossip algorithms now appear in many applications, including distributed data fusion in sensor networks [88], load balancing [10] and clock synchronization [19]. For a survey of selected relevant work prior to 2010, we refer the reader to the work of Dimakis et al. [14]. For more recent results on randomized gossip algorithms we suggest [94, 40, 63, 43, 55, 3]. See also [15, 4, 64, 29, 30].

1.1 Main contributions

In this work, we connect two areas of research which until now have remained remarkably disjoint in the literature: randomized iterative (projection) methods for solving linear systems and randomized gossip protocols for solving the average consensus. This connection enables us to make contributions by borrowing from each body of literature to the other and using it we propose a new framework for the design and analysis of novel efficient randomized gossip protocols.

The main contributions of our work include:

- **RandNLA.** We show how classical randomized iterative methods for solving linear systems can be interpreted as gossip algorithms when applied to special systems encoding the underlying network and explain in detail their decentralized nature. Through our general framework we recover a comprehensive array of well-known gossip protocols as special cases. In addition our approach allows for the development of novel block and dual variants of all of these methods. From a randomized numerical linear algebra (RandNLA) viewpoint our work is the first that explores in depth, the decentralized nature of randomized iterative methods for solving linear systems and proposes them as efficient methods for solving the average consensus problem (and its weighted variant).
- **Weighted AC.** The methods presented in this work solve the more general *weighted* average consensus (Weighted AC) problem (Section 3.1) popular in the area of distributed cooperative spectrum sensing networks. The proposed protocols are the first randomized gossip algorithms that directly solve this problem with finite-time convergence rate analysis. In particular, we prove linear convergence of the proposed protocols and explain how we can obtain further acceleration using momentum. To the best of our knowledge, the existing decentralized protocols that solve the weighted average consensus problem show convergence but without convergence analysis.
- **Acceleration.** We present novel and provably *accelerated* randomized gossip protocols. In each step, of the proposed algorithms, all nodes of the network update their values using their own information but only a subset of them exchange messages. The protocols are inspired by the recently proposed accelerated variants of randomized Kaczmarz-type methods and use momentum terms on top of the sketch and project update rule (gossip communication) to obtain better theoretical and practical performance. To the best of our knowledge, our accelerated protocols are the first randomized gossip algorithms that converge to a consensus with

a provably accelerated linear rate without making any further assumptions on the structure of the network. Achieving an accelerated linear rate in this setting using randomized gossip protocols was an open problem.

- **Duality.** We reveal a hidden duality of randomized gossip algorithms, with the dual iterative process maintaining variables attached to the edges of the network. We show how the randomized coordinate descent and randomized Newton methods work as edge-based dual randomized gossip algorithms.
- **Experiments.** We corroborate our theoretical results with extensive experimental testing on typical wireless network topologies. We numerically verify the linear convergence of the our protocols for solving the weighted AC problem. We explain the benefit of using block variants in the gossip protocols where more than two nodes update their values in each iteration. We explore the performance of the proposed provably accelerated gossip protocols and show that they significantly outperform the standard pairwise gossip algorithm and existing fast pairwise gossip protocols with momentum. An experiment showing the importance of over-relaxation in the gossip setting is also presented.

This paper contains a synthesis and a unified presentation of the randomized gossip protocols proposed in Loizou and Richtárik [46], Loizou and Richtárik [49] and Loizou et al. [45]. In [46], building upon the results from [26], a connection between the area of randomized iterative methods for linear systems and gossip algorithms was established and block gossip algorithms were developed. Then in [49] and [45] faster and provably accelerated gossip algorithms were proposed using the heavy ball momentum and Nesterov’s acceleration technique, respectively. This paper expands upon these results and presents proofs for theorems that are referenced in the above papers. We also conduct several new experiments.

We believe that this work could potentially open up new avenues of research in the area of decentralized gossip protocols.

1.2 Structure of the paper

This work is organized as follows. Section 2 introduces the necessary background on basic randomized iterative methods for linear systems that will be used for the development of randomized gossip protocols. Related work on the literature of linear system solvers, randomized gossip algorithms for averaging and gossip algorithms for consensus optimization is presented. In Section 3 the more general weighted average consensus problem is described and the connections between the two areas of research (randomized projection methods for linear systems and gossip algorithms) is established. In particular we explain how methods for solving linear systems can be interpreted as gossip algorithms when applied to special systems encoding the underlying network and elaborate in detail their distributed nature. Novel block gossip variants are also presented. In Section 4 we describe and analyze fast and provably accelerated randomized gossip algorithms. In each step of these protocols all nodes of the network update their values but only a subset of them exchange their private values. Section 5 describes dual randomized gossip algorithms that operate with values that are associated to the edges of the network and Section 6 highlights further connections between methods for solving linear systems and gossip algorithms. Numerical evaluation of the new gossip protocols is presented in Section 7. Finally, concluding remarks are given in Section 8.

1.3 Notation

For convenience, a table of the most frequently used notation is included in the Appendix B. In particular, with boldface upper-case letters denote matrices; \mathbf{I} is the identity matrix. By $\|\cdot\|$ and $\|\cdot\|_F$ we denote the Euclidean norm and the Frobenius norm, respectively. For a positive integer number n , we write $[n] := \{1, 2, \dots, n\}$. By \mathcal{L} we denote the solution set of the linear system $\mathbf{A}x = b$, where $\mathbf{A} \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$.

We shall often refer to specific matrix expressions involving several matrices. In order to keep these expressions brief throughout the paper it will be useful to define the following two matrices:

$$\mathbf{H} := \mathbf{S}(\mathbf{S}^\top \mathbf{A} \mathbf{B}^{-1} \mathbf{A}^\top \mathbf{S})^\dagger \mathbf{S}^\top \quad \text{and} \quad \mathbf{Z} := \mathbf{A}^\top \mathbf{H} \mathbf{A}, \quad (1)$$

depending on a random matrix $\mathbf{S} \in \mathbb{R}^{m \times q}$ drawn from a given distribution \mathcal{D} and on an $n \times n$ positive definite matrix \mathbf{B} which defines the geometry of the space. In particular we define \mathbf{B} -inner product in \mathbb{R}^n via $\langle x, z \rangle_{\mathbf{B}} := \langle \mathbf{B}x, z \rangle = x^\top \mathbf{B}z$ and an induced \mathbf{B} -norm, $\|x\|_{\mathbf{B}} := (x^\top \mathbf{B}x)^{1/2}$. By $\mathbf{A}_{i\cdot}$ and $\mathbf{A}_{\cdot j}$ we denote the i^{th} row and the j^{th} column of matrix \mathbf{A} , respectively. By \dagger we denote the Moore-Penrose pseudoinverse.

The complexity of all gossip protocols presented in this paper is described by the spectrum of matrix

$$\mathbf{W} = \mathbf{B}^{-1/2} \mathbf{A}^\top \mathbb{E}[\mathbf{H}] \mathbf{A} \mathbf{B}^{-1/2} \stackrel{(1)}{=} \mathbf{B}^{-1/2} \mathbb{E}[\mathbf{Z}] \mathbf{B}^{-1/2}, \quad (2)$$

where the expectation is taken over $\mathbf{S} \sim \mathcal{D}$. With λ_{\min}^+ and λ_{\max} we indicate the smallest nonzero and the largest eigenvalue of matrix \mathbf{W} , respectively.

Vector $x^k = (x_1^k, \dots, x_n^k) \in \mathbb{R}^n$ represents the vector with the private values of the n nodes of the network at the k^{th} iteration while with x_i^k we denote the value of node $i \in [n]$ at the k^{th} iteration. $\mathcal{N}_i \subseteq \mathcal{V}$ denotes the set of nodes that are neighbors of node $i \in \mathcal{V}$. By $\alpha(\mathcal{G})$ we denote the algebraic connectivity of graph \mathcal{G} .

Throughout the paper, x^* is the projection of x^0 onto \mathcal{L} in the \mathbf{B} -norm. We write $x^* = \Pi_{\mathcal{L}, \mathbf{B}}(x^0)$. An explicit formula for the projection of x onto set \mathcal{L} is given by

$$\Pi_{\mathcal{L}, \mathbf{B}}(x) := \arg \min_{x' \in \mathcal{L}} \|x' - x\|_{\mathbf{B}} = x - \mathbf{B}^{-1} \mathbf{A}^\top (\mathbf{A} \mathbf{B}^{-1} \mathbf{A}^\top)^\dagger (\mathbf{A}x - b).$$

Finally, with $\mathbf{Q} \in \mathbb{R}^{|\mathcal{E}| \times n}$ we define the incidence matrix and with $\mathbf{L} \in \mathbb{R}^{n \times n}$ the Laplacian matrix of the network. Note that it holds that $\mathbf{L} = \mathbf{Q}^\top \mathbf{Q}$. Further, with \mathbf{D} we denote the degree matrix of the graph. That is, $\mathbf{D} = \text{Diag}(d_1, d_2, \dots, d_n) \in \mathbb{R}^{n \times n}$ where d_i is the degree of node $i \in \mathcal{V}$.

2 Background - Technical Preliminaries

Solving linear systems is a central problem in numerical linear algebra and plays an important role in computer science, control theory, scientific computing, optimization, computer vision, machine learning, and many other fields. With the advent of the age of big data, practitioners are looking for ways to solve linear systems of unprecedented sizes. In this large scale setting, randomized iterative methods are preferred mainly because of their cheap per iteration cost and because they can easily scale to extreme dimensions.

2.1 Randomized iterative methods for linear systems

Kaczmarz-type methods are very popular for solving linear systems $\mathbf{A}x = b$ with many equations. The (deterministic) Kaczmarz method for solving consistent linear systems was originally introduced by Kaczmarz in 1937 [34]. Despite the fact that a large volume of papers was written on the topic, the first provably linearly convergent variant of the Kaczmarz method—the randomized Kaczmarz Method (RK)—was developed more than 70 years later, by Strohmer and Vershynin [77]. This result sparked renewed interest in design of randomized methods for solving linear systems [56, 57, 18, 51, 93, 58, 74, 41]. More recently, Gower and Richtárik [25] provide a unified analysis for several randomized iterative methods for solving linear systems using a sketch-and-project framework. We adopt this framework in this paper.

In particular, the analysis in [25] was done under the assumption that matrix \mathbf{A} has full column rank. This assumption was lifted in [26], and a duality theory for the method developed. Later, in [73], it was shown that the sketch and project method of [25] can be interpreted as stochastic gradient descent applied to a suitable stochastic optimization problem and relaxed variants of the proposed methods have been presented.

The sketch-and-project algorithm [25, 73] for solving a consistent linear system $\mathbf{A}x = b$ has the form

$$\begin{aligned} x^{k+1} &= x^k - \omega \mathbf{B}^{-1} \mathbf{A}^\top \mathbf{S}_k (\mathbf{S}_k^\top \mathbf{A} \mathbf{B}^{-1} \mathbf{A}^\top \mathbf{S}_k)^\dagger \mathbf{S}_k^\top (\mathbf{A}x^k - b) \\ &\stackrel{(1)}{=} x^k - \omega \mathbf{B}^{-1} \mathbf{A}^\top \mathbf{H}_k (\mathbf{A}x^k - b), \end{aligned} \quad (3)$$

where in each iteration, matrix $\mathbf{S}_k \in \mathbb{R}^{m \times q}$ is sampled afresh from an arbitrary distribution \mathcal{D} .¹ In [25] it was shown that many popular randomized algorithms for solving linear systems, including RK method and randomized coordinate descent method (a.k.a Gauss-Seidel method) can be cast as special cases of the above update by choosing an appropriate combination of the distribution \mathcal{D} and the positive definite matrix \mathbf{B} .

In the special case that $\omega = 1$ (no relaxation), the update rule of equation (3) can be equivalently written as follows:

$$\begin{aligned} x^{k+1} &:= \operatorname{argmin}_{x \in \mathbb{R}^n} \|x - x^k\|_{\mathbf{B}}^2 \\ \text{subject to} &\quad \mathbf{S}_k^\top \mathbf{A}x = \mathbf{S}_k^\top b. \end{aligned} \quad (4)$$

This equivalent presentation of the method justifies the name *Sketch and Project*. In particular, the method is a two step procedure: (i) Draw random matrix $\mathbf{S}_k \in \mathbb{R}^{m \times q}$ from distribution \mathcal{D} and formulate a *sketched* system $\mathbf{S}_k^\top \mathbf{A}x = \mathbf{S}_k^\top b$, (ii) *Project* the last iterate x^k into the solution set of the sketched system.

A formal presentation of the Sketch and Project method is shown in Algorithm 1.

Algorithm 1 Sketch and Project Method [73]

- 1: **Parameters:** Distribution \mathcal{D} from which method samples matrices; stepsize/relaxation parameter $\omega \in \mathbb{R}$; momentum parameter β .
 - 2: **Initialize:** $x^0, x^1 \in \mathbb{R}^n$
 - 3: **for** $k = 0, 1, 2, \dots$ **do**
 - 4: Draw a fresh $\mathbf{S}_k \sim \mathcal{D}$
 - 5: Set $x^{k+1} = x^k - \omega \mathbf{B}^{-1} \mathbf{A}^\top \mathbf{S}_k (\mathbf{S}_k^\top \mathbf{A} \mathbf{B}^{-1} \mathbf{A}^\top \mathbf{S}_k)^\dagger \mathbf{S}_k^\top (\mathbf{A}x^k - b)$
 - 6: **end for**
 - 7: **Output:** The last iterate x^k
-

In this work, we are mostly interested in two special cases of the sketch and project framework—the randomized Kaczmarz (RK) method and its block variant, the randomized block Kaczmarz (RBK) method. In addition, in the following sections we present novel scaled and accelerated variants of these two selected cases and interpret their gossip nature. In particular, we focus on explaining how these methods can solve the average consensus problem and its more general version, the weighted average consensus (subsection 3.1).

Let $e_i \in \mathbb{R}^m$ be the i^{th} unit coordinate vector in \mathbb{R}^m and let $\mathbf{I}_{:C}$ be column submatrix of the $m \times m$ identity matrix with columns indexed by $C \subseteq [m]$. Then RK and RBK methods can be obtained as special cases of the update rule (3) as follows:

- RK: Let $\mathbf{B} = \mathbf{I}$ and $\mathbf{S}_k = e_i$, where $i \in [m]$ is chosen independently at each iteration, with probability $p_i > 0$. In this setup the update rule (3) simplifies to

$$x^{k+1} = x^k - \omega \frac{\mathbf{A}_{i:} x^k - b_i}{\|\mathbf{A}_{i:}\|^2} \mathbf{A}_{i:}^\top. \quad (5)$$

- RBK: Let $\mathbf{B} = \mathbf{I}$ and $\mathbf{S} = \mathbf{I}_{:C}$, where set $C \subseteq [m]$ is chosen independently at each iteration, with probability $p_C \geq 0$. In this setup the update rule (3) simplifies to

$$x^{k+1} = x^k - \omega \mathbf{A}_{C:}^\top (\mathbf{A}_{C:} \mathbf{A}_{C:}^\top)^\dagger (\mathbf{A}_{C:} x^k - b_C). \quad (6)$$

¹We stress that there is no restriction on the number of columns of matrix \mathbf{S}_k (q can be varied)

In several papers [26, 73, 48, 50], it was shown that even in the case of consistent linear systems with *multiple* solutions, Kaczmarz-type methods converge linearly to one particular solution: the projection (on \mathbf{B} -norm) of the initial iterate x^0 onto the solution set of the linear system. This naturally leads to the formulation of the *best approximation problem*:

$$\min_{x=(x_1, \dots, x_n) \in \mathbb{R}^n} \frac{1}{2} \|x - x^0\|_{\mathbf{B}}^2 \quad \text{subject to} \quad \mathbf{A}x = b. \quad (7)$$

where $\mathbf{A} \in \mathbb{R}^{m \times n}$. In the rest of this manuscript, x^* denotes the solution of (7) and we write $x^* = \Pi_{\mathcal{L}, \mathbf{B}}(x^0)$.

Exactness. An important assumption that is required for the convergence analysis of the randomized iterative methods under study is *exactness*. That is:

$$\text{Null}(\mathbb{E}_{\mathbf{S} \sim \mathcal{D}}[\mathbf{Z}]) = \text{Null}(\mathbf{A}). \quad (8)$$

The exactness property is of key importance in the sketch and project framework, and should be seen as an assumption on the distribution \mathcal{D} and not on matrix \mathbf{A} .

Clearly, an assumption on the distribution \mathcal{D} of the random matrices \mathbf{S} should be required for the convergence of Algorithm 1. For an instance, if \mathcal{D} is such that, $\mathbf{S} = e_1$ with probability 1, where $e_1 \in \mathbb{R}^m$ be the 1st unit coordinate vector in \mathbb{R}^m , then the algorithm will select the same row of matrix \mathbf{A} in each step. For this choice of distribution it is clear that the algorithm will not converge to a solution of the linear system. The exactness assumption guarantees that this will not happen.

For necessary and sufficient conditions for exactness, we refer the reader to [73]. Here it suffices to remark that the exactness condition is very weak, allowing \mathcal{D} to be virtually any reasonable distribution of random matrices. For instance, a sufficient condition for exactness is for the matrix $\mathbb{E}[\mathbf{H}]$ to be positive definite [26]. This is indeed a weak condition since it is easy to see that this matrix is symmetric and positive semidefinite without the need to invoke any assumptions; simply by design.

A much stronger condition than exactness is $\mathbb{E}[\mathbf{Z}] \succ 0$ which has been used for the analysis of the sketch and project method in [25]. In this case, the matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ of the linear system requires to have full column rank and as a result the consistent linear system has a unique solution.

The convergence performance of the Sketch and Project method (Algorithm 1) under the exactness assumption for solving the best approximation problem is described by the following theorem.

Theorem 1 ([73]). *Let assume exactness and let $\{x^k\}_{k=0}^\infty$ be the iterates produced by the sketch and project method (Algorithm 1) with step-size $\omega \in (0, 2)$. Set, $x^* = \Pi_{\mathcal{L}, \mathbf{B}}(x^0)$. Then,*

$$\mathbb{E}[\|x^k - x^*\|_{\mathbf{B}}^2] \leq \rho^k \|x^0 - x^*\|_{\mathbf{B}}^2, \quad (9)$$

where

$$\rho := 1 - \omega(2 - \omega)\lambda_{\min}^+ \in [0, 1]. \quad (10)$$

Recall that λ_{\min}^+ denotes the minimum nonzero eigenvalue of matrix $\mathbf{W} := \mathbb{E}[\mathbf{B}^{-1/2} \mathbf{A}^\top \mathbf{H} \mathbf{A} \mathbf{B}^{-1/2}]$.

In other words, using standard arguments, from Theorem 1 we observe that for a given $\epsilon \in (0, 1)$ we have that:

$$k \geq \frac{1}{1 - \rho} \log \left(\frac{1}{\epsilon} \right) \quad \Rightarrow \quad \mathbb{E}[\|x^k - x^*\|_{\mathbf{B}}^2] \leq \epsilon \|x^0 - x^*\|_{\mathbf{B}}^2.$$

We say that the iteration complexity of sketch and project method is,

$$O \left(\frac{1}{1 - \rho} \log \left(\frac{1}{\epsilon} \right) \right).$$

2.2 Other related work

On Sketch and Project Methods. Variants of the sketch-and-project methods have been recently proposed for solving several other problems. [22] and [27] use similar ideas for the development of linearly convergent randomized iterative methods for computing/estimating the inverse and pseudoinverse of a large matrix, respectively. A limited memory variant of the stochastic block BFGS method for solving the empirical risk minimization problem arising in machine learning was proposed by [23]. Tu et al. [82] utilize the sketch-and-project framework to show that breaking block locality can accelerate block Gauss-Seidel methods. In addition, they develop an accelerated variant of the method for a specific distribution \mathcal{D} . A sketch and project method with the heavy ball momentum was studied in [48, 47] and an accelerated (in the sense of Nesterov) variant of the method proposed in [24] for the more general Euclidean setting and applied to matrix inversion and quasi-Newton updates. Inexact variants of Algorithm 1 have been proposed in [50]. As we have already mentioned, in [73], through the development of stochastic reformulations, a stochastic gradient descent interpretation of the sketch and project method has been proposed. Recently, using a different stochastic reformulation, [21] performed a tight convergence analysis of stochastic gradient descent in a more general convex setting. The analysis proposed in [21] recovers the linear convergence rate of sketch and project method (Theorem 1) as special case.

Gossip algorithms for average consensus The problem of average consensus has been extensively studied in the automatic control and signal processing literature for the past two decades [14], and was first introduced for decentralized processing in the seminal work [81]. A clear connection between the rate of convergence and spectral characteristics of the underlying network topology over which message passing occurs was first established in [6] for pairwise randomized gossip algorithms.

Motivated by network topologies with salient properties of wireless networks (e.g., nodes can communicate directly only with other nearby nodes), several methods were proposed to accelerate the convergence of gossip algorithms. For instance, [5] proposed averaging among a set of nodes forming a path in the network (this protocol can be seen as special case of our block variants in Section 3.4). Broadcast gossip algorithms have also been analyzed [4] where the nodes communicate with more than one of their neighbors by broadcasting their values.

While the gossip algorithms studied in [6, 5, 4] are all first-order (the update of x^{k+1} only depends on x^k), a faster randomized pairwise gossip protocol was proposed in [7] which suggested to incorporate additional memory to accelerate convergence. The first analysis of this protocol was later proposed in [40] under strong conditions. It is worth to mention that in the setting of deterministic gossip algorithms theoretical guarantees for accelerated convergence were obtained in [65, 35]. In Section 4 we propose fast and provably accelerated randomized gossip algorithms with memory and compare them in more detail with the fast randomized algorithm proposed in [7, 40].

Gossip algorithms for multiagent consensus optimization. In the past decade there has been substantial interest in consensus-based multiagent optimization methods that use gossip updates in their update rule [55, 90, 76]. In multiagent consensus optimization setting, n agents or nodes, cooperate to solve an optimization problem. In particular, a local objective function $f_i : \mathbb{R}^d \rightarrow \mathbb{R}$ is associated with each node $i \in [n]$ and the goal is for all nodes to solve the optimization problem

$$\min_{x \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^n f_i(x) \quad (11)$$

by communicate only with their neighbors. In this setting gossip algorithms works in two steps by first executing some local computation followed by communication over the network [55]. Note that the average consensus problem with c_i as node i initial value can be case as a special case of the optimization problem (11) when the function values are $f_i(x) = (x - c_i)^2$.

Recently there has been an increasing interest in applying multitagent optimization methods to solve convex and non-convex optimization problems arising in machine learning [80, 39, 1, 2, 9, 36, 32]. In this setting most consensus-based optimization methods make use of standard, first-order

gossip, such as those described in [6], and incorporating momentum into their updates to improve their practical performance.

3 Sketch and Project Methods as Gossip Algorithms

In this section we show how by carefully choosing the linear system in the constraints of the best approximation problem (7) and the combination of the parameters of the Sketch and Project method (Algorithm 1) we can design efficient randomized gossip algorithms. We show that the proposed protocols can actually solve the weighted average consensus problem, a more general version of the average consensus problem described in Section 1. In particular we focus, on a scaled variant of the RK method (5) and on the RBK (6) and understand the convergence rates of these methods in the consensus setting, their distributed nature and how they are connected with existing gossip protocols.

3.1 Weighted average consensus

In the *weighted average consensus* (Weighted AC) problem we are given an undirected connected network $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with node set $\mathcal{V} = \{1, 2, \dots, n\}$ and edges \mathcal{E} . Each node $i \in \mathcal{V}$ holds a private value $c_i \in \mathbb{R}$ and its weight w_i . The goal of this problem is for every node to compute the weighted average of the private values,

$$\bar{c} := \frac{\sum_{i=1}^n w_i c_i}{\sum_{i=1}^n w_i},$$

in a distributed fashion. That is, the exchange of information can only occur between connected nodes (neighbors).

Note that in the special case when the weights of all nodes are the same ($w_i = r$ for all $i \in [n]$) the weighted average consensus is reduced to the standard average consensus problem. However, there are more special cases that could be interesting. For instance the weights can represent the degree of the nodes ($w_i = d_i$) or they can denote a probability vector and satisfy $\sum_i w_i = 1$ with $w_i > 0$.

It can be easily shown that the weighted average consensus problem can be expressed as optimization problem as follows:

$$\min_{x=(x_1, \dots, x_n) \in \mathbb{R}^n} \frac{1}{2} \|x - c\|_{\mathbf{B}}^2 \quad \text{subject to} \quad x_1 = x_2 = \dots = x_n \quad (12)$$

where matrix $\mathbf{B} = \text{Diag}(w_1, w_2, \dots, w_n)$ is a diagonal positive definite matrix (that is $w_i > 0$ for all $i \in [n]$) and $c = (c_1, \dots, c_n)^\top$ the vector with the initial values c_i of all nodes $i \in \mathcal{V}$. The optimal solution of this problem is $x_i^* = \frac{\sum_{i=1}^n w_i c_i}{\sum_{i=1}^n w_i}$ for all $i \in [n]$ which is exactly the solution of the weighted average consensus.

As we have explained, the standard average consensus problem can be cast as a special case of weighted average consensus. However, in the situation when the nodes have access to global information related to the network, such as the size of the network (number of nodes $n = |\mathcal{V}|$) and the sum of the weights $\sum_{i=1}^n w_i$, then any algorithm that solves the standard average consensus can be used to solve the weighted average consensus problem with the initial private values of the nodes changed from c_i to $\frac{n w_i c_i}{\sum_{i=1}^n w_i}$.

The weighted AC problem is popular in the area of distributed cooperative spectrum sensing networks [33, 66, 91, 92]. In this setting, one of the goals is to develop decentralized protocols for solving the cooperative sensing problem in cognitive radio systems. The weights in this case represent a ratio related to the channel conditions of each node/agent [33]. The development of methods for solving the weighted AC problem is an active area of research (check [33] for a recent comparison of existing algorithms). However, to the best of our knowledge, existing analysis for the proposed algorithms focuses on showing convergence and not on providing convergence rates. Our framework allows us to obtain novel randomized gossip algorithms for solving the weighted AC problem. In addition, we provide a tight analysis of their convergence rates. In particular, we show

convergence with a linear rate. See Section 7.1 for an experiment confirming linear convergence of one of our proposed protocols on typical wireless network topologies.

3.2 Gossip algorithms through sketch and project framework

We propose that randomized gossip algorithms should be viewed as special case of the Sketch and Project update to a particular problem of the form (7). In particular, we let $c = (c_1, \dots, c_n)$ be the initial values stored at the nodes of \mathcal{G} , and choose \mathbf{A} and b so that the constraint $\mathbf{A}x = b$ is equivalent to the requirement that $x_i = x_j$ (the value stored at node i is equal to the value stored at node j) for all $(i, j) \in \mathcal{E}$.

Definition 2. We say that $\mathbf{A}x = b$ is an “average consensus (AC) system” when $\mathbf{A}x = b$ iff $x_i = x_j$ for all $(i, j) \in \mathcal{E}$.

It is easy to see that $\mathbf{A}x = b$ is an AC system precisely when $b = 0$ and the nullspace of \mathbf{A} is $\{t\mathbf{1}_n : t \in \mathbb{R}\}$, where $\mathbf{1}_n$ is the vector of all ones in \mathbb{R}^n . Hence, \mathbf{A} has rank $n - 1$. Moreover in the case that $x^0 = c$, it is easy to see that for any AC system, the solution of (7) necessarily is $x^* = \bar{c} \cdot \mathbf{1}_n$ — this is why we singled out AC systems. In this sense, *any* algorithm for solving (7) will “find” the (weighted) average \bar{c} . However, in order to obtain a distributed algorithm we need to make sure that only “local” (with respect to \mathcal{G}) exchange of information is allowed.

It can be shown that many linear systems satisfy the above definition.

For example, we can choose $b = 0$ and $\mathbf{A} = \mathbf{Q} \in \mathbb{R}^{|\mathcal{E}| \times n}$ to be the incidence matrix of \mathcal{G} . That is, $\mathbf{Q} \in \mathbb{R}^{|\mathcal{E}| \times n}$ such that $\mathbf{Q}x = 0$ directly encodes the constraints $x_i = x_j$ for $(i, j) \in \mathcal{E}$. That is, row $e = (i, j) \in \mathcal{E}$ of matrix \mathbf{Q} contains value 1 in column i , value -1 in column j (we use an arbitrary but fixed order of nodes defining each edge in order to fix \mathbf{Q}) and zeros elsewhere. A different choice is to pick $b = 0$ and $\mathbf{A} = \mathbf{L} = \mathbf{Q}^\top \mathbf{Q}$, where \mathbf{L} is the Laplacian matrix of network \mathcal{G} . Depending on what AC system is used, the sketch and project methods can have different interpretations as gossip protocols.

In this work we mainly focus on the above two AC systems but we highlight that other choices are possible². In Section 4.2 for the provably accelerated gossip protocols we also use a normalized variant ($\|\mathbf{A}_i\|^2 = 1$) of the Incidence matrix.

3.2.1 Standard form and mass preservation

Assume that $\mathbf{A}x = b$ is an AC system. Note that since $b = 0$, the general sketch-and-project update rule (3) simplifies to:

$$x^{k+1} = [\mathbf{I} - \omega \mathbf{A}^\top \mathbf{H}_k \mathbf{A}] x^k = [\mathbf{I} - \omega \mathbf{Z}_k] x^k. \quad (13)$$

This is the standard form in which randomized gossip algorithms are written. What is new here is that the iteration matrix $\mathbf{I} - \omega \mathbf{Z}_k$ has a specific structure which guarantees convergence to x^* under very weak assumptions (see Theorem 1). Note that if $x^0 = c$, i.e., the starting primal iterate is the vector of private values (as should be expected from any gossip algorithm), then the iterates of (13) enjoy a mass preservation property (the proof follows the fact that $\mathbf{A}\mathbf{1}_n = 0$):

Theorem 3 (Mass preservation). *If $\mathbf{A}x = b$ is an AC system, then the iterates produced by (13) satisfy: $\frac{1}{n} \sum_{i=1}^n x_i^k = \bar{c}$, for all $k \geq 0$.*

Proof. Let fix $k \geq 0$ then,

$$\frac{1}{n} \mathbf{1}_n^\top x^{k+1} = \frac{1}{n} \mathbf{1}_n^\top (\mathbf{I} - \omega \mathbf{A}^\top \mathbf{H}_k \mathbf{A}) x^k = \frac{1}{n} \mathbf{1}_n^\top \mathbf{I} x^k - \frac{1}{n} \mathbf{1}_n^\top \omega \mathbf{A}^\top \mathbf{H}_k \mathbf{A} x^k \stackrel{\mathbf{A}\mathbf{1}_n=0}{=} \frac{1}{n} \mathbf{1}_n^\top x^k.$$

□

²Novel gossip algorithms can be proposed by using different AC systems to formulate the average consensus problem. For example one possibility is using the random walk normalized Laplacian $\mathbf{L}^{rw} = \mathbf{D}^{-1} \mathbf{L}$. For the case of degree-regular networks the symmetric normalized Laplacian matrix $\mathbf{L}^{sym} = \mathbf{D}^{-1/2} \mathbf{L} \mathbf{D}^{-1/2}$ can also be used.

3.2.2 ε -Averaging time

Let $z^k := \|x^k - x^*\|$. The typical measure of convergence speed employed in the randomized gossip literature, called ε -averaging time and here denoted by $T_{ave}(\varepsilon)$, represents the smallest time k for which x^k gets within εz^0 from x^* , with probability greater than $1 - \varepsilon$, uniformly over all starting values $x^0 = c$. More formally, we define

$$T_{ave}(\varepsilon) := \sup_{c \in \mathbb{R}^n} \inf \left\{ k : \mathbb{P} \left(z^k > \varepsilon z^0 \right) \leq \varepsilon \right\}.$$

This definition differs slightly from the standard one in that we use z^0 instead of $\|c\|$.

Inequality (9), together with Markov inequality, can be used to give a bound on $K(\varepsilon)$, formalized next:

Theorem 4. *Assume $\mathbf{A}x = b$ is an AC system. Let $x^0 = c$ and \mathbf{B} be positive definite diagonal matrix. Assume exactness. Then for any $0 < \varepsilon < 1$ we have*

$$T_{ave}(\varepsilon) \leq 3 \frac{\log(1/\varepsilon)}{\log(1/\rho)} \leq 3 \frac{\log(1/\varepsilon)}{1 - \rho},$$

where ρ is defined in (10).

Proof. See Appendix A.1. □

Note that under the assumptions of the above theorem, $\mathbf{W} = \mathbf{B}^{-1/2} \mathbb{E}[\mathbf{Z}] \mathbf{B}^{-1/2}$ only has a single zero eigenvalue, and hence $\lambda_{\min}^+(\mathbf{W})$ is the second smallest eigenvalue of \mathbf{W} . Thus, ρ is the second largest eigenvalue of $\mathbf{I} - \mathbf{W}$. The bound on $K(\varepsilon)$ appearing in Thm 4 is often written with ρ replaced by $\lambda_2(\mathbf{I} - \mathbf{W})$ [6].

In the rest of this section we show how two special cases of the sketch and project framework, the randomized Kaczmarz (RK) and its block variant, randomized block Kaczmarz (RBK) work as gossip algorithms for the two AC systems described above.

3.3 Randomized Kaczmarz method as gossip algorithm

As we described before the sketch and project update rule (3) has several parameters that should be chosen in advance by the user. These are the stepsize ω (relaxation parameter), the positive definite matrix \mathbf{B} and the distribution \mathcal{D} of the random matrices \mathbf{S} .

In this section we focus on one particular special case of the sketch and project framework, a scaled/weighted variant of the randomized Kaczmarz method (RK) presented in (5), and we show how this method works as gossip algorithm when applied to special systems encoding the underlying network. In particular, the linear systems that we solve are the two AC systems described in the previous section where the matrix is either the incidence matrix \mathbf{Q} or the Laplacian matrix \mathbf{L} of the network.

As we described in (5) the standard RK method can be cast as special case of the sketch and project update (3) by choosing $\mathbf{B} = \mathbf{I}$ and $\mathbf{S} = e_i$. In this section, we focus on a small modification of this algorithm and we choose the positive definite matrix \mathbf{B} to be $\mathbf{B} = \mathbf{Diag}(w_1, w_2, \dots, w_n)$, the diagonal matrix of the weights presented in the weighted average consensus problem.

Scaled RK: Let us have a general consistent linear system $\mathbf{A}x = b$ with $\mathbf{A} \in \mathbb{R}^{m \times n}$. Let us also choose $\mathbf{B} = \mathbf{Diag}(w_1, w_2, \dots, w_n)$ and $\mathbf{S}_k = e_i$, where $i \in [m]$ is chosen in each iteration independently, with probability $p_i > 0$. In this setup the update rule (3) simplifies to

$$x^{k+1} = x^k - \omega \frac{e_i^\top (\mathbf{A}x^k - b)}{e_i^\top \mathbf{A} \mathbf{B}^{-1} \mathbf{A}^\top e_i} \mathbf{B}^{-1} \mathbf{A}^\top e_i = x^k - \omega \frac{\mathbf{A}_{i:} x^k - b_i}{\|\mathbf{B}^{-1/2} \mathbf{A}_{i:}^\top\|_2^2} \mathbf{B}^{-1} \mathbf{A}_{i:}^\top. \quad (14)$$

This small modification of RK allow us to solve the more general weighted average consensus presented in Section 3.1 (and at the same time the standard average consensus problem if $\mathbf{B} = r\mathbf{I}$ where $r \in \mathbb{R}$). To the best of our knowledge, even if this variant is special case of the general Sketch and project update, was never precisely presented before in any setting.

3.3.1 AC system with incidence matrix \mathbf{Q}

Let us represent the constraints of problem (12) as linear system with matrix $\mathbf{A} = \mathbf{Q} \in \mathbb{R}^{|\mathcal{E}| \times n}$ be the Incidence matrix of the graph and right hand side $b = 0$. Lets also assume that the random matrices $\mathbf{S} \sim \mathcal{D}$ are unit coordinate vectors in $\mathbb{R}^m = \mathbb{R}^{|\mathcal{E}|}$.

Let $e = (i, j) \in \mathcal{E}$ then from the definition of matrix \mathbf{Q} we have that $\mathbf{Q}_{e:}^\top = f_i - f_j$ where f_i, f_j are unit coordinate vectors in \mathbb{R}^n . In addition, from the definition the diagonal positive definite matrix \mathbf{B} we have that

$$\|\mathbf{B}^{-1/2} \mathbf{Q}_{e:}^\top\|^2 = \|\mathbf{B}^{-1/2}(f_i - f_j)\|^2 = \frac{1}{w_1} + \frac{1}{w_j}. \quad (15)$$

Thus in this case the update rule (14) simplifies:

$$\begin{aligned} x^{k+1} &\stackrel{b=0, \mathbf{A}=\mathbf{Q}, (14)}{=} x^k - \omega \frac{\mathbf{Q}_{e:} x^k}{\|\mathbf{B}^{-1/2} \mathbf{Q}_{e:}^\top\|^2} \mathbf{B}^{-1} \mathbf{Q}_{e:}^\top \\ &\stackrel{(15)}{=} x^k - \omega \frac{\mathbf{Q}_{e:} x^k}{\frac{1}{w_1} + \frac{1}{w_j}} \mathbf{B}^{-1} \mathbf{Q}_{e:}^\top \\ &= x^k - \frac{\omega(x_i^k - x_j^k)}{\frac{1}{w_i} + \frac{1}{w_j}} \left(\frac{1}{w_i} f_i - \frac{1}{w_j} f_j \right). \end{aligned} \quad (16)$$

From (16) it can be easily seen that only the values of coordinates i and j update their values. These coordinates correspond to the private values x_i^k and x_j^k of the nodes of the selected edge $e = (i, j)$. In particular the values of x_i^k and x_j^k are updated as follows:

$$x_i^{k+1} = \left(1 - \omega \frac{w_j}{w_j + w_i}\right) x_i^k + \omega \frac{w_j}{w_j + w_i} x_j^k \quad \text{and} \quad x_j^{k+1} = \omega \frac{w_i}{w_j + w_i} x_i^k + \left(1 - \omega \frac{w_i}{w_j + w_i}\right) x_j^k. \quad (17)$$

Remark 1. In the special case that $\mathbf{B} = r\mathbf{I}$ where $r \in \mathbb{R}$ (we solve the standard average consensus problem) the update of the two nodes is simplified to

$$x_i^{k+1} = \left(1 - \frac{\omega}{2}\right) x_i^k + \frac{\omega}{2} x_j^k \quad \text{and} \quad x_j^{k+1} = \frac{\omega}{2} x_i^k + \left(1 - \frac{\omega}{2}\right) x_j^k.$$

If we further select $\omega = 1$ then this becomes:

$$x_i^{k+1} = x_j^{k+1} = \frac{x_i^k + x_j^k}{2}, \quad (18)$$

which is the update of the standard pairwise randomized gossip algorithm first presented and analyzed in [6].

3.3.2 AC system with Laplacian matrix \mathbf{L}

The AC system takes the form $\mathbf{L}x = 0$, where matrix $\mathbf{L} \in \mathbb{R}^{n \times n}$ is the Laplacian matrix of the network. In this case, each row of the matrix corresponds to a node. Using the definition of the Laplacian, we have that $\mathbf{L}_{i:}^\top = d_i f_i - \sum_{j \in \mathcal{N}_i} f_j$, where f_i, f_j are unit coordinate vectors in \mathbb{R}^n and d_i is the degree of node $i \in \mathcal{V}$.

Thus, by letting $\mathbf{B} = \text{Diag}(w_1, w_2, \dots, w_n)$ to be the diagonal matrix of the weights we obtain:

$$\|\mathbf{B}^{-1/2} \mathbf{L}_{i:}^\top\|^2 = \left\| \mathbf{B}^{-1/2} (d_i f_i - \sum_{j \in \mathcal{N}_i} f_j) \right\|^2 = \frac{d_i^2}{w_i} + \sum_{j \in \mathcal{N}_i} \frac{1}{w_j}. \quad (19)$$

In this case, the update rule (14) simplifies to:

$$\begin{aligned}
x^{k+1} &\stackrel{b=0, \mathbf{A}=\mathbf{L}, (14)}{=} x^k - \omega \frac{\mathbf{L}_{i,:} x^k}{\|\mathbf{B}^{-1/2} \mathbf{L}_{i,:}^\top\|_2^2} \mathbf{B}^{-1} \mathbf{L}_{i,:}^\top \\
&\stackrel{(19)}{=} x^k - \omega \frac{\mathbf{L}_{i,:} x^k}{\frac{d_i^2}{w_i} + \sum_{j \in \mathcal{N}_i} \frac{1}{w_j}} \mathbf{B}^{-1} \mathbf{L}_{i,:}^\top \\
&= x^k - \frac{\omega \left(d_i x_i^k - \sum_{j \in \mathcal{N}_i} x_j^k \right)}{\frac{d_i^2}{w_i} + \sum_{j \in \mathcal{N}_i} \frac{1}{w_j}} \left(\frac{d_i}{w_i} f_i - \sum_{j \in \mathcal{N}_i} \frac{1}{w_j} f_j \right). \tag{20}
\end{aligned}$$

From (20), it is clear that only coordinates $\{i\} \cup \mathcal{N}_i$ update their values. All the other coordinates remain unchanged. In particular, the value of the selected node i (coordinate i) is updated as follows:

$$x_i^{k+1} = x_i^k - \frac{\omega \left(d_i x_i^k - \sum_{j \in \mathcal{N}_i} x_j^k \right)}{\frac{d_i^2}{w_i} + \sum_{j \in \mathcal{N}_i} \frac{1}{w_j}} \frac{d_i}{w_i}, \tag{21}$$

while the values of its neighbors $j \in \mathcal{N}_i$ are updated as:

$$x_j^{k+1} = x_j^k + \frac{\omega \left(d_i x_i^k - \sum_{\ell \in \mathcal{N}_i} x_\ell^k \right)}{\frac{d_i^2}{w_i} + \sum_{\ell \in \mathcal{N}_i} \frac{1}{w_\ell}} \frac{1}{w_j}. \tag{22}$$

Remark 2. Let $\omega = 1$ and $\mathbf{B} = r\mathbf{I}$ where $r \in \mathbb{R}$ then the selected nodes update their values as follows:

$$x_i^{k+1} = \frac{\sum_{\ell \in \{i\} \cup \mathcal{N}_i} x_\ell^k}{d_i + 1} \quad \text{and} \quad x_j^{k+1} = x_j^k + \frac{(d_i x_i^k - \sum_{\ell \in \mathcal{N}_i} x_\ell^k)}{d_i^2 + d_i}. \tag{23}$$

That is, the selected node i updates its value to the average of its neighbors and itself, while all the nodes $j \in \mathcal{N}_i$ update their values using the current value of node i and all nodes in \mathcal{N}_i .

In a wireless network, to implement such an update, node i would first broadcast its current value to all of its neighbors. Then it would need to receive values from each neighbor to compute the sums over \mathcal{N}_i , after which node i would broadcast the sum to all neighbors (since there may be two neighbors $j_1, j_2 \in \mathcal{N}_i$ for which $(j_1, j_2) \notin \mathcal{E}$). In a wired network, using standard concepts from the MPI library, such an update rule could be implemented efficiently by defining a process group consisting of $\{i\} \cup \mathcal{N}_i$, and performing one **Broadcast** in this group from i (containing x_i) followed by an **AllReduce** to sum x_ℓ over $\ell \in \mathcal{N}_i$. Note that the terms involving diagonal entries of \mathbf{B} and the degrees d_i could be sent once, cached, and reused throughout the algorithm execution to reduce communication overhead.

3.3.3 Details on complexity results

Recall that the convergence rate of the sketch and project method (Algorithm 1) is equivalent to:

$$\rho := 1 - \omega(2 - \omega)\lambda_{\min}^+(\mathbf{W}),$$

where $\omega \in (0, 2)$ and $\mathbf{W} = \mathbf{B}^{-1/2} \mathbf{A}^\top \mathbb{E}[\mathbf{H}] \mathbf{A} \mathbf{B}^{-1/2}$ (from Theorem 1). In this subsection we explain how the convergence rate of the scaled RK method (14) is modified for different choices of the main parameters of the method.

Let us choose $\omega = 1$ (no over-relaxation). In this case, the rate is simplified to $\rho = 1 - \lambda_{\min}^+$.

Note that the different ways of modeling the problem (AC system) and the selection of the main parameters (weight matrix \mathbf{B} and distribution \mathcal{D}) determine the convergence rate of the method through the spectrum of matrix \mathbf{W} .

Recall that in the k^{th} iterate of the scaled RK method (14) a random vector $\mathbf{S}_k = e_i$ is chosen with probability $p_i > 0$. For convenience, let us choose³:

$$p_i = \frac{\|\mathbf{B}^{-1/2} \mathbf{A}_{i:}^\top\|^2}{\|\mathbf{B}^{-1/2} \mathbf{A}^\top\|_F^2}. \quad (24)$$

Then we have that:

$$\begin{aligned} \mathbb{E}[\mathbf{H}] &= \mathbb{E}[\mathbf{S}(\mathbf{S}^\top \mathbf{A} \mathbf{B}^{-1} \mathbf{A}^\top \mathbf{S})^\dagger \mathbf{S}^\top] \\ &= \sum_{i=1}^m p_i \frac{e_i e_i^\top}{e_i^\top \mathbf{A} \mathbf{B}^{-1} \mathbf{A}^\top e_i} = \sum_{i=1}^m p_i \frac{e_i e_i^\top}{\|\mathbf{A}_{i:}^\top\|_{\mathbf{B}^{-1}}^2} = \sum_{i=1}^m p_i \frac{e_i e_i^\top}{\|\mathbf{B}^{-1/2} \mathbf{A}_{i:}^\top\|^2} \\ &\stackrel{(24)}{=} \sum_{i=1}^m \frac{e_i e_i^\top}{\|\mathbf{B}^{-1/2} \mathbf{A}^\top\|_F^2} = \frac{1}{\|\mathbf{B}^{-1/2} \mathbf{A}^\top\|_F^2} \mathbf{I}, \end{aligned} \quad (25)$$

and

$$\mathbf{W} \stackrel{(24),(25)}{=} \frac{\mathbf{B}^{-1/2} \mathbf{A}^\top \mathbf{A} \mathbf{B}^{-1/2}}{\|\mathbf{B}^{-1/2} \mathbf{A}^\top\|_F^2}. \quad (26)$$

Incidence Matrix: Let us choose the AC system to be the one with the incidence matrix $\mathbf{A} = \mathbf{Q}$. Then $\|\mathbf{B}^{-1/2} \mathbf{Q}^\top\|_F^2 = \sum_{i=1}^n \frac{d_i}{b_i}$ and we obtain

$$\mathbf{W} \stackrel{\mathbf{A}=\mathbf{Q},(26)}{=} \frac{\mathbf{B}^{-1/2} \mathbf{L} \mathbf{B}^{-1/2}}{\|\mathbf{B}^{-1/2} \mathbf{Q}^\top\|_F^2} = \frac{\mathbf{B}^{-1/2} \mathbf{L} \mathbf{B}^{-1/2}}{\sum_{i=1}^n \frac{d_i}{\mathbf{B}_{ii}}}.$$

If we further have $\mathbf{B} = \mathbf{D}$, then $\mathbf{W} = \frac{\mathbf{D}^{-1/2} \mathbf{L} \mathbf{D}^{-1/2}}{n}$ and the convergence rate simplifies to:

$$\rho = 1 - \frac{\lambda_{\min}^+(\mathbf{D}^{-1/2} \mathbf{L} \mathbf{D}^{-1/2})}{n} = 1 - \frac{\lambda_{\min}^+(\mathbf{L}^{\text{sym}})}{n}.$$

If $\mathbf{B} = r\mathbf{I}$ where $r \in \mathbb{R}$ (solve the standard average consensus problem), then $\mathbf{W} = \frac{\mathbf{L}}{\|\mathbf{Q}\|_F^2} = \frac{\mathbf{L}}{\sum_{i=1}^n d_i} = \frac{\mathbf{L}}{2m}$ and the convergence rate simplifies to

$$\rho = 1 - \frac{\lambda_{\min}^+(\mathbf{L})}{2m} = 1 - \frac{\alpha(\mathcal{G})}{2m}. \quad (27)$$

The convergence rate (27) is identical to the rate proposed for the convergence of the standard pairwise gossip algorithm in [6]. Recall that in this special case the proposed gossip protocol has exactly the same update rule with the algorithm presented in [6], see equation (18).

Laplacian Matrix: If we choose to formulate the AC system using the Laplacian matrix \mathbf{L} , that is $\mathbf{A} = \mathbf{L}$, then $\|\mathbf{B}^{-1/2} \mathbf{L}^\top\|_F^2 = \sum_{i=1}^n \frac{d_i(d_i+1)}{\mathbf{B}_{ii}}$ and we have:

$$\mathbf{W} \stackrel{\mathbf{A}=\mathbf{L},(26)}{=} \frac{\mathbf{B}^{-1/2} \mathbf{L}^\top \mathbf{L} \mathbf{B}^{-1/2}}{\sum_{i=1}^n \frac{d_i(d_i+1)}{\mathbf{B}_{ii}}}.$$

If $\mathbf{B} = \mathbf{D}$, then the convergence rate simplifies to:

$$\rho = 1 - \frac{\lambda_{\min}^+(\mathbf{D}^{-1/2} \mathbf{L}^\top \mathbf{L} \mathbf{D}^{-1/2})}{\sum_{i=1}^n (d_i + 1)} = 1 - \frac{\lambda_{\min}^+(\mathbf{D}^{-1/2} \mathbf{L}^2 \mathbf{D}^{-1/2})}{n + \sum_{i=1}^n d_i} \stackrel{\sum_{i=1}^n d_i = 2m}{=} 1 - \frac{\lambda_{\min}^+(\mathbf{D}^{-1/2} \mathbf{L}^2 \mathbf{D}^{-1/2})}{n + 2m}.$$

If $\mathbf{B} = r\mathbf{I}$, where $r \in \mathbb{R}$, then $\mathbf{W} = \frac{\mathbf{L}^2}{\|\mathbf{L}\|_F^2} = \frac{\mathbf{L}^2}{\sum_{i=1}^n d_i(d_i+1)}$ and the convergence rate simplifies to

$$\rho = 1 - \frac{\lambda_{\min}^+(\mathbf{L}^2)}{\sum_{i=1}^n d_i(d_i + 1)} = 1 - \frac{\alpha(\mathcal{G})^2}{\sum_{i=1}^n d_i(d_i + 1)}.$$

³Similar probabilities have been chosen in [25] for the convergence of the standard RK method ($\mathbf{B} = \mathbf{I}$). The distribution \mathcal{D} of the matrices \mathbf{S} used in equation (24) is common in the area of randomized iterative methods for linear systems and is used to simplify the analysis and the expressions of the convergence rates. For more choices of distributions we refer the interested reader to [25]. It is worth to mention that the probability distribution that optimizes the convergence rate of the RK and other projection methods can be expressed as the solution to a convex semidefinite program [25, 11].

3.4 Block gossip algorithms

Up to this point we focused on the basic connections between the convergence analysis of the sketch and project methods and the literature of randomized gossip algorithms. We show how specific variants of the randomized Kaczmarz method (RK) can be interpreted as gossip algorithms for solving the weighted and standard average consensus problems.

In this part we extend the previously described methods to their block variants related to randomized block Kaczmarz (RBK) method (6). In particular, in each step of the sketch and project method (3), the random matrix \mathbf{S} is selected to be a random column submatrix of the $m \times m$ identity matrix corresponding to columns indexed by a random subset $C \subseteq [m]$. That is, $\mathbf{S} = \mathbf{I}_{:,C}$, where a set $C \subseteq [m]$ is chosen in each iteration independently, with probability $p_C \geq 0$ (see equation (6)). Note that in the special case that set C is a singleton with probability 1 the algorithm is simply the randomized Kaczmarz method of the previous section.

To keep things simple, we assume that $\mathbf{B} = \mathbf{I}$ (standard average consensus, without weights) and choose the stepsize $\omega = 1$. In the next section, we will describe gossip algorithms with heavy ball momentum and explain in detail how the gossip interpretation of RBK change in the more general case of $\omega \in (0, 2)$.

Similar to the previous subsections, we formulate the consensus problem using either $\mathbf{A} = \mathbf{Q}$ or $\mathbf{A} = \mathbf{L}$ as the matrix in the AC system. In this setup, the iterative process (3) has the form:

$$x^{k+1} \stackrel{(6),(13)}{=} x^k - \mathbf{A}^\top \mathbf{I}_{:,C} (\mathbf{I}_{:,C}^\top \mathbf{A} \mathbf{A}^\top \mathbf{I}_{:,C})^\dagger \mathbf{I}_{:,C}^\top \mathbf{A} x^k = x^k - \mathbf{A}_{C,:}^\top (\mathbf{A}_{C,:} \mathbf{A}_{C,:}^\top)^\dagger \mathbf{A}_{C,:} x^k, \quad (28)$$

which, as explained in the introduction, can be equivalently written as:

$$x^{k+1} = \underset{x \in \mathbb{R}^n}{\operatorname{argmin}} \{ \|x - x^k\|^2 : \mathbf{I}_{:,C}^\top \mathbf{A} x = 0 \}. \quad (29)$$

Essentially in each step of this method the next iterate is evaluated to be the projection of the current iterate x^k onto the solution set of a row subsystem of $\mathbf{A}x = 0$.

AC system with Incidence Matrix: In the case that $\mathbf{A} = \mathbf{Q}$ the selected rows correspond to a random subset $C \subseteq \mathcal{E}$ of selected edges. While (28) may seem to be a complicated algebraic (resp. variational) characterization of the method, due to our choice of $\mathbf{A} = \mathbf{Q}$ we have the following result which gives a natural interpretation of RBK as a gossip algorithm (see also Figure 1).

Theorem 5 (RBK as Gossip algorithm: RBKG). *Consider the AC system with the constraints being expressed using the Incidence matrix \mathbf{Q} . Then each iteration of RBK (Algorithm (28)) works as gossip algorithm as follows:*

1. Select a random set of edges $C \subseteq \mathcal{E}$,
2. Form subgraph \mathcal{G}_k of \mathcal{G} from the selected edges
3. For each connected component of \mathcal{G}_k , replace node values with their average.

Proof. See Appendix A.2. □

Using the convergence result of general Theorem 1 and the form of matrix \mathbf{W} (recall that in this case we assume $\mathbf{B} = \mathbf{I}$, $\mathbf{S} = \mathbf{I}_{:,C} \sim \mathcal{D}$ and $\omega = 1$), we obtain the following complexity for the algorithm:

$$\mathbb{E}[\|x^k - x^*\|^2] \leq \left[1 - \lambda_{\min}^+ \left(\mathbb{E} \left[\mathbf{Q}_{C,:}^\top (\mathbf{Q}_{C,:} \mathbf{Q}_{C,:}^\top)^\dagger \mathbf{Q}_{C,:} \right] \right) \right]^k \|x^0 - x^*\|^2. \quad (30)$$

For more details on the above convergence rate of randomized block Kaczmarz method with meaningfully bounds on the rate in a more general setting we suggest the papers [57, 58].

There is a very closed relationship between the gossip interpretation of RBK explained in Theorem 5 and several existing randomized gossip algorithms that in each step update the values of more than two nodes. For example the *path averaging* algorithm proposed in [5] is a special case

of RBK, when set C is restricted to correspond to a path of vertices. That is, in path averaging, in each iteration a path of nodes is selected and the nodes that belong to it update their values to their exact average. A different example is the recently proposed clique gossiping [44] where the network is already divided into cliques and through a random procedure a clique is activated and the nodes of it update their values to their exact average. In [6] a synchronous variant of gossip algorithm is presented where in each step multiple node pairs communicate exactly at the same time with the restriction that these simultaneously active node pairs are disjoint.

It is easy to see that all of the above algorithms can be cast as special cases of RBK if the distribution \mathcal{D} of the random matrices is chosen carefully to be over random matrices \mathbf{S} (column sub-matrices of Identity) that update specific set of edges in each iteration. As a result our general convergence analysis can recover the complexity results proposed in the above works.

Finally, as we mentioned, in the special case in which set C is always a singleton, Algorithm (28) reduces to the standard randomized Kaczmarz method. This means that only a random edge is selected in each iteration and the nodes incident with this edge replace their local values with their average. This is the pairwise gossip algorithm of Boyd et al. [6] presented in equation (18). Theorem 5 extends this interpretation to the case of the RBK method.

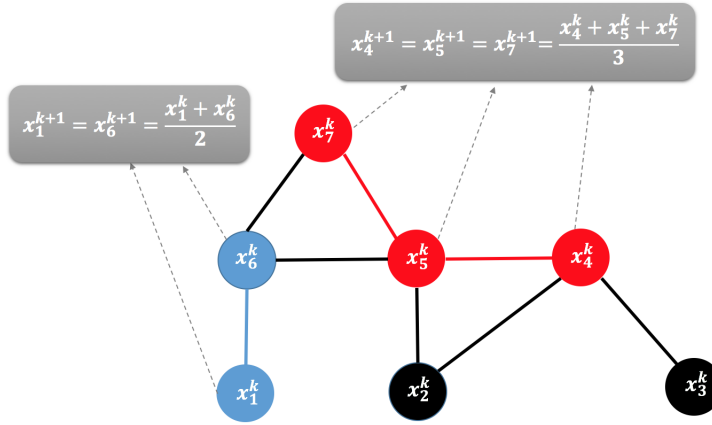


Figure 1: Example of how the RBK method works as gossip algorithm in case of AC system with Incidence matrix. In the presented network 3 edges are randomly selected and a subgraph of two connected components (blue and red) is formed. Then the nodes of each connected component update their private values to their average.

AC system with Laplacian Matrix: For this choice of AC system the update is more complicated. To simplify the way that the block variant work as gossip we make an extra assumption. We assume that the selected rows of the constraint $\mathbf{I}_C^\top \mathbf{L} \mathbf{x} = 0$ in update (29) have no-zero elements at different coordinates. This allows to have a direct extension of the serial variant presented in Remark 2. Thus, in this setup, the RBK update rule (28) works as gossip algorithm as follows:

1. $|C|$ nodes are activated (with restriction that the nodes are not neighbors and they do not share common neighbors)
2. For each node $i \in C$ we have the following update:

$$x_i^{k+1} = \frac{\sum_{\ell \in \{i \cup \mathcal{N}_i\}} x_\ell^k}{d_i + 1} \quad \text{and} \quad x_j^{k+1} = x_j^k + \frac{(d_i x_i^k - \sum_{\ell \in \mathcal{N}_i} x_\ell^k)}{d_i^2 + d_i}. \quad (31)$$

The above update rule can be seen as a parallel variant of update (23). Similar to the convergence in the case of Incidence matrix, the RBK for solving the AC system with a Laplacian matrix converges to x^* with the following rate (using result of Theorem 1):

$$\mathbb{E}[\|x^k - x^*\|^2] \leq \left[1 - \lambda_{\min}^+ \left(\mathbb{E} \left[\mathbf{L}_C^\top (\mathbf{L}_C \mathbf{L}_C^\top)^\dagger \mathbf{L}_C \right] \right) \right]^k \|x^0 - x^*\|^2.$$

4 Faster and Provably Accelerated Randomized Gossip Algorithms

The main goal in the design of gossip protocols is for the computation and communication to be done as quickly and efficiently as possible. In this section, our focus is precisely this. We design randomized gossip protocols which converge to consensus fast with provable accelerated linear rates. To the best of our knowledge, the proposed protocols are the first randomized gossip algorithms that converge to consensus with an accelerated linear rate.

In particular, we present novel protocols for solving the average consensus problem where in each step all nodes of the network update their values but only a subset of them exchange their private values. The protocols are inspired from the recently developed accelerated variants of randomized Kaczmarz-type methods for solving consistent linear systems where the addition of momentum terms on top of the sketch and project update rule provides better theoretical and practical performance.

In the area of optimization algorithms, there are two popular ways to accelerate an algorithm using momentum. The first one is using the Polyak’s heavy ball momentum [67] and the second one is using the theoretically much better understood momentum introduced by Nesterov [59, 61]. Both momentum approaches have been recently proposed and analyzed to improve the performance of randomized iterative methods for solving linear systems.

To simplify the presentation, the accelerated algorithms and their convergence rates are presented for solving the standard average consensus problem ($\mathbf{B} = \mathbf{I}$). Using a similar approach as in the previous section, the update rules and the convergence rates can be easily modified to solve the more general weighted average consensus problem. For the protocols in this section we use the incidence matrix $\mathbf{A} = \mathbf{Q}$ or its normalized variant to formulate the AC system.

4.1 Gossip algorithms with heavy ball momentum

The recent works [47, 48] propose and analyze heavy ball momentum variants of several stochastic optimization algorithms for solving stochastic quadratic optimization problems and linear systems. One of the proposed algorithms is the sketch and project method (Algorithm 1) with heavy ball momentum. In particular, the authors focus on explaining how the method can be interpreted as SGD with momentum—also known as the stochastic heavy ball method (SHB). SHB is a well known algorithm in the optimization literature for solving stochastic optimization problems, and extremely popular in areas such as deep learning [78, 79, 37, 85]. However, even if SHB is used extensively in practice, its theoretical convergence behavior is not well understood. [47, 48] were the first works that prove linear convergence of SHB in any setting.

In this subsection we focus on the sketch and project method with heavy ball momentum. We present the main theorems showing its convergence performance as presented in [47, 48] and explain how special cases of the general method work as gossip algorithms when are applied to a special system encoding the underlying network.

4.1.1 Sketch and project with heavy ball momentum

The update rule of the sketch and project method with heavy ball momentum as proposed and analyzed in [47, 48] is formally presented in the following algorithm:

Algorithm 2 Sketch and Project with Heavy Ball Momentum

- 1: **Parameters:** Distribution \mathcal{D} from which method samples matrices; stepsize/relaxation parameter $\omega \in \mathbb{R}$; momentum parameter β .
 - 2: **Initialize:** $x^0, x^1 \in \mathbb{R}^n$
 - 3: **for** $k = 1, 2, \dots$ **do**
 - 4: Draw a fresh $\mathbf{S}_k \sim \mathcal{D}$
 - 5: Set

$$x^{k+1} = x^k - \omega \mathbf{B}^{-1} \mathbf{A}^\top \mathbf{S}_k (\mathbf{S}_k^\top \mathbf{A} \mathbf{B}^{-1} \mathbf{A}^\top \mathbf{S}_k)^\dagger \mathbf{S}_k^\top (\mathbf{A} x^k - b) + \beta (x^k - x^{k-1}). \quad (32)$$
 - 6: **end for**
 - 7: **Output:** The last iterate x^k
-

Using, $\mathbf{B} = \mathbf{I}$ and the same choice of distribution \mathcal{D} as in equations (5) and (6) we can now obtain momentum variants of the RK and RBK as special case of the above algorithm as follows:

- RK with momentum (mRK):

$$x^{k+1} = x^k - \omega \frac{\mathbf{A}_{i:} x^k - b_i}{\|\mathbf{A}_{i:}\|^2} \mathbf{A}_{i:}^\top + \beta(x^k - x^{k-1}). \quad (33)$$

- RBK with momentum (mRBK):

$$x^{k+1} = x^k - \omega \mathbf{A}_{C:}^\top (\mathbf{A}_C \mathbf{A}_{C:}^\top)^\dagger (\mathbf{A}_C x^k - b_C) + \beta(x^k - x^{k-1}). \quad (34)$$

In [48], two main theoretical results describing the behavior of Algorithm 2 (and as a result also the special cases mRK and mRBK) have been presented⁴.

Theorem 6 (Theorem 1, [48]). *Choose $x^0 = x^1 \in \mathbb{R}^n$. Let $\{x^k\}_{k=0}^\infty$ be the sequence of random iterates produced by Algorithm 2 and let assume exactness. Assume $0 < \omega < 2$ and $\beta \geq 0$ and that the expressions $a_1 := 1 + 3\beta + 2\beta^2 - (\omega(2 - \omega) + \omega\beta)\lambda_{\min}^+$ and $a_2 := \beta + 2\beta^2 + \omega\beta\lambda_{\max}$ satisfy $a_1 + a_2 < 1$. Set $x^* = \Pi_{\mathcal{L}, \mathbf{B}}(x^0)$. Then*

$$\mathbb{E}[\|x^k - x^*\|^2] \leq q^k(1 + \delta)\|x^0 - x^*\|^2, \quad (35)$$

where $q = \frac{1}{2}(a_1 + \sqrt{a_1^2 + 4a_2})$ and $\delta = q - a_1$. Moreover, $a_1 + a_2 \leq q < 1$.

Theorem 7 (Theorem 4, [48]). *Assume exactness. Let $\{x^k\}_{k=0}^\infty$ be the sequence of random iterates produced by Algorithm 2, started with $x^0 = x^1 \in \mathbb{R}^n$, with relaxation parameter (stepsize) $0 < \omega \leq 1/\lambda_{\max}$ and momentum parameter $(1 - \sqrt{\omega\lambda_{\min}^+})^2 < \beta < 1$. Let $x^* = \Pi_{\mathcal{L}, \mathbf{I}}(x^0)$. Then there exists a constant $C > 0$ such that for all $k \geq 0$ we have*

$$\|\mathbb{E}[x^k - x^*]\|^2 \leq \beta^k C.$$

Using Theorem 7 and by a proper combination of the stepsize ω and the momentum parameter β , Algorithm 2 enjoys an accelerated linear convergence rate in mean [48].

Corollary 1. (i) *If $\omega = 1$ and $\beta = (1 - \sqrt{0.99\lambda_{\min}^+})^2$, then the iteration complexity of Algorithm 2 becomes: $O\left(\sqrt{1/\lambda_{\min}^+} \log(1/\epsilon)\right)$.*

(ii) *If $\omega = 1/\lambda_{\max}$ and $\beta = (1 - \sqrt{0.99\lambda_{\min}^+/\lambda_{\max}})^2$, then the iteration complexity of Algorithm 2 becomes: $O\left(\sqrt{\lambda_{\max}/\lambda_{\min}^+} \log(1/\epsilon)\right)$.*

Having presented Algorithm 2 and its convergence analysis results, let us now describe its behavior as a randomized gossip protocol when applied on the AC system $\mathbf{A}x = 0$ with $\mathbf{A} = \mathbf{Q} \in |\mathcal{E}| \times n$ (incidence matrix of the network).

Note that since $b = 0$ (from the AC system definition), method (32) can be simplified to:

$$x^{k+1} = \left[\mathbf{I} - \omega \mathbf{A}^\top \mathbf{S}_k (\mathbf{S}_k^\top \mathbf{A} \mathbf{A}^\top \mathbf{S}_k)^\dagger \mathbf{S}_k^\top \mathbf{A} \right] x^k + \beta(x^k - x^{k-1}). \quad (36)$$

In the rest of this section we focus on two special cases of (36): RK with heavy ball momentum (equation (33) with $b_i = 0$) and RBK with heavy ball momentum (equation (34) with $b_C = 0$).

Algorithm 3 mRK: Randomized Kaczmarz with momentum as a gossip algorithm

- 1: **Parameters:** Distribution \mathcal{D} from which method samples matrices; stepsize/relaxation parameter $\omega \in \mathbb{R}$; heavy ball/momentum parameter β .
 - 2: **Initialize:** $x^0, x^1 \in \mathbb{R}^n$
 - 3: **for** $k = 1, 2, \dots$ **do**
 - 4: Pick an edge $e = (i, j)$ following the distribution \mathcal{D}
 - 5: The values of the nodes are updated as follows:
 - Node i : $x_i^{k+1} = \frac{2-\omega}{2}x_i^k + \frac{\omega}{2}x_j^k + \beta(x_i^k - x_i^{k-1})$
 - Node j : $x_j^{k+1} = \frac{2-\omega}{2}x_j^k + \frac{\omega}{2}x_i^k + \beta(x_j^k - x_j^{k-1})$
 - Any other node ℓ : $x_\ell^{k+1} = x_\ell^k + \beta(x_\ell^k - x_\ell^{k-1})$
 - 6: **end for**
 - 7: **Output:** The last iterate x^k
-

4.1.2 Randomized Kaczmarz gossip with heavy ball momentum

As we have seen in previous section when the standard RK is applied to solve the AC system $\mathbf{Q}x = 0$, one can recover the famous pairwise gossip algorithm [6]. Algorithm 3 describes how a relaxed variant of randomized Kaczmarz with heavy ball momentum ($0 < \omega < 2$ and $0 \leq \beta < 1$) behaves as a gossip algorithm. See also Figure (2) for a graphical illustration of the method.

Remark 3. *In the special case that $\beta = 0$ (zero momentum) only the two nodes of edge $e = (i, j)$ update their values. In this case the two selected nodes do not update their values to their exact average but to a convex combination that depends on the stepsize $\omega \in (0, 2)$. To obtain the pairwise gossip algorithm of [6], one should further choose $\omega = 1$.*

Distributed Nature of the Algorithm: Here we highlight a few ways to implement mRK in a distributed fashion.

- *Pairwise broadcast gossip:* In this protocol each node $i \in \mathcal{V}$ of the network \mathcal{G} has a clock that ticks at the times of a rate 1 Poisson process. The inter-tick times are exponentially distributed, independent across nodes, and independent across time. This is equivalent to a global clock ticking at a rate n Poisson process which wakes up an edge of the network at random. In particular, in this implementation mRK works as follows: In the k^{th} iteration (time slot) the clock of node i ticks and node i randomly contact one of its neighbors and simultaneously broadcast a signal to inform the nodes of the whole network that is updating (this signal does not contain any private information of node i). The two nodes (i, j) share their information and update their private values following the update rule of Algorithm 3 while all the other nodes update their values using their own information. In each iteration only one pair of nodes exchange their private values.
- *Synchronous pairwise gossip:* In this protocol a single global clock is available to all nodes. The time is assumed to be slotted commonly across nodes and in each time slot only a pair of nodes of the network is randomly activated and exchange their information following the update rule of Algorithm 3. The remaining not activated nodes update their values using their own last two private values. Note that this implementation of mRK comes with the disadvantage that it requires a central entity which in each step requires to choose the activated pair of nodes⁵.
- *Asynchronous pairwise gossip with common counter:* Note that the update rule of the selected

⁴Note that in [48] the analysis have been made on the same framework of Theorem 1 with general positive definite matrix \mathbf{B} .

⁵We speculate that a completely distributed synchronous gossip algorithm that finds pair of nodes in a distributed manner without any additional computational burden can be design following the same procedure proposed in Section III.C of [6].

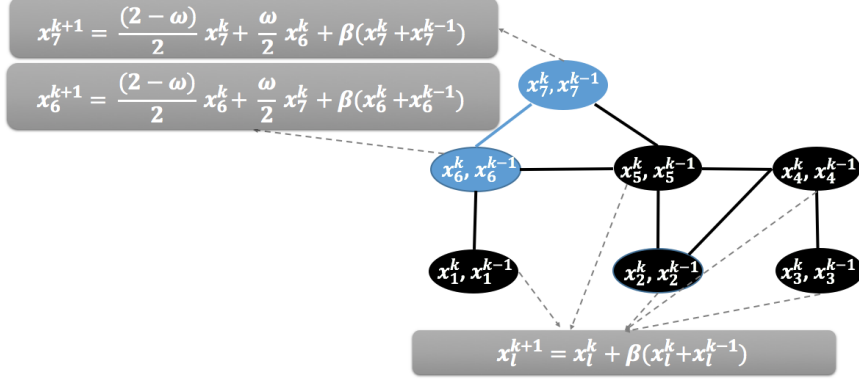


Figure 2: Example of how mRK works as gossip algorithm. In the presented network the edge that connects nodes 6 and 7 is randomly selected. The pair of nodes exchange their information and update their values following the update rule of the Algorithm 3 while the rest of the nodes, $\ell \in [5]$, update their values using only their own previous private values.

pair of nodes (i, j) in Algorithm 3 can be rewritten as follows:

$$x_i^{k+1} = x_i^k + \beta(x_i^k - x_i^{k-1}) + \frac{\omega}{2}(x_j^k - x_i^k),$$

$$x_j^{k+1} = x_j^k + \beta(x_j^k - x_j^{k-1}) + \frac{\omega}{2}(x_i^k - x_j^k).$$

In particular observe that the first part of the above expressions $x_i^k + \beta(x_i^k - x_i^{k-1})$ (for the case of node i) is exactly the same with the update rule of the non activate nodes at k^{th} iterate (check step 5 of Algorithm 3). Thus, if we assume that all nodes share a common counter that keeps track of the current iteration count and that each node $i \in \mathcal{V}$ remembers the iteration counter k_i of when it was last activated, then step 5 of Algorithm 3 takes the form:

- $x_i^{k+1} = i_k \left[x_i^k + \beta(x_i^k - x_i^{k-1}) \right] + \frac{\omega}{2}(x_j^k - x_i^k),$
- $x_j^{k+1} = j_k \left[x_j^k + \beta(x_j^k - x_j^{k-1}) \right] + \frac{\omega}{2}(x_i^k - x_j^k),$
- $k_i = k_j = k + 1,$
- Any other node ℓ : $x_\ell^{k+1} = x_\ell^k,$

where $i_k = k - k_i$ ($j_k = k - k_j$) denotes the number of iterations between the current iterate and the last time that the i^{th} (j^{th}) node is activated. In this implementation only a pair of nodes communicate and update their values in each iteration (thus the justification of asynchronous), however it requires the nodes to share a common counter that keeps track the current iteration count in order to be able to compute the value of $i_k = k - k_i$.

4.1.3 Connections with existing fast randomized gossip algorithms

In the randomized gossip literature there is one particular method closely related to our approach. It was first proposed in [7] and its analysis under strong conditions was presented in [40]. In this work local memory is exploited by installing shift registers at each agent. In particular we are interested in the case of two registers where the first stores the agent's current value and the second the agent's value before the latest update. The algorithm can be described as follows. Suppose that edge $e = (i, j)$ is chosen at time k . Then,

- Node i : $x_i^{k+1} = \omega \left(\frac{x_i^k + x_j^k}{2} \right) + (1 - \omega)x_i^{k-1},$

- Node j : $x_i^{k+1} = \omega(\frac{x_i^k + x_j^k}{2}) + (1 - \omega)x_j^{k-1}$,
- Any other node ℓ : $x_\ell^{k+1} = x_\ell^k$,

where $\omega \in [1, 2)$. The method was analyzed in [40] under a strong assumption on the probabilities of choosing the pair of nodes, that as the authors mentioned, is unrealistic in practical scenarios, and for networks like the random geometric graphs. At this point we should highlight that the results presented in [48] hold for essentially any distribution \mathcal{D} ⁶ and as a result in the proposed gossip variants with heavy ball momentum such problem cannot occur.

Note that, in the special case that we choose $\beta = \omega - 1$ in the update rule of Algorithm 3 is simplified to:

- Node i : $x_i^{k+1} = \omega(\frac{x_i^k + x_j^k}{2}) + (1 - \omega)x_i^{k-1}$,
- Node j : $x_j^{k+1} = \omega(\frac{x_i^k + x_j^k}{2}) + (1 - \omega)x_j^{k-1}$,
- Any other node ℓ : $x_\ell^{k+1} = \omega x_\ell^k + (1 - \omega)x_\ell^{k-1}$.

In order to apply Theorem 6, we need to assume that $0 < \omega < 2$ and $\beta = \omega - 1 \geq 0$ which also means that $\omega \in [1, 2)$. Thus for $\omega \in [1, 2)$ and momentum parameter $\beta = \omega - 1$ it is easy to see that our approach is very similar to the shift-register algorithm. Both methods update the selected pair of nodes in the same way. However, in Algorithm 3 the not selected nodes of the network do not remain idle but instead update their values using their own previous information.

By defining the momentum matrix $\mathbf{M} = \text{Diag}(\beta_1, \beta_2, \dots, \beta_n)$, the above closely related algorithms can be expressed, in vector form, as:

$$x^{k+1} = x^k - \frac{\omega}{2}(x_i^k - x_j^k)(e_i - e_j) + \mathbf{M}(x^k - x^{k-1}). \quad (37)$$

In particular, in mRK every diagonal element of matrix \mathbf{M} is equal to $\omega - 1$, while in the algorithm of [7, 40] all the diagonal elements are zeros except the two values that correspond to nodes i and j that are equal to $\beta_i = \beta_j = \omega - 1$.

Remark 4. *The shift register algorithm of [40] and Algorithm 3 of this work can be seen as the two limit cases of the update rule (37). As we mentioned, the shift register method [40] uses only two non-zero diagonal elements in \mathbf{M} , while our method has a full diagonal. We believe that further methods can be developed in the future by exploring the cases where more than two but not all elements of the diagonal matrix \mathbf{M} are non-zero. It might be possible to obtain better convergence if one carefully chooses these values based on the network topology. We leave this as an open problem for future research.*

4.1.4 Randomized block Kaczmarz gossip with heavy ball momentum

Recall that Theorem 5 explains how RBK (with no momentum and no relaxation) can be interpreted as a gossip algorithm. In this subsection by using this result we explain how relaxed RBK with momentum works. Note that the update rule of RBK with momentum can be rewritten as follows:

$$x^{k+1} \stackrel{(36),(34)}{=} \omega \left(\mathbf{I} - \mathbf{A}_C^\top (\mathbf{A}_C \mathbf{A}_C^\top)^\dagger \mathbf{A}_C \right) x^k + (1 - \omega)x^k + \beta(x^k - x^{k-1}), \quad (38)$$

and recall that $x^{k+1} = (\mathbf{I} - \mathbf{A}_C^\top (\mathbf{A}_C \mathbf{A}_C^\top)^\dagger \mathbf{A}_C) x^k$ is the update rule of the standard RBK (28).

Thus, in analogy to the standard RBK, in the k^{th} step, a random set of edges is selected and $q \leq n$ connected components are formed as a result. This includes the connected components that belong to both sub-graph \mathcal{G}_k and also the singleton connected components (nodes outside the \mathcal{G}_k). Let us define the set of the nodes that belong in the $r \in [q]$ connected component at the k^{th} step \mathcal{V}_r^k , such that $\mathcal{V} = \cup_{r \in [q]} \mathcal{V}_r^k$ and $|\mathcal{V}| = \sum_{r=1}^q |\mathcal{V}_r^k|$ for any $k > 0$.

Algorithm 4 mRBK: Randomized Block Kaczmarz Gossip with momentum

- 1: **Parameters:** Distribution \mathcal{D} from which method samples matrices; stepsize/relaxation parameter $\omega \in \mathbb{R}$; heavy ball/momentum parameter β .
 - 2: **Initialize:** $x^0, x^1 \in \mathbb{R}^n$
 - 3: **for** $k = 1, 2, \dots$ **do**
 - 4: Select a random set of edges $\mathcal{S} \subseteq \mathcal{E}$
 - 5: Form subgraph \mathcal{G}_k of \mathcal{G} from the selected edges
 - 6: Node values are updated as follows:
 - For each connected component \mathcal{V}_r^k of \mathcal{G}_k , replace the values of its nodes with:
$$x_i^{k+1} = \omega \frac{\sum_{j \in \mathcal{V}_r^k} x_j^k}{|\mathcal{V}_r^k|} + (1 - \omega)x_i^k + \beta(x_i^k - x_i^{k-1}). \quad (39)$$
 - Any other node ℓ : $x_\ell^{k+1} = x_\ell^k + \beta(x_\ell^k - x_\ell^{k-1})$
 - 7: **end for**
 - 8: **Output:** The last iterate x^k
-

Using the update rule (38), Algorithm 4 shows how mRBK is updating the private values of the nodes of the network (see also Figure 3 for the graphical interpretation).

Note that in the update rule of mRBK the nodes that are not attached to a selected edge (do not belong in the sub-graph \mathcal{G}_k) update their values via $x_\ell^{k+1} = x_\ell^k + \beta(x_\ell^k - x_\ell^{k-1})$. By considering these nodes as singleton connected components their update rule is exactly the same with the nodes of sub-graph \mathcal{G}_k . This is easy to see as follows:

$$\begin{aligned} x_\ell^{k+1} &\stackrel{(39)}{=} \omega \frac{\sum_{j \in \mathcal{V}_r^k} x_j^k}{|\mathcal{V}_r^k|} + (1 - \omega)x_\ell^k + \beta(x_\ell^k - x_\ell^{k-1}) \\ &\stackrel{|\mathcal{V}_r^k|=1}{=} \omega x_\ell^k + (1 - \omega)x_\ell^k + \beta(x_\ell^k - x_\ell^{k-1}) \\ &= x_\ell^k + \beta(x_\ell^k - x_\ell^{k-1}). \end{aligned} \quad (40)$$

Remark 5. In the special case that only one edge is selected in each iteration ($\mathbf{S}_k \in \mathbb{R}^{m \times 1}$) the update rule of mRBK is simplified to the update rule of mRK. In this case the sub-graph \mathcal{G}_k is the pair of the two selected edges.

Remark 6. In previous section we explained how several existing gossip protocols for solving the average consensus problem are special cases of the RBK (Theorem 5). For example two gossip algorithms that can be cast as special cases of the standard RBK are the path averaging proposed in [5] and the clique gossiping [44]. In path averaging, in each iteration a path of nodes is selected and its nodes update their values to their exact average ($\omega = 1$). In clique gossiping, the network is already divided into cliques and through a random procedure a clique is activated and the nodes of it update their values to their exact average ($\omega = 1$). Since mRBK contains the standard RBK as a special case (when $\beta = 0$), we expect that these special protocols can also be accelerated with the addition of momentum parameter $\beta \in (0, 1)$.

4.1.5 Mass preservation

One of the key properties of some of the most efficient randomized gossip algorithms is mass preservation. That is, the sum (and as a result the average) of the private values of the nodes remains fixed during the iterative procedure ($\sum_{i=1}^n x_i^k = \sum_{i=1}^n x_i^0, \quad \forall k \geq 1$). The original pairwise gossip algorithm proposed in [6] satisfied the mass preservation property, while existing fast gossip algorithms [7, 40] preserving a scaled sum. In this subsection we show that mRK and mRBK gossip protocols presented above satisfy the mass preservation property. In particular, we prove mass preservation for the case of the block randomized gossip protocol (Algorithm 4) with momentum.

⁶The only restriction is the exactness condition to be satisfied. See Theorem 6.

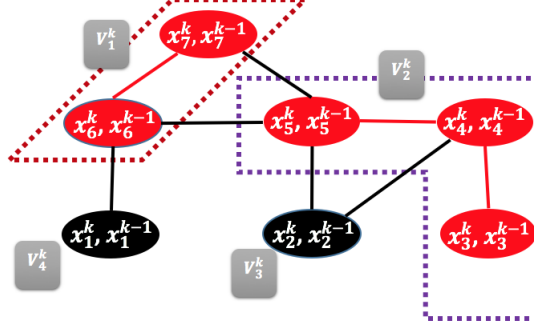


Figure 3: Example of how the mRBK method works as gossip algorithm. In the presented network the red edges are randomly chosen in the k^{th} iteration, and they form subgraph \mathcal{G}_k and four connected component. In this figure V_1^k and V_2^k are the two connected components that belong in the subgraph \mathcal{G}_k while V_3^k and V_4^k are the singleton connected components. Then the nodes update their values by communicate with the other nodes of their connected component using the update rule (39). For example the node number 5 that belongs in the connected component V_2^k will update its value using the values of node 4 and 3 that also belong in the same component as follows: $x_5^{k+1} = \omega \frac{x_3^k + x_4^k + x_5^k}{3} + (1 - \omega)x_5^k + \beta(x_5^k - x_5^{k-1})$.

This is sufficient since the randomized Kaczmarz gossip with momentum (mRK), Algorithm 3 can be cast as special case.

Theorem 8. Assume that $x^0 = x^1 = c$. That is, the two registers of each node have the same initial value. Then for the Algorithms 3 and 4 we have $\sum_{i=1}^n x_i^k = \sum_{i=1}^n c_i$ for any $k \geq 0$ and as a result, $\frac{1}{n} \sum_{i=1}^n x_i^k = \bar{c}$.

Proof. We prove the result for the more general Algorithm 4. Assume that in the k^{th} step of the method q connected components are formed. Let the set of the nodes of each connected component be \mathcal{V}_r^k so that $\mathcal{V} = \cup_{r=\{1,2,\dots,q\}} \mathcal{V}_r^k$ and $|\mathcal{V}| = \sum_{r=1}^q |\mathcal{V}_r^k|$ for any $k > 0$. Thus:

$$\sum_{i=1}^n x_i^{k+1} = \sum_{i \in \mathcal{V}_1^k} x_i^{k+1} + \dots + \sum_{i \in \mathcal{V}_q^k} x_i^{k+1}. \quad (41)$$

Let us first focus, without loss of generality, on connected component $r \in [q]$ and simplify the expression for the sum of its nodes:

$$\begin{aligned} \sum_{i \in \mathcal{V}_r^k} x_i^{k+1} &\stackrel{(39)}{=} \sum_{i \in \mathcal{V}_r^k} \omega \frac{\sum_{j \in \mathcal{V}_r^k} x_j^k}{|\mathcal{V}_r^k|} + (1 - \omega) \sum_{i \in \mathcal{V}_r^k} x_i^k + \beta \sum_{i \in \mathcal{V}_r^k} (x_i^k - x_i^{k-1}) \\ &= |\mathcal{V}_r^k| \frac{\omega \sum_{j \in \mathcal{V}_r^k} x_j^k}{|\mathcal{V}_r^k|} + (1 - \omega) \sum_{i \in \mathcal{V}_r^k} x_i^k + \beta \sum_{i \in \mathcal{V}_r^k} (x_i^k - x_i^{k-1}) \\ &= (1 + \beta) \sum_{i \in \mathcal{V}_r^k} x_i^k - \beta \sum_{i \in \mathcal{V}_r^k} x_i^{k-1}. \end{aligned} \quad (42)$$

By substituting this for all $r \in [q]$ into the right hand side of (41) and from the fact that $\mathcal{V} = \cup_{r \in [q]} \mathcal{V}_r^k$, we obtain:

$$\sum_{i=1}^n x_i^{k+1} = (1 + \beta) \sum_{i=1}^n x_i^k - \beta \sum_{i=1}^n x_i^{k-1}.$$

Since $x^0 = x^1$, we have $\sum_{i=1}^n x_i^0 = \sum_{i=1}^n x_i^1$, and as a result $\sum_{i=1}^n x_i^k = \sum_{i=1}^n x_i^0$ for all $k \geq 0$. \square

4.2 Provably accelerated randomized gossip algorithms

In the results of this subsection we focus on one specific case of the Sketch and Project framework, the RK method (5). We present two accelerated variants of the randomized Kaczmarz (RK)

where the Nesterov's momentum is used, for solving consistent linear systems and we describe their theoretical convergence results. Based on these methods we propose two provably accelerated gossip protocols, along with some remarks on their implementation.

4.2.1 Accelerated Kaczmarz methods using Nesterov's momentum

There are two different but very similar ways to provably accelerate the randomized Kaczmarz method using Nesterov's acceleration. The first paper that proves *asymptotic convergence* with an accelerated linear rate is [41]. The proof technique is similar to the framework developed by Nesterov in [60] for the acceleration of coordinate descent methods. In [82, 24] a modified version for the selection of the parameters was proposed and a *non-asymptotic* accelerated linear rate was established. In Algorithm 5, pseudocode of the Accelerated Kaczmarz method (AccRK) is presented where both variants can be cast as special cases, by choosing the parameters with the correct way.

Algorithm 5 Accelerated Randomized Kaczmarz Method (AccRK)

- 1: Data: Matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$; vector $b \in \mathbb{R}^m$
 - 2: Choose $x^0 \in \mathbb{R}^n$ and set $v^0 = x^0$
 - 3: Parameters: Evaluate the sequences of the scalars $\alpha_k, \beta_k, \gamma_k$ following one of two possible options.
 - 4: **for** $k = 0, 1, 2, \dots, K$ **do**
 - 5: $y^k = \alpha_k v^k + (1 - \alpha_k)x^k$
 - 6: Draw a fresh sample $i_k \in [m]$ with equal probability
 - 7: $x^{k+1} = y^k - \frac{\mathbf{A}_{i_k:} y^k - b_{i_k}}{\|\mathbf{A}_{i_k:}\|^2} \mathbf{A}_{i_k:}^\top$
 - 8: $v^{k+1} = \beta_k v^k + (1 - \beta_k)y^k - \gamma_k \frac{\mathbf{A}_{i_k:} y^k - b_{i_k}}{\|\mathbf{A}_{i_k:}\|^2} \mathbf{A}_{i_k:}^\top$
 - 9: **end for**
-

There are two options for selecting the parameters of the AccRK for solving consistent linear systems with normalized matrices, which we describe next.

1. From [41]: Choose $\lambda \in [0, \lambda_{\min}^+(\mathbf{A}^\top \mathbf{A})]$ and set $\gamma_{-1} = 0$. Generate the sequence $\{\gamma_k : k = 0, 1, \dots, K+1\}$ by choosing γ_k to be the largest root of

$$\gamma_k^2 - \frac{\gamma_k}{m} = (1 - \frac{\gamma_k}{\lambda} m) \gamma_{k-1}^2,$$

and generate the sequences $\{\alpha_k : k = 0, 1, \dots, K+1\}$ and $\{\beta_k : k = 0, 1, \dots, K+1\}$ by setting

$$\alpha_k = \frac{m - \gamma_k \lambda}{\gamma_k (m^2 - \lambda)}, \quad \beta_k = 1 - \frac{\gamma_k \lambda}{m}.$$

2. From [24]: Let

$$\nu = \max_{u \in \text{Range}(\mathbf{A}^\top)} \frac{u^\top [\sum_{i=1}^m \mathbf{A}_{i:}^\top \mathbf{A}_{i:} (\mathbf{A}^\top \mathbf{A})^\dagger \mathbf{A}_{i:}^\top \mathbf{A}_{i:}] u}{u^\top \frac{\mathbf{A}^\top \mathbf{A}}{m} u}. \quad (43)$$

Choose the three sequences to be fixed constants as follows: $\beta_k = \beta = 1 - \sqrt{\frac{\lambda_{\min}^+(\mathbf{W})}{\nu}}$, $\gamma_k = \gamma = \sqrt{\frac{1}{\lambda_{\min}^+(\mathbf{W})\nu}}$, $\alpha_k = \alpha = \frac{1}{1+\gamma\nu} \in (0, 1)$ where $\mathbf{W} = \frac{\mathbf{A}^\top \mathbf{A}}{m}$.

4.2.2 Theoretical guarantees of AccRK

The two variants (Option 1 and Option 2) of AccRK are closely related, however their convergence analyses are different. Below we present the theoretical guarantees of the two options as presented in [41] and [24].

Theorem 9 ([41]). Let $\{x^k\}_{k=0}^\infty$ be the sequence of random iterates produced by Algorithm 5 with the Option 1 for the parameters. Let \mathbf{A} be normalized matrix and let $\lambda \in [0, \lambda_{\min}^+(\mathbf{A}^\top \mathbf{A})]$. Set $\sigma_1 = 1 + \frac{\sqrt{\lambda}}{2m}$ and $\sigma_2 = 1 - \frac{\sqrt{\lambda}}{2m}$. Then for any $k \geq 0$ we have that:

$$\mathbb{E}[\|x^k - x^*\|^2] \leq \frac{4\lambda}{(\sigma_1^k - \sigma_2^k)^2} \|x^0 - x^*\|_{(\mathbf{A}^\top \mathbf{A})^\dagger}^2.$$

Corollary 2 ([41]). Note that as $k \rightarrow \infty$, we have that $\sigma_2^k \rightarrow 0$. This means that the decrease of the right hand side is governed mainly by the behavior of the term σ_1 in the denominator and as a result the method converge asymptotically with a decrease factor per iteration: $\sigma_1^{-2} = (1 + \frac{\sqrt{\lambda}}{2m})^{-2} \approx 1 - \frac{\sqrt{\lambda}}{m}$. That is, as $k \rightarrow \infty$:

$$\mathbb{E}[\|x^k - x^*\|^2] \leq \left(1 - \sqrt{\lambda}/m\right)^k 4\lambda \|x^0 - x^*\|_{(\mathbf{A}^\top \mathbf{A})^\dagger}^2$$

Thus, by choosing $\lambda = \lambda_{\min}^+$ and for the case that λ_{\min}^+ is small, Algorithm 5 will have significantly faster convergence rate than RK. Note that the above convergence results hold only for normalized matrices $\mathbf{A} \in \mathbb{R}^{m \times n}$, that is matrices that have $\|\mathbf{A}_{i:}\| = 1$ for any $i \in m$.

Using Corollary 2, Algorithm 5 with the first choice of the parameters converges linearly with rate $\left(1 - \sqrt{\lambda}/m\right)$. That is, it requires $O\left(m/\sqrt{\lambda} \log(1/\epsilon)\right)$ iterations to obtain accuracy $\mathbb{E}[\|x^k - x^*\|^2] \leq \epsilon 4\lambda \|x^0 - x^*\|_{(\mathbf{A}^\top \mathbf{A})^\dagger}^2$.

Theorem 10 ([24]). Let $\mathbf{W} = \frac{\mathbf{A}^\top \mathbf{A}}{m}$ and let assume exactness⁷. Let $\{x^k, y^k, v^k\}$ be the iterates of Algorithm 5 with the Option 2 for the parameters. Then

$$\Psi^k \leq \left(1 - \sqrt{\lambda_{\min}^+(\mathbf{W})/\nu}\right)^k \Psi^0,$$

where $\Psi^k = \mathbb{E} \left[\|v^k - x^*\|_{\mathbf{W}^\dagger}^2 + \frac{1}{\mu} \|x^k - x^*\|^2 \right]$.

The above result implies that Algorithm 5 converges linearly with rate $1 - \sqrt{\lambda_{\min}^+(\mathbf{W})/\nu}$, which translates to a total of $O\left(\sqrt{\nu/\lambda_{\min}^+(\mathbf{W})} \log(1/\epsilon)\right)$ iterations to bring the quantity Ψ^k below $\epsilon > 0$. It can be shown that $1 \leq \nu \leq 1/\lambda_{\min}^+(\mathbf{W})$, (Lemma 2 in [24]) where ν is as defined in (43). Thus, $\sqrt{\frac{1}{\lambda_{\min}^+(\mathbf{W})}} \leq \sqrt{\frac{\nu}{\lambda_{\min}^+(\mathbf{W})}} \leq \frac{1}{\lambda_{\min}^+(\mathbf{W})}$, which means that the rate of AccRK (Option 2) is always better than that of the RK with unit stepsize which is equal to $O\left(\frac{1}{\lambda_{\min}^+(\mathbf{W})} \log(1/\epsilon)\right)$ (see Theorem 1).

In [24], Theorem 10 has been proposed for solving more general consistent linear systems (the matrix \mathbf{A} of the system is not assumed to be normalized). In this case $\mathbf{W} = \mathbb{E}[\mathbf{Z}]$ and the parameter ν is slightly more complicated than the one of equation (43). We refer the interested reader to [24] for more details.

Comparison of the convergence rates: Before describe the distributed nature of the AccRK and explain how it can be interpreted as a gossip algorithm, let us compare the convergence rates of the two options of the parameters for the case of general normalized consistent linear systems ($\|\mathbf{A}_{i:}\| = 1$ for any $i \in [m]$).

Using Theorems 9 and 10, it is clear that the iteration complexity of AccRK is

$$O\left(\frac{m}{\sqrt{\lambda}} \log(1/\epsilon)\right)^{\lambda=\lambda_{\min}^+(\mathbf{A}^\top \mathbf{A})} O\left(\frac{m}{\sqrt{\lambda_{\min}^+(\mathbf{A}^\top \mathbf{A})}} \log(1/\epsilon)\right), \quad (44)$$

⁷Note that in this setting $\mathbf{B} = \mathbf{I}$, which means that $\mathbf{W} = \mathbb{E}[\mathbf{Z}]$, and the exactness assumption takes the form $\text{Null}(\mathbf{W}) = \text{Null}(\mathbf{A})$.

and

$$O\left(\sqrt{\frac{\nu m}{\lambda_{\min}^+(\mathbf{A}^\top \mathbf{A})}} \log(1/\epsilon)\right), \quad (45)$$

for the Option 1 and Option 2 for the parameters, respectively.

In the following derivation we compare the iteration complexity of the two methods.

Lemma 11. *Let matrices $\mathbf{C} \in \mathbb{R}^{n \times n}$ and $\mathbf{C}_i \in \mathbb{R}^{n \times n}$ where $i \in [m]$ be positive semidefinite, and satisfying $\sum_{i=1}^m \mathbf{C}_i = \mathbf{C}$. Then*

$$\sum_{i=1}^m \mathbf{C}_i \mathbf{C}^\dagger \mathbf{C}_i \preceq \mathbf{C}.$$

Proof. From the definition of the matrices it holds that $\mathbf{C}_i \preceq \mathbf{C}$ for any $i \in [m]$. Using the properties of Moore-Penrose pseudoinverse, this implies that

$$\mathbf{C}_i^\dagger \succeq \mathbf{C}^\dagger. \quad (46)$$

Therefore

$$\mathbf{C}_i = \mathbf{C}_i \mathbf{C}_i^\dagger \mathbf{C}_i \stackrel{(46)}{\succeq} \mathbf{C}_i \mathbf{C}^\dagger \mathbf{C}_i. \quad (47)$$

From the definition of the matrices by taking the sum over all $i \in [m]$ we obtain:

$$\mathbf{C} = \sum_{i=1}^m \mathbf{C}_i \stackrel{(47)}{\succeq} \sum_{i=1}^m \mathbf{C}_i \mathbf{C}^\dagger \mathbf{C}_i,$$

which completes the proof. \square

Let us now choose $\mathbf{C}_i = \mathbf{A}_{i:}^\top \mathbf{A}_{i:}$ and $\mathbf{C} = \mathbf{A}^\top \mathbf{A}$. Note that from their definition the matrices are positive semidefinite and satisfy $\sum_{i=1}^m \mathbf{A}_{i:}^\top \mathbf{A}_{i:} = \mathbf{A}^\top \mathbf{A}$. Using Lemma 11 it is clear that:

$$\sum_{i=1}^m \mathbf{A}_{i:}^\top \mathbf{A}_{i:} (\mathbf{A}^\top \mathbf{A})^\dagger \mathbf{A}_{i:}^\top \mathbf{A}_{i:} \preceq \mathbf{A}^\top \mathbf{A},$$

or in other words, for any vector $v \notin \text{Null}(\mathbf{A})$ we set the inequality

$$\frac{v^\top \left[\sum_{i=1}^m \mathbf{A}_{i:}^\top \mathbf{A}_{i:} (\mathbf{A}^\top \mathbf{A})^\dagger \mathbf{A}_{i:}^\top \mathbf{A}_{i:} \right] v}{v^\top [\mathbf{A}^\top \mathbf{A}] v} \leq 1.$$

Multiplying both sides by m , we set:

$$\frac{v^\top \left[\sum_{i=1}^m \mathbf{A}_{i:}^\top \mathbf{A}_{i:} (\mathbf{A}^\top \mathbf{A})^\dagger \mathbf{A}_{i:}^\top \mathbf{A}_{i:} \right] v}{v^\top \left[\frac{\mathbf{A}^\top \mathbf{A}}{m} \right] v} \leq m.$$

Using the above derivation, it is clear from the definition of the parameter ν (43), that $\nu \leq m$. By combining our finding with the bounds already obtained in [24] for the parameter ν , we have that:

$$1 \leq \nu \leq \min \left\{ m, \frac{1}{\lambda_{\min}^+(\mathbf{W})} \right\}. \quad (48)$$

Thus, by comparing the two iteration complexities of equations (44) and (45) it is clear that Option 2 for the parameters [24] is always faster in theory than Option 1 [41]. To the best of our knowledge, such comparison of the two choices of the parameters for the AccRK was never presented before.

4.2.3 Accelerated randomized gossip algorithms

Having presented the complexity analysis guarantees of AccRK for solving consistent linear systems with normalized matrices, let us now explain how the two options of AccRK behave as gossip algorithms when they are used to solve the linear system $\mathbf{A}x = 0$ where $\mathbf{A} \in \mathbb{R}^{|\mathcal{E}| \times n}$ is the normalized incidence matrix of the network. That is, each row $e = (i, j)$ of \mathbf{A} can be represented as $(\mathbf{A}_{e:})^\top = \frac{1}{\sqrt{2}}(e_i - e_j)$ where e_i (resp. e_j) is the i^{th} (resp. j^{th}) unit coordinate vector in \mathbb{R}^n .

By using this particular linear system, the expression $\frac{\mathbf{A}_{i:}y^k - b_i}{\|\mathbf{A}_{i:}\|^2} \mathbf{A}_{i:}^\top$ that appears in steps 7 and 8 of AccRK takes the following form when the row $e = (i, j) \in \mathcal{E}$ is sampled:

$$\frac{\mathbf{A}_{e:}y^k - b_i}{\|\mathbf{A}_{e:}\|^2} \mathbf{A}_{e:}^\top \stackrel{b=0}{=} \frac{\mathbf{A}_{e:}y^k}{\|\mathbf{A}_{e:}\|^2} \mathbf{A}_{e:}^\top \stackrel{\text{form of A}}{=} \frac{y_i^k - y_j^k}{2}(e_i - e_j).$$

Recall that with \mathbf{L} we denote the Laplacian matrix of the network. For solving the above AC system (see Definition 2), the standard RK requires $O\left(\left(\frac{2m}{\lambda_{\min}^+(\mathbf{L})}\right) \log(1/\epsilon)\right)$ iterations to achieve expected accuracy $\epsilon > 0$. To understand the acceleration in the gossip framework this should be compared to the

$$O\left(m \sqrt{\frac{2}{\lambda_{\min}^+(\mathbf{L})}} \log(1/\epsilon)\right)$$

of AccRK (Option 1) and the

$$O\left(\sqrt{\frac{2m\nu}{\lambda_{\min}^+(\mathbf{L})}} \log(1/\epsilon)\right)$$

of AccRK (Option 2).

Algorithm 6 describes in a single framework how the two variants of AccRK of Section 4.2.1 behave as gossip algorithms when are used to solve the above linear system. Note that each node $\ell \in \mathcal{V}$ of the network has two local registers to save the quantities v_ℓ^k and x_ℓ^k . In each step using these two values every node $\ell \in \mathcal{V}$ of the network (activated or not) computes the quantity $y_\ell^k = \alpha_k v_\ell^k + (1 - \alpha_k)x_\ell^k$. Then in the k^{th} iteration the activated nodes i and j of the randomly selected edge $e = (i, j)$ exchange their values y_i^k and y_j^k and update the values of x_i^k, x_j^k and v_i^k, v_j^k as shown in Algorithm 6. The rest of the nodes use only their own y_ℓ^k to update the values of v_i^k and x_i^k without communicate with any other node.

The parameter $\lambda_{\min}^+(\mathbf{L})$ can be estimated by all nodes in a decentralized manner using the method described in [8]. In order to implement this algorithm, we assume that all nodes have synchronized clocks and that they know the rate at which gossip updates are performed, so that inactive nodes also update their local values. This may not be feasible in all applications, but when it is possible (e.g., if nodes are equipped with inexpensive GPS receivers, or have reliable clocks) then they can benefit from the significant speedup achieved.

5 Dual Randomized Gossip Algorithms

An important tool in optimization literature is duality. In our setting, instead of solving the original minimization problem (primal problem) one may try to develop dual in nature methods that have as a goal to directly solve the dual maximization problem. Then the primal solution can be recovered through the use of optimality conditions and the development of an affine mapping between the two spaces (primal and dual).

Like in most optimization problems (see [54, 75, 70, 17]), dual methods have been also developed in the literature of randomized algorithms for solving linear systems [86, 26, 48]. In this section, using existing dual methods and the connection already established between the two areas of research (methods for linear systems and gossip algorithms), we present a different viewpoint that allows the development of novel dual randomized gossip algorithms.

Without loss of generality we focus on the case of $\mathbf{B} = \mathbf{I}$ (no weighted average consensus). For simplicity, we formulate the AC system as the one with the incidence matrix of the network

Algorithm 6 Accelerated Randomized Gossip Algorithm (AccGossip)

- 1: **Data:** Matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ (normalized incidence matrix); vector $b = 0 \in \mathbb{R}^m$
- 2: Choose $x^0 \in \mathbb{R}^n$ and set $v^0 = x^0$
- 3: **Parameters:** Evaluate the sequences of the scalars $\alpha_k, \beta_k, \gamma_k$ following one of two possible options.
- 4: **for** $k = 0, 1, 2, \dots, K$ **do**
- 5: Each node $\ell \in \mathcal{V}$ evaluate $y_\ell^k = \alpha_k v_\ell^k + (1 - \alpha_k)x_\ell^k$.
- 6: Pick an edge $e = (i, j)$ uniformly at random.
- 7: Then the nodes update their values as follows:
 - The selected node i and node j :

$$x_i^{k+1} = x_j^{k+1} = (y_i^k + y_j^k)/2$$

$$v_i^{k+1} = \beta_k v_i^k + (1 - \beta_k)y_i^k - \gamma_k(y_i^k - y_j^k)/2$$

$$v_j^{k+1} = \beta_k v_j^k + (1 - \beta_k)y_j^k - \gamma_k(y_j^k - y_i^k)/2$$

- Any other node $\ell \in \mathcal{V}$:

$$x_\ell^{k+1} = y_\ell^k, \quad v_\ell^{k+1} = \beta_k v_\ell^k + (1 - \beta_k)y_\ell^k$$

8: **end for**

($\mathbf{A} = \mathbf{Q}$) and focus on presenting the distributed nature of dual randomized gossip algorithms with no momentum. While we focus only on no-momentum protocols, we note that accelerated variants of the dual methods could be easily obtained using tools from Section 4.

5.1 Dual problem and SDSA

The Lagrangian dual of the best approximation problem (7) is the (bounded) unconstrained concave quadratic maximization problem [26]:

$$\max_{y \in \mathbb{R}^m} D(y) := (b - \mathbf{A}x^0)^\top y - \frac{1}{2} \|\mathbf{A}^\top y\|_{\mathbf{B}^{-1}}^2. \quad (49)$$

A direct method for solving the dual problem – Stochastic Dual Subspace Accent (SDSA)– was first proposed in [26]. SDSA is a randomized iterative algorithm for solving (49), which updates the dual vectors y^k as follows:

$$y^{k+1} = y^k + \omega \mathbf{S}_k \lambda^k,$$

where \mathbf{S}_k is a matrix chosen in an i.i.d. fashion throughout the iterative process from an arbitrary but fixed distribution (which is a parameter of the method) and λ^k is a vector chosen *afterwards* so that $D(y^k + \mathbf{S}_k \lambda)$ is maximized in λ . In particular, SDSA proceeds by moving in random subspaces spanned by the random columns of \mathbf{S}_k .

In general, the maximizer in λ is not unique. In SDSA, λ^k is chosen to be the least-norm maximizer, which leads to the iteration

$$y^{k+1} = y^k + \omega \mathbf{S}_k \left(\mathbf{S}_k^\top \mathbf{A} \mathbf{B}^{-1} \mathbf{A}^\top \mathbf{S}_k \right)^\dagger \mathbf{S}_k^\top \left(b - \mathbf{A}(x^0 + \mathbf{B}^{-1} \mathbf{A}^\top y^k) \right). \quad (50)$$

It can be shown that the iterates $\{x^k\}_{k \geq 0}$ of the sketch and project method (Algorithm 1) can be arised as affine images of the iterates $\{y^k\}_{k \geq 0}$ of the dual method (50) as follows [26, 48]:

$$x^k = x(y^k) = x^0 + \mathbf{B}^{-1} \mathbf{A}^\top y^k. \quad (51)$$

In [26] the dual method was analyzed for the case of unit stepsize ($\omega = 1$). Later in [48] the analysis extended to capture the cases of $\omega \in (0, 2)$.

It can be shown,[26, 48], that if \mathbf{S}_k is chosen randomly from the set of unit coordinate/basis vectors in \mathbb{R}^m , then the dual method (50) is randomized coordinate descent [38, 72], and the corresponding primal method is RK (5). More generally, if \mathbf{S}_k is a random column submatrix of the $m \times m$ identity matrix, the dual method is the randomized Newton method [69], and the corresponding primal method is RBK (6). In Section 5.2 we shall describe the more general block case in more detail.

The basic convergence result for the dual iterative process (50) is presented in the following theorem. We set $y^0 = 0$ so that $x^0 = c$ and the affine mapping of equation (51) is satisfied. Note that, SDSA for solving the dual problem (49) and sketch and project method for solving the best approximation problem (7) (primal problem) converge exactly with the same convergence rate.

Theorem 12 (Complexity of SDSA [26, 48]). *Assume exactness and let $y^0 = 0$. Let $\{y^k\}$ be the iterates produced by SDSA (equation (50)) with step-size $\omega \in (0, 2)$. Then,*

$$\mathbb{E}[D(y^*) - D(y^k)] \leq \rho^k [D(y^*) - D(y^0)], \quad (52)$$

where $\rho := 1 - \omega(2 - \omega)\lambda_{\min}^+(\mathbf{W}) \in (0, 1)$.

5.2 Randomized Newton method as a dual gossip algorithm

In this subsection we bring a new insight into the randomized gossip framework by presenting how the dual iterative process that is associated to RBK method solves the AC problem with $\mathbf{A} = \mathbf{Q}$ (incidence matrix). Recall that the right hand side of the linear system is $b = 0$. For simplicity, we focus on the case of $\mathbf{B} = \mathbf{I}$ and $\omega = 1$.

Under this setting ($\mathbf{A} = \mathbf{Q}$, $\mathbf{B} = \mathbf{I}$ and $\omega = 1$) the dual iterative process (50) takes the form:

$$y^{k+1} = y^k - \mathbf{I}_{C:}(\mathbf{I}_{C:}^\top \mathbf{Q} \mathbf{Q}^\top \mathbf{I}_{C:})^\dagger \mathbf{Q}(x^0 + \mathbf{Q}^\top y^k), \quad (53)$$

and from Theorem 12 converges to a solution of the dual problem as follows:

$$\mathbb{E}[D(y^*) - D(y^k)] \leq \left[1 - \lambda_{\min}^+ \left(\mathbb{E} \left[\mathbf{Q}_{C:}^\top (\mathbf{Q}_C \mathbf{Q}_C^\top)^\dagger \mathbf{Q}_{C:}\right]\right)\right]^k [D(y^*) - D(y^0)].$$

Note that the convergence rate is exactly the same with the rate of the RBK under the same assumptions (see (30)).

This algorithm is a randomized variant of the Newton method applied to the problem of maximizing the quadratic function $D(y)$ defined in (49). Indeed, in each iteration we perform the update $y^{k+1} = y^k + \mathbf{I}_{C:}\lambda^k$, where λ^k is chosen greedily so that $D(y^{k+1})$ is maximized. In doing so, we invert a random principal submatrix of the Hessian of D , whence the name.

Randomized Newton Method (RNM) was first proposed by Qu et al. [69]. RNM was first analyzed as an algorithm for minimizing *smooth strongly convex functions*. In [26] it was also extended to the case of a *smooth but weakly convex quadratics*. This method was not previously associated with any gossip algorithm.

The most important distinction of RNM compared to existing gossip algorithms is that it operates with values that are associated to the *edges* of the network. To the best of our knowledge, it is the first *randomized dual gossip method*⁸. In particular, instead of iterating over values stored at the nodes, RNM uses these values to update “dual weights” $y^k \in \mathbb{R}^m$ that correspond to the edges \mathcal{E} of the network. However, deterministic dual distributed averaging algorithms were proposed before [71, 20]. Edge-based methods have also been proposed before; in particular in [84] an asynchronous distributed ADMM algorithm presented for solving the more general consensus optimization problem with convex functions.

Natural Interpretation. In iteration k , *RNM* (Algorithm (53)) executes the following steps: 1) Select a random set of edges $\mathcal{S}_k \subseteq \mathcal{E}$, 2) Form a subgraph \mathcal{G}_k of \mathcal{G} from the selected edges, 3)

⁸Since our first paper [46] appeared online, papers [29] and [32] propose dual randomized gossip algorithms as well. In [29] the authors focus on presenting privacy preserving dual randomized gossip algorithms and [32] solve the dual problem (49) using accelerated coordinate descent.

The values of the edges in each connected component of \mathcal{G}_k are updated: their new values are a linear combination of the private values of the nodes belonging to the connected component and of the adjacent edges of their connected components. (see also example of Figure 4).

Dual Variables as Advice. The weights y^k of the edges have a natural interpretation as *advice* that each selected node receives from the network in order to update its value (to one that will eventually converge to the desired average).

Consider RNM performing the k^{th} iteration and let \mathcal{V}_r denote the set of nodes of the selected connected component that node i belongs to. Then, from Theorem 5 we know that $x_i^{k+1} = \sum_{i \in \mathcal{V}_r} x_i^k / |\mathcal{V}_r|$. Hence, by using (51), we obtain the following identity:

$$(\mathbf{A}^\top y^{k+1})_i = \frac{1}{|\mathcal{V}_r|} \sum_{i \in \mathcal{V}_r} (c_i + (\mathbf{A}^\top y^k)_i) - c_i. \quad (54)$$

Thus in each step $(\mathbf{A}^\top y^{k+1})_i$ represents the term (advice) that must be added to the initial value c_i of node i in order to update its value to the average of the values of the nodes of the connected component i belongs to.

Importance of the dual perspective: It was shown in [69] that when RNM (and as a result, RBK, through the affine mapping (51)) is viewed as a family of methods indexed by the size $\tau = |\mathcal{S}|$ (we choose \mathcal{S} of fixed size in the experiments), then $\tau \rightarrow 1/(1 - \rho)$, where ρ is defined in (10), decreases *superlinearly* fast in τ . That is, as τ increases by some factor, the iteration complexity drops by a factor that is at least as large. Through preliminary numerical experiments in Section 7.2 we experimentally show that this is true for the case of AC systems as well.

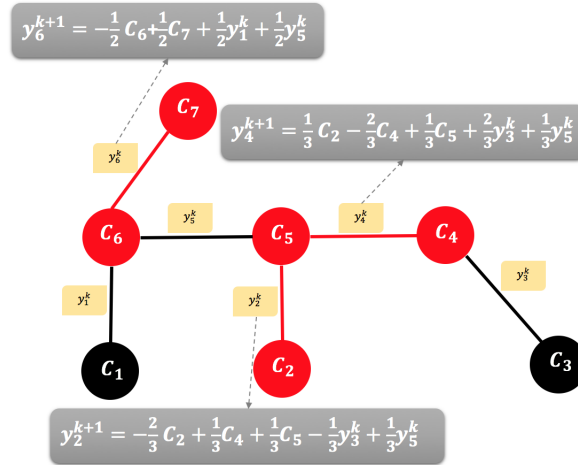


Figure 4: Example of how the RNM method works as gossip algorithm. In this specific case 3 edges are selected and form a sub-graph with two connected components. Then the values at the edges update their values using the private values of the nodes belonging to their connected component and the values associate to the adjacent edges of their connected components.

6 Further Connections Between Methods for Solving Linear Systems and Gossip Algorithms

In this section we highlight some further interesting connections between linear systems solvers and gossip protocols for average consensus:

- **Eavesdrop gossip as special case of Kaczmarz-Motzkin method.** In [83] greedy gossip with eavesdropping (GGE), a novel randomized gossip algorithm for distributed computation of the average consensus problem was proposed and analyzed. In particular it was shown that that greedy updates of GGE lead to rapid convergence. In this protocol, the greedy updates are made possible by exploiting the broadcast nature of wireless communications. During the operation of GGE, when a node decides to gossip, instead of choosing one of its neighbors at

random, it makes a greedy selection, choosing the node which has the value most different from its own. In particular the method behaves as follows:

At the k^{th} iteration of GGE, a node i_k is chosen uniformly at random from $[n]$. Then, i_k identifies a neighboring node $j_k \in \mathcal{N}_i$ satisfying:

$$j_k \in \max_{j \in \mathcal{N}_i} \left\{ \frac{1}{2} (x_i^k - x_j^k)^2 \right\}$$

which means that the selected node i_k identifies a neighbor that currently has the most different value from its own. This choice is possible because each node $i \in \mathcal{V}$ maintains not only its own local variable x_i^k , but also a copy of the current values at its neighbors x_j^k for $j \in \mathcal{N}_i$. In the case that node i_k has multiple neighbors whose values are all equally (and maximally) different from its current value, it chooses one of these neighbors at random. Then node i_k and j_k update their values to:

$$x_i^{k+1} = x_j^{k+1} = \frac{1}{2} (x_i^k + x_j^k).$$

In the area of randomized methods for solving large linear system there is one particular method, the Kaczmarz-Motzkin algorithm [12, 28] that can work as gossip algorithm with the same update as the GGE when is use to solve the homogeneous linear system with matrix the Incidence matrix of the network.

Update rule of Kaczmarz-Motzkin algorithm (KMA) [12, 28]:

1. Choose sample of d_k constraints, P_k , uniformly at random from among the rows of matrix \mathbf{A} .
2. From among these d_k constraints, choose $t_k = \operatorname{argmax}_{i \in P_k} \mathbf{A}_{i,:} x^k - b_i$.
3. Update the value: $x^{k+1} = x^k - \frac{\mathbf{A}_{t_k,:} x^k - b_{t_k}}{\|\mathbf{A}_{t_k,:}\|^2} \mathbf{A}_{t_k,:}^\top$.

It is easy to verify that when the Kaczmarz-Motzkin algorithm is used for solving the AC system with $\mathbf{A} = \mathbf{Q}$ (incidence matrix) and in each step of the method the chosen constraints d_k of the linear system correspond to edges attached to one node it behaves exactly like the GGE. From numerical analysis viewpoint an easy way to choose the constraints d_k that are compatible to the desired edges is in each iteration to find the indexes of the non-zeros of a uniformly at random selected column (node) and then select the rows corresponding to these indexes.

Therefore, since GGE [83] is a special case of the KMA (when the later applied to special AC system with Incidence matrix) it means that we can obtain the convergence rate of GGE by simply use the tight conergence analysis presented in [12, 28]⁹. In [83] it was mentioned that analyzing the convergence behavior of GGE is non-trivial and not an easy task. By establishing the above connection the convergence rates of GGE can be easily obtained as special case of the theorems presented in [12].

In Section 4 we presented provably accelerated variants of the pairwise gossip algorithm and of its block variant. Following the same approach one can easily develop accelerated variants of the GGE using the recently proposed analysis for the accelerated Kaczmarz-Motzkin algorithm presented in [52].

• Inexact Sketch and Project Methods:

Recently in [50], several inexact variants of the sketch and project method (1) have been proposed. As we have already mentioned the sketch and project method is a two step procedure algorithm where first the sketched system is formulated and then the last iterate x^k is *exactly*

⁹Note that the convergence theorems of [12, 28] use $d_k = d$. However, with a small modification in the original proof the theorem can capture the case of different d_k .

projected into the solution set of the sketched system. In [50] the authors replace the exact projection with an inexact variant and suggested to run a different algorithm (this can be the sketch and project method itself) in the sketched system to obtain an approximate solution. It was shown that in terms of time the inexact updates can be faster than their exact variants.

In the setting of randomized gossip algorithms for the AC system with Incidence matrix ($\mathbf{A} = \mathbf{Q}$), $\mathbf{B} = \mathbf{I}$ and $\omega = 1$ a variant of the inexact sketch and project method will work as follows (similar to the update proved in Theorem 5):

1. Select a random set of edges $C \subseteq \mathcal{E}$.
2. Form subgraph \mathcal{G}_k of \mathcal{G} from the selected edges.
3. Run the pairwise gossip algorithm of [6] (or any variant of the sketch and project method) on the subgraph \mathcal{G}_k until an accuracy ϵ is achieved (reach a neighborhood of the exact average).

- **Non-randomized gossip algorithms as special cases of Kaczmarz methods:**

In the gossip algorithms literature there are efficient protocols that are not randomized [53, 31, 42, 89]. Typically, in these algorithms the pairwise exchanges between nodes it happens in a deterministic, such as predefined cyclic, order. For example, T -periodic gossiping is a protocol which stipulates that each node must interact with each of its neighbours exactly once every T time units. It was shown that under suitable connectivity assumptions of the network \mathcal{G} , the T -periodic gossip sequence will converge at a rate determined by the magnitude of the second largest eigenvalue of the stochastic matrix determined by the sequence of pairwise exchanges which occurs over a period. It has been shown that if the underlying graph is a tree, the mentioned eigenvalue is constant for all possible T -periodic gossip protocols.

In this work we focus only on randomized gossip protocols. However we speculate that the above non-randomized gossip algorithms would be able to express as special cases of popular non-randomized projection methods for solving linear systems [68, 62, 16]. Establishing connections like that is an interesting future direction of research and can possibly lead to the development of novel block and accelerated variants of many non-randomized gossip algorithms, similar to the protocols we present in Sections 3 and 4.

7 Numerical Evaluation

In this section, we empirically validate our theoretical results and evaluate the performance of the proposed randomized gossip algorithms. The section is divided into four main parts, in each of which we highlight a different aspect of our contributions.

In the first experiment, we numerically verify the linear convergence of the Scaled RK algorithm (see equation (14)) for solving the weighted average consensus problem presented in Section 3.1. In the second part, we explain the benefit of using block variants in the gossip protocols where more than two nodes update their values in each iteration (protocols presented in Section 3.4). In the third part, we explore the performance of the faster and provably accelerated gossip algorithms proposed in Section 4. In the last experiment, we numerically show that relaxed variants of the pairwise randomized gossip algorithm converge faster than the standard randomized pairwise gossip with unit stepsize (no relaxation). This gives a specific setting where the phenomenon of over-relaxation of iterative methods for solving linear systems is beneficial.

In the comparison of all gossip algorithms we use the relative error measure $\|x^k - x^*\|_{\mathbf{B}}^2 / \|x^0 - x^*\|_{\mathbf{B}}^2$ where $x^0 = c \in \mathbb{R}^n$ is the starting vector of the values of the nodes and matrix \mathbf{B} is the positive definite diagonal matrix with weights in its diagonal (recall that in the case of standard average consensus this can be simply $\mathbf{B} = \mathbf{I}$). Depending on the experiment, we choose the values of the starting vector $c \in \mathbb{R}^n$ to follow either a Gaussian distribution or uniform distribution or to be integer values such that $c_i = i \in \mathbb{R}$. In the plots, the horizontal axis represents the number of iterations except in the figures of subsection 7.2, where the horizontal axis represents the block size.

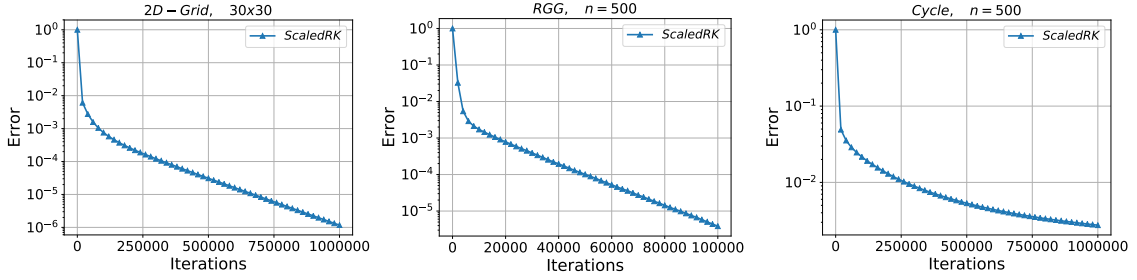


Figure 5: Performance of ScaledRK in a 2-dimension grid, random geometric graph (RGG) and a cycle graph for solving the weighted average consensus problem. The weight matrix is chosen to be $\mathbf{B} = \mathbf{D}$, the degree matrix of the network. The n in the title of each plot indicates the number of nodes of the network. For the grid graph this is $n \times n$.

In our implementations we use three popular graph topologies from the area of wireless sensor networks. These are the cycle (ring graph), the 2-dimension grid and the random geometric graph (RGG) with radius $r = \sqrt{\log(n)/n}$. In all experiments we formulate the average consensus problem (or its weighted variant) using the incidence matrix. That is, $\mathbf{A} = \mathbf{Q}$ is used as the AC system. Code was written in Julia 0.6.3.

7.1 Convergence on weighted average consensus

As we explained in Section 3, the sketch and project method (Algorithm 1) can solve the more general weighted AC problem. In this first experiment we numerically verify the linear convergence of the Scaled RK algorithm (14) for solving this problem in the case of $\mathbf{B} = \mathbf{D}$. That is, the matrix \mathbf{B} of the weights is the degree matrix \mathbf{D} of the graph ($\mathbf{B}_{ii} = d_i, \forall i \in [n]$). In this setting the exact update rule of the method is given in equation (17), where in order to have convergence to the weighted average the chosen nodes are required to share not only their private values but also their weight \mathbf{B}_{ii} (in our experiment this is equal to the degree of the node d_i). In this experiment the starting vector of values $x^0 = c \in \mathbb{R}^n$ is a Gaussian vector. The linear convergence of the algorithm is clear in Figure 5.

7.2 Benefit of block variants

We devote this experiment to evaluate the performance of the randomized block gossip algorithms presented in Sections 3.4 and 5. In particular, we would like to highlight the benefit of using larger block size in the update rule of randomized Kaczmarz method and as a result through our established connection of the randomized pairwise gossip algorithm [6] (see equation (18)).

Recall that in Section 5 we show that both RBK and RNM converge to the solution of the primal and dual problems respectively with the same rate and that their iterates are related via a simple affine transform (51). In addition note that an interesting feature of the RNM [69], is that when the method viewed as algorithm indexed by the size $\tau = |C|$, it enjoys superlinear speedup in τ . That is, as τ (block size) increases by some factor, the iteration complexity drops by a factor that is at least as large (see Section 5.2). Since RBK and RNM share the same rates this property naturally holds for RBK as well.

We show that for a connected network \mathcal{G} , the complexity improves superlinearly in $\tau = |C|$, where C is chosen as a subset of \mathcal{E} of size τ , uniformly at random (recall the in the update rule of RBK the random matrix is $\mathbf{S} = \mathbf{I}_C$). Similar to the rest of this section in comparing the number of iterations for different values of τ , we use the relative error $\varepsilon = \|x^k - x^*\|^2 / \|x^0 - x^*\|^2$. We let $x_i^0 = c_i = i$ for each node $i \in \mathcal{V}$ (vector of integers). We run RBK until the relative error becomes smaller than 0.01. The blue solid line in the figures denotes the actual number of iterations (after running the code) needed in order to achieve $\varepsilon \leq 10^{-2}$ for different values of τ . The green dotted line represents the function $f(\tau) := \frac{\ell}{\tau}$, where ℓ is the number of iterations of RBK with $\tau = 1$ (i.e., the pairwise gossip algorithm). The green line depicts linear speedup; the fact that the blue line (obtained through experiments) is below the green line points to superlinear speedup. In this

experiment we use the Cycle graph with $n = 30$ and $n = 100$ nodes (Figure 6) and the 4×4 two dimension grid graph (Figure 7). Note that, when $|C| = m$ the convergence rate of the method becomes $\rho = 0$ and as a result it converges in one step.

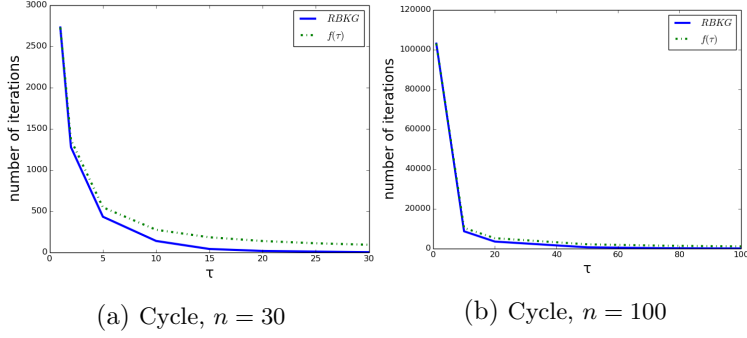


Figure 6: Superlinear speedup of RBK on cycle graphs.

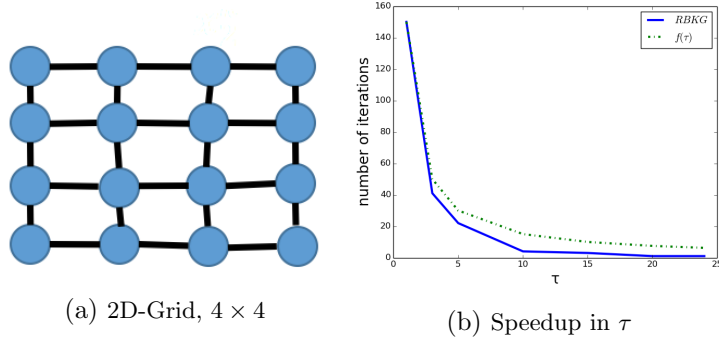


Figure 7: Superlinear speedup of RBK on a 4×4 two dimension grid graph.

7.3 Accelerated gossip algorithms

We devote this subsection to experimentally evaluate the performance of the proposed accelerated gossip algorithms: mRK (Algorithm 3), mRBK (Algorithm 4) and AccGossip with the two options of the parameters (Algorithm 6). In particular we perform four experiments. In the first two we focus on the performance of the mRK and how the choice of stepsize (relaxation parameter) ω and heavy ball momentum parameter β affect the performance of the method. In the next experiment we show that the addition of heavy ball momentum can be also beneficial for the performance of the block variant mRBK. In the last experiment we compare the standard pairwise gossip algorithm (baseline method) from [6], the mRK and the AccGossip and show that the probably accelerated gossip algorithm, AccGossip outperforms the other algorithms and converge as predicted from the theory with an accelerated linear rate.

7.3.1 Impact of momentum parameter on mRK

As we have already presented in the standard pairwise gossip algorithm (equation (18)) the two selected nodes that exchange information update their values to their exact average while all the other nodes remain idle. In our framework this update can be cast as special case of mRK when $\beta = 0$ and $\omega = 1$.

In this experiment we keep the stepsize fixed and equal to $\omega = 1$ which means that the pair of the chosen nodes update their values to their exact average and we show that by choosing a suitable momentum parameter $\beta \in (0, 1)$ we can obtain faster convergence to the consensus for all networks under study. The momentum parameter β is chosen following the suggestions made in [48] for solving general consistent linear systems. See Figure 8 for more details. It is worth to point

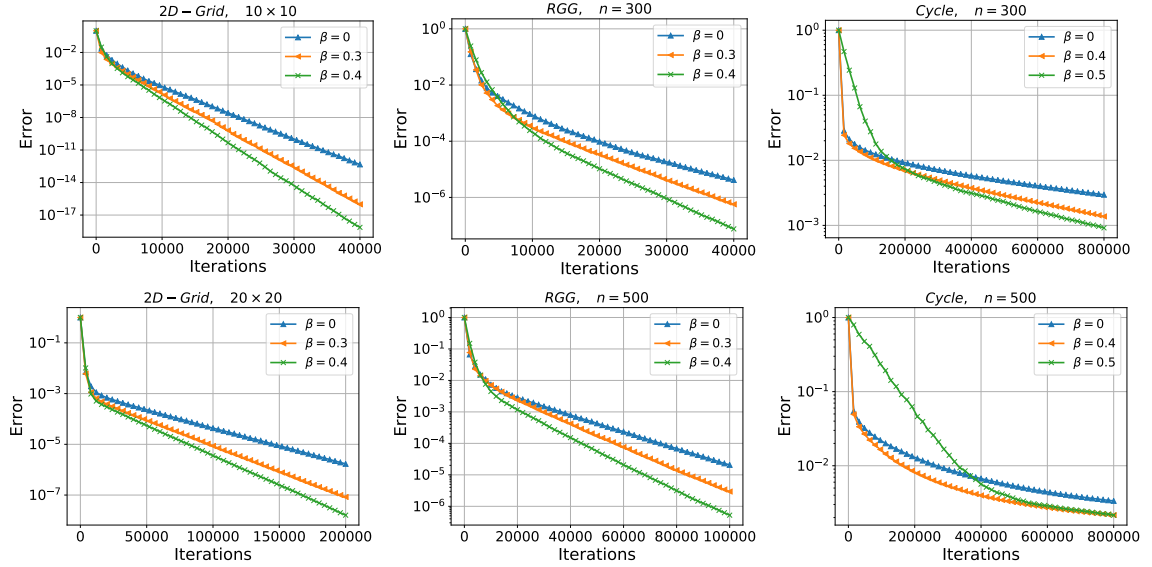


Figure 8: Performance of mRK for fixed step-size $\omega = 1$ and several momentum parameters β in a 2-dimension grid, random geometric graph (RGG) and a cycle graph. The choice $\beta = 0$ corresponds to the randomized pairwise gossip algorithm proposed in [6]. The starting vector $x^0 = c \in \mathbb{R}^n$ is a Gaussian vector. The n in the title of each plot indicates the number of nodes of the network. For the grid graph this is $n \times n$.

out that for all networks under study the addition of a heavy ball momentum term is beneficial in the performance of the method.

7.3.2 Comparison of mRK and shift-Register algorithm [40]

In this experiment we compare mRK with the shift register gossip algorithm (pairwise momentum method, abbreviation: Pmom) analyzed in [40]. We choose the parameters ω and β of mRK in such a way in order to satisfy the connection established in Section 4.1.3. That is, we choose $\beta = \omega - 1$ for any choice of $\omega \in (1, 2)$. Observe that in all plots of Figure 9 mRK outperforms the corresponding shift-register algorithm.

7.3.3 Impact of momentum parameter on mRBK

In this experiment our goal is to show that the addition of heavy ball momentum accelerates the RBK gossip algorithm presented in Section 3.4. Without loss of generality we choose the block size to be equal to $\tau = 5$. That is, the random matrix $\mathbf{S}_k \sim \mathcal{D}$ in the update rule of mRBK is a $m \times 5$ column submatrix of the identity $m \times m$ matrix. Thus, in each iteration 5 edges of the network are chosen to form the subgraph \mathcal{G}_k and the values of the nodes are updated according to Algorithm 4. Note that similar plots can be obtained for any choice of block size. We run all algorithms with fixed stepsize $\omega = 1$. From Figure 10, it is obvious that for all networks under study, choosing a suitable momentum parameter $\beta \in (0, 1)$ gives faster convergence than having no momentum, $\beta = 0$.

7.3.4 Performance of AccGossip

In the last experiment on faster gossip algorithms we evaluate the performance of the proposed provably accelerated gossip protocols of Section 4.2. In particular we compare the standard RK (pairwise gossip algorithm of [6]) the mRK (Algorithm 3) and the AccGossip (Algorithm 6) with the two options for the selection of the parameters presented in Section 4.2.1.

The starting vector of values $x^0 = c$ is taken to be a Gaussian vector. For the implementation of mRK we use the same parameters with the ones suggested in the stochastic heavy ball (SGB) setting in [48]. For the AccRK (Option 1) we use $\lambda = \lambda_{\min}^+(\mathbf{A}^\top \mathbf{A})$ and for AccRK (Option 2) we

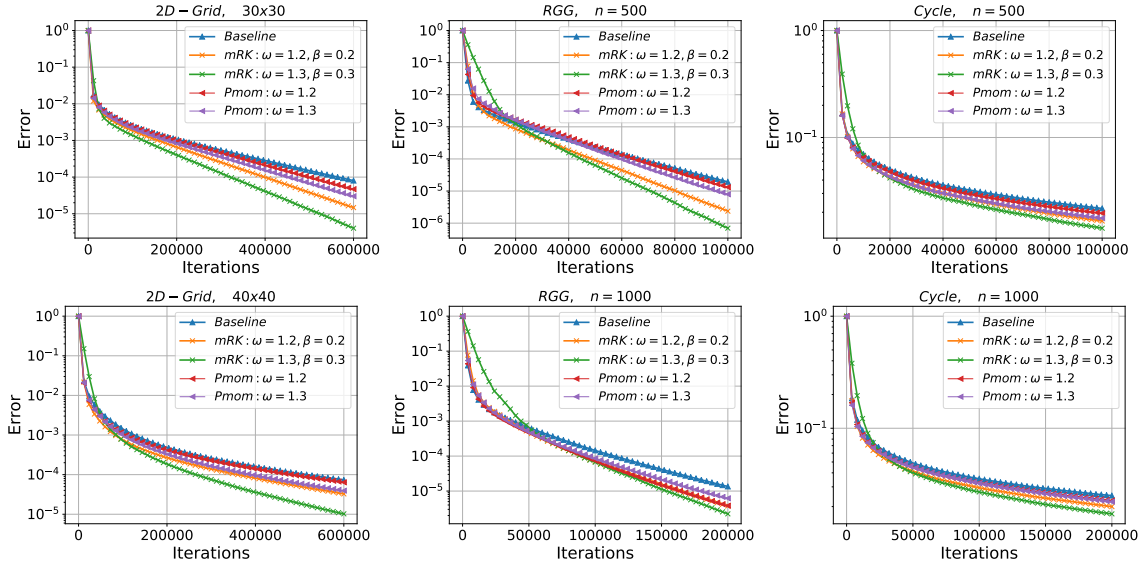


Figure 9: Comparison of mRK and the pairwise momentum method (Pmom), proposed in [40] (shift-register algorithm of Section 4.1.3). Following the connection between mRK and Pmom established in Section 4.1.3 the momentum parameter of mRK is chosen to be $\beta = \omega - 1$ and the stepsizes are selected to be either $\omega = 1.2$ or $\omega = 1.3$. The baseline method is the standard randomized pairwise gossip algorithm from [6]. The starting vector $x^0 = c \in \mathbb{R}^n$ is a Gaussian vector. The n in the title of each plot indicates the number of nodes of the network. For the grid graph this is $n \times n$.

select $\nu = m^{10}$. From Figure 11 it is clear that for all networks under study the two randomized gossip protocols with Nesterov momentum are faster than both the pairwise gossip algorithm of [6] and the mRK/SHB (Algorithm 3). To the best of our knowledge Algorithm 6 (Option 1 and Option 2) is the first randomized gossip protocol that converges with provably accelerated linear rate and as we can see from our experiment its faster convergence is also obvious in practice.

7.4 Relaxed randomized gossip without momentum

In the area of randomized iterative methods for linear systems it is known that over-relaxation (using of larger step-sizes) can be particularly helpful in practical scenarios. However, to the best of our knowledge there is not theoretical justification of why this is happening.

In our last experiment we explore the performance of relaxed randomized gossip algorithms ($\omega \neq 1$) without momentum and show that in this setting having larger stepsize can be particularly beneficial.

As we mentioned before (see Theorem 1) the sketch and project method (Algorithm 1) converges with linear rate when the step-size (relaxation parameter) of the method is $\omega \in (0, 2)$ and the best theoretical rate is achieved when $\omega = 1$. In this experiment we explore the performance of the standard pairwise gossip algorithm when the step-size of the update rule is chosen in $(1, 2)$. Since there is no theoretical proof of why over-relaxation can be helpful we perform the experiments using different starting values of the nodes. In particular we choose the values of vector $c \in \mathbb{R}^n$ to follow (i) Gaussian distribution, (ii) Uniform Distribution and (iii) to be integers values such that $c_i = i \in \mathbb{R}$. Our findings are presented in Figure 12. Note that for all networks under study and for all choices of starting values having larger stepsize, $\omega \in (1, 2)$ can lead to better performance. Interesting observation from Figure 12 is that the stepsizes $\omega = 1.8$ and $\omega = 1.9$ give the best performance (among the selected choices of stepsizes) for all networks and for all choices of starting vector $x^0 = c$.

¹⁰For the networks under study we have $m < \frac{1}{\lambda_{\min}^+(\mathbf{W})}$. Thus, by choosing $\nu = m$ we select the pessimistic upper bound of the parameter (48) and not its exact value (43). As we can see from the experiments, the performance is still accelerated and almost identical to the performance of AccRK (Option 1) for this choice of ν .

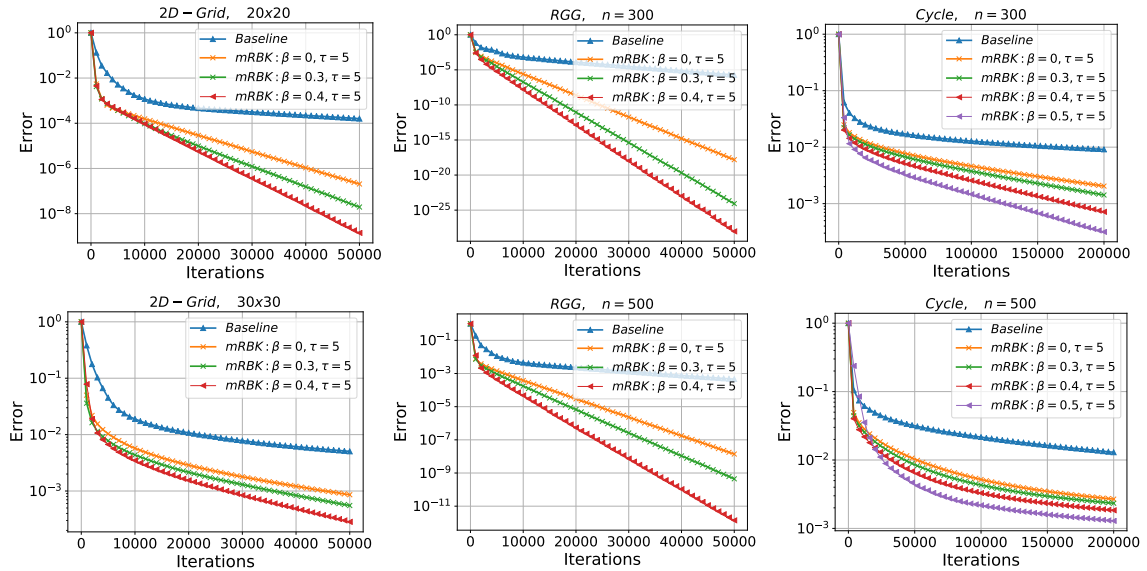


Figure 10: Comparison of mRBK with its no momentum variant RBK ($\beta = 0$). The stepsize for all methods is $\omega = 1$ and the block size is $\tau = 5$. The baseline method in the plots denotes the standard randomized pairwise gossip algorithm (block $\tau = 1$) and is plotted to highlight the benefits of having larger block sizes (at least in terms of iterations). The starting vector $x^0 = c \in \mathbb{R}^n$ is a Gaussian vector. The n in the title of each plot indicates the number of nodes. For the grid graph this is $n \times n$.

8 Conclusion and Future Direction of Research

In this work we present a general framework for the analysis and design of randomized gossip algorithms. Using tools from numerical linear algebra and the area of randomized projection methods for solving linear systems we propose novel serial, block and accelerated gossip protocols for solving the average consensus and weighted average consensus problems.

We believe that this work could open up several future avenues for research. Using similar approach with the one presented in this manuscript, many popular projection methods can be interpreted as gossip algorithms when used to solve linear systems encoding the underlying network. This can lead to the development of novel distributed protocols for average consensus.

In addition, we speculate that the gossip protocols presented in this work can be extended to the more general setting of multi-agent consensus optimization where the goal is to minimize the average of convex or non-convex functions $\frac{1}{n} \sum_{i=1}^n f_i(x)$ in a decentralized way [55].

9 Acknowledgments

The authors would like to thank Dr. Michael Rabbat for useful discussions related to the literature of gossip algorithms and for his comments during the writing of this paper.

References

- [1] M. Assran, N. Loizou, N. Ballas, and M. Rabbat. Stochastic gradient push for distributed deep learning. *arXiv preprint arXiv:1811.10792*, 2018.
- [2] M. Assran and M. Rabbat. Asynchronous subgradient-push. *arXiv preprint arXiv:1803.08950*, 2018.
- [3] N. S. Aybat and M. Gürbüzbalaban. Decentralized computation of effective resistances and acceleration of consensus algorithms. In *Signal and Information Processing (GlobalSIP), 2017 IEEE Global Conference on*, pages 538–542. IEEE, 2017.
- [4] T.C. Aysal, M.E. Yildiz, A.D. Sarwate, and A. Scaglione. Broadcast gossip algorithms for consensus. *IEEE Trans. Signal Process.*, 57(7):2748–2761, 2009.
- [5] F. Bénézit, A.G. Dimakis, P. Thiran, and M. Vetterli. Order-optimal consensus through randomized path averaging. *IEEE Trans. Inf. Theory*, 56(10):5150–5167, 2010.

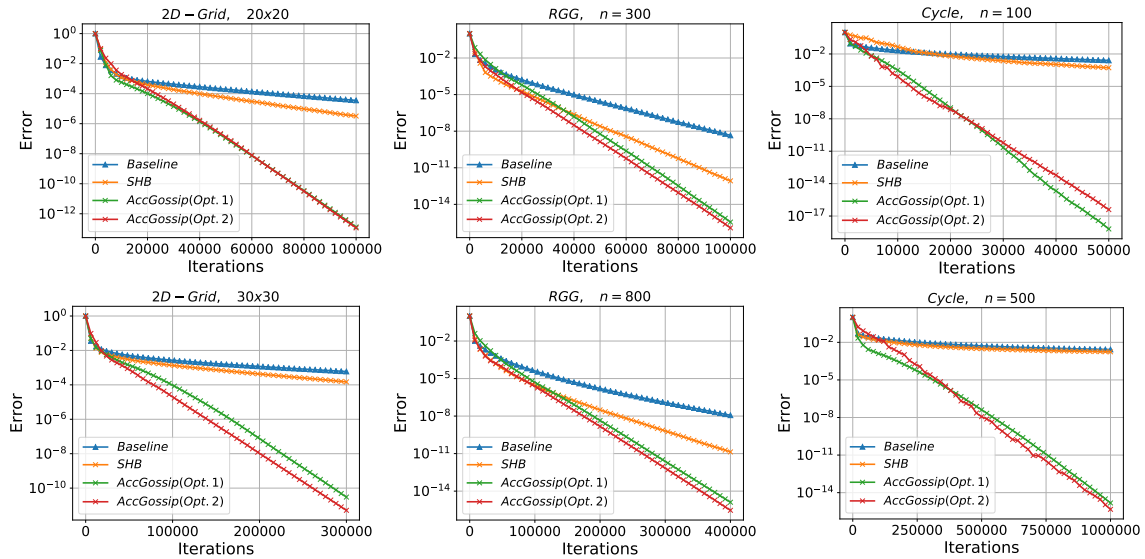


Figure 11: Performance of AccGossip (Option 1 and Option 2 for the parameters) in a 2-dimension grid, random geometric graph (RGG) and a cycle graph. The Baseline method corresponds to the randomized pairwise gossip algorithm proposed in [6] and the SHB represents the mRK (Algorithm 3) with the best choice of parameters as proposed in [48] ; The n in the title of each plot indicates the number of nodes of the network. For the grid graph this is $n \times n$.

- [6] S. Boyd, A. Ghosh, B. Prabhakar, and D. Shah. Randomized gossip algorithms. *IEEE Transactions on Information Theory*, 14(SI):2508–2530, 2006.
- [7] M. Cao, D.A. Spielman, and E.M. Yeh. Accelerated gossip algorithms for distributed computation. In *Proc. of the 44th Annual Allerton Conference on Communication, Control, and Computation*, pages 952–959, 2006.
- [8] T. Charalambous, M.G. Rabbat, M. Johansson, and C.N. Hadjicostis. Distributed finite-time computation of digraph parameters: Left eigenvector, out-degree and spectrum. *IEEE Trans. Control of Network Systems*, 3(2):137–148, June 2016.
- [9] I. Colin, A. Bellet, J. Salmon, and S. Cl  men  on. Gossip dual averaging for decentralized optimization of pairwise functions. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning-Volume 48*, pages 1388–1396. JMLR. org, 2016.
- [10] G. Cybenko. Dynamic load balancing for distributed memory multiprocessors. *J. Parallel Distrib. Comput.*, 7(2):279–301, 1989.
- [11] L. Dai, M. Soltanalian, and K. Pelckmans. On the randomized Kaczmarz algorithm. *IEEE Signal Processing Letters*, 21(3):330–333, 2014.
- [12] J. A. De Loera, J. Haddock, and D. Needell. A sampling Kaczmarz–Motzkin algorithm for linear feasibility. *SIAM Journal on Scientific Computing*, 39(5):S66–S87, 2017.
- [13] Morris H DeGroot. Reaching a consensus. *Journal of the American Statistical Association*, 69(345):118–121, 1974.
- [14] A.G. Dimakis, S. Kar, J.M.F. Moura, M.G. Rabbat, and A. Scaglione. Gossip algorithms for distributed signal processing. *Proceedings of the IEEE*, 98(11):1847–1864, 2010.
- [15] A.G. Dimakis, A.D. Sarwate, and M.J. Wainwright. Geographic gossip: Efficient averaging for sensor networks. *IEEE Trans. Signal Process.*, 56(3):1205–1216, 2008.
- [16] K. Du. Tight upper bounds for the convergence of the randomized extended Kaczmarz and Gauss–Seidel algorithms. *Numerical Linear Algebra with Applications*, 26(3):e2233, 2019.
- [17] J.C. Duchi, A. Agarwal, and M.J. Wainwright. Dual averaging for distributed optimization: convergence analysis and network scaling. *IEEE Transactions on Automatic Control*, 57(3):592–606, 2012.
- [18] Y.C. Eldar and D. Needell. Acceleration of randomized Kaczmarz method via the Johnson–Lindenstrauss lemma. *Numerical Algorithms*, 58(2):163–177, 2011.
- [19] N.M. Freris and A. Zouzias. Fast distributed smoothing of relative measurements. In *Decision and Control (CDC), 2012 IEEE 51st Annual Conference on*, pages 1411–1416. IEEE, 2012.
- [20] E. Ghadimi, A. Teixeira, M.G. Rabbat, and M. Johansson. The admm algorithm for distributed averaging: Convergence rates and optimal parameter selection. In *2014 48th Asilomar Conference on Signals, Systems and Computers*, pages 783–787. IEEE, 2014.
- [21] R. M. Gower, N. Loizou, X. Qian, A. Sailanbayev, E. Shulgin, and P. Richtarik. SGD: General analysis and improved rates. *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2019.

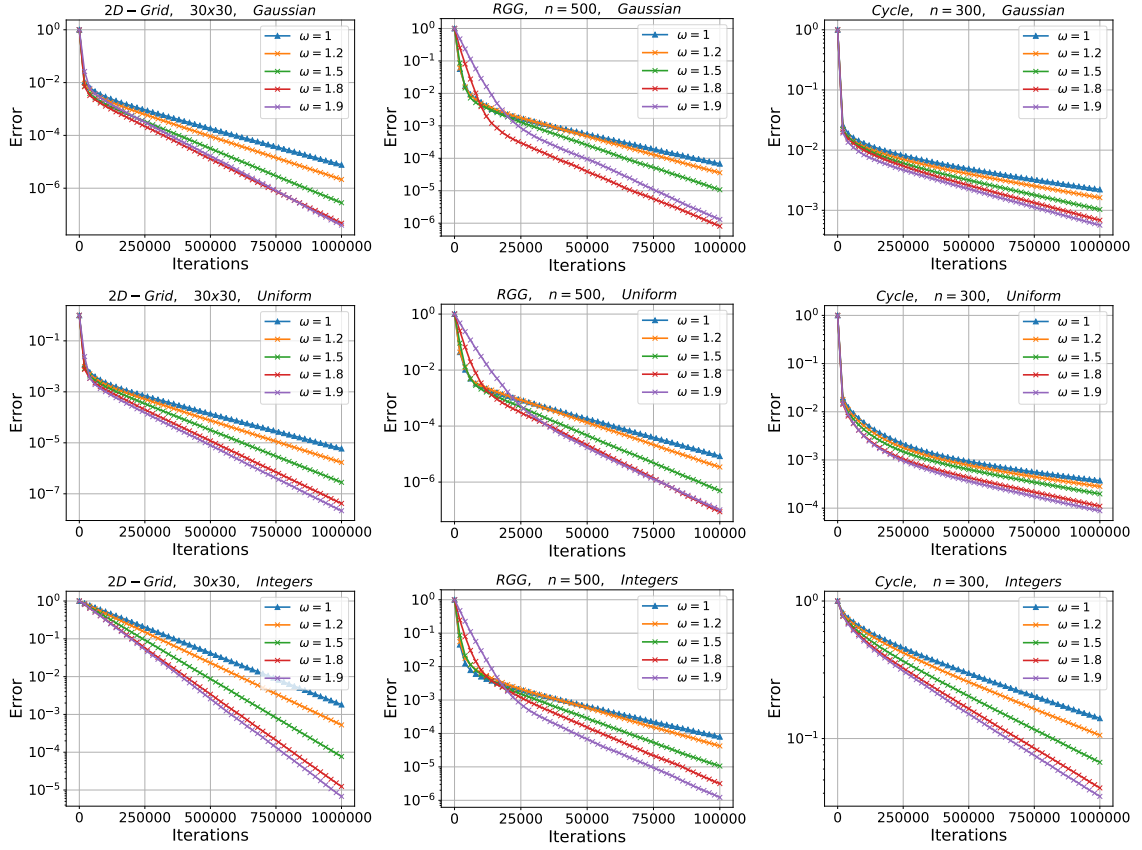


Figure 12: Performance of Relaxed randomized pairwise Gossip algorithm in a 2-dimension grid, random geometric graph (RGG) and a cycle graph. The case $\omega = 1$ corresponds to the randomized pairwise gossip algorithm proposed in [6]; The n in the title of each plot indicates the number of nodes of the network. For the grid graph this is $n \times n$. The title of each plot indicates the vector of starting values that is used.

- [22] R. M. Gower and P. Richtárik. Randomized quasi-newton updates are linearly convergent matrix inversion algorithms. *SIAM Journal on Matrix Analysis and Applications*, 38(4):1380–1409, 2017.
- [23] R.M. Gower, D. Goldfarb, and P. Richtárik. Stochastic block BFGS: squeezing more curvature out of data. In *ICML*, pages 1869–1878, 2016.
- [24] R.M. Gower, F. Hanzely, P. Richtárik, and S. U. Stich. Accelerated stochastic matrix inversion: general theory and speeding up BFGS rules for faster second-order optimization. In *Advances in Neural Information Processing Systems*, pages 1619–1629, 2018.
- [25] R.M. Gower and P. Richtárik. Randomized iterative methods for linear systems. *SIAM Journal on Matrix Analysis and Applications*, 36(4):1660–1690, 2015.
- [26] R.M. Gower and P. Richtárik. Stochastic dual ascent for solving linear systems. *arXiv preprint arXiv:1512.06890*, 2015.
- [27] R.M. Gower and P. Richtárik. Linearly convergent randomized iterative methods for computing the pseudoinverse. *arXiv preprint arXiv:1612.06255*, 2016.
- [28] J. Haddock and D. Needell. On motzkins method for inconsistent linear systems. *BIT Numerical Mathematics*, pages 1–15, 2018.
- [29] F. Hanzely, J. Konečný, N. Loizou, P. Richtárik, and D. Grishchenko. Privacy preserving randomized gossip algorithms. *arXiv preprint arXiv:1706.07636*, 2017.
- [30] F. Hanzely, J. Konečný, N. Loizou, P. Richtárik, and D. Grishchenko. A privacy preserving randomized gossip algorithm via controlled noise insertion. *NeurIPS Privacy Preserving Machine Learning Workshop*, 2018.
- [31] F. He, A. S. Morse, J. Liu, and S. Mou. Periodic gossiping. *IFAC Proceedings Volumes*, 44(1):8718–8723, 2011.
- [32] H. Hendrikx, L. Massoulié, and F. Bach. Accelerated decentralized optimization with local updates for smooth and strongly convex objectives. *arXiv preprint arXiv:1810.02660*, 2018.
- [33] A. G. Hernandez, M. L. Proença Jr, and T. Abrão. Improved weighted average consensus in distributed cooperative spectrum sensing networks. *Transactions on Emerging Telecommunications Technologies*, 29(3):e3259, 2018.
- [34] S. Kaczmarz. Angenäherte auflösung von systemen linearer gleichungen. *Bulletin International de l'Academie Polonaise des Sciences et des Lettres*, 35:355–357, 1937.

- [35] E. Kokiopoulou and P. Frossard. Polynomial filtering for fast convergence in distributed consensus. *IEEE Transactions on Signal Processing*, 57(1):342–354, 2009.
- [36] A. Koloskova, S. U. Stich, and M. Jaggi. Decentralized stochastic optimization and gossip algorithms with compressed communication. *arXiv preprint arXiv:1902.00340*, 2019.
- [37] A. Krizhevsky, I. Sutskever, and G.E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1097–1105, 2012.
- [38] D. Leventhal and A.S. Lewis. Randomized methods for linear constraints: convergence rates and conditioning. *Mathematics of Operations Research*, 35(3):641–654, 2010.
- [39] X. Lian, C. Zhang, C.-J. Hsieh, W. Zhang, and J. Liu. Asynchronous decentralized parallel stochastic gradient descent. In *International Conference on Machine Learning*, pages 3049–3058, 2018.
- [40] J. Liu, B.D.O. Anderson, M. Cao, and A.S. Morse. Analysis of accelerated gossip algorithms. *Automatica*, 49(4):873–883, 2013.
- [41] J. Liu and S. Wright. An accelerated randomized Kaczmarz algorithm. *Mathematics of Computation*, 85(297):153–178, 2016.
- [42] Ji Liu, Shaoshuai Mou, A Stephen Morse, Brian DO Anderson, and Changbin Yu. Deterministic gossiping. *Proceedings of the IEEE*, 99(9):1505–1524, 2011.
- [43] Y. Liu, J. Wu, I. Manchester, and G. Shi. Privacy-preserving gossip algorithms. *arXiv preprint arXiv:1808.00120*, 2018.
- [44] Yang Liu, Bo Li, Brian Anderson, and Guodong Shi. Clique gossiping. *arXiv preprint arXiv:1706.02540*, 2017.
- [45] N. Loizou, M. Rabbat, and P. Richtárik. Provably accelerated randomized gossip algorithms. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7505–7509. IEEE, 2019.
- [46] N. Loizou and P. Richtárik. A new perspective on randomized gossip algorithms. In *2016 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 440–444. IEEE, 2016.
- [47] N. Loizou and P. Richtárik. Linearly convergent stochastic heavy ball method for minimizing generalization error. *NIPS-Workshop on Optimization for Machine Learning*, 2017.
- [48] N. Loizou and P. Richtárik. Momentum and stochastic momentum for stochastic gradient, Newton, proximal point and subspace descent methods. *arXiv preprint arXiv:1712.09677*, 2017.
- [49] N. Loizou and P. Richtárik. Accelerated gossip via stochastic heavy ball method. In *2018 56th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 927–934. IEEE, 2018.
- [50] N. Loizou and P. Richtárik. Convergence analysis of inexact randomized iterative methods. *arXiv preprint arXiv:1903.07971*, 2019.
- [51] A. Ma, D. Needell, and A. Ramdas. Convergence properties of the randomized extended Gauss-Seidel and Kaczmarz methods. *SIAM Journal on Matrix Analysis and Applications*, 36(4):1590–1604, 2015.
- [52] M. S. Morshed, M.S. Islam, et al. Accelerated sampling Kaczmarz Motzkin algorithm for linear feasibility problem. *arXiv preprint arXiv:1902.03502*, 2019.
- [53] S. Mou, C. Yu, B.D.O Anderson, and A. S. Morse. Deterministic gossiping with a periodic protocol. In *Decision and Control (CDC), 2010 49th IEEE Conference on*, pages 5787–5791. IEEE, 2010.
- [54] I. Necoara and V. Nedelcu. Rate analysis of inexact dual first-order methods application to dual decomposition. *IEEE Transactions on Automatic Control*, 59(5):1232–1243, 2014.
- [55] A. Nedić, A. Olshevsky, and M. G. Rabbat. Network topology and communication-computation tradeoffs in decentralized optimization. *Proceedings of the IEEE*, 106(5):953–976, 2018.
- [56] D. Needell. Randomized Kaczmarz solver for noisy linear systems. *BIT Numerical Mathematics*, 50(2):395–403, 2010.
- [57] D. Needell and J.A. Tropp. Paved with good intentions: analysis of a randomized block Kaczmarz method. *Linear Algebra Appl.*, 441:199–221, 2014.
- [58] D. Needell, R. Zhao, and A. Zouzias. Randomized block Kaczmarz method with projection for solving least squares. *Linear Algebra Appl.*, 484:322–343, 2015.
- [59] Y. Nesterov. A method of solving a convex programming problem with convergence rate $o(1/k^2)$. *Soviet Mathematics Doklady*, 27:372–376, 1983.
- [60] Y. Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012.
- [61] Y. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*, volume 87. Springer Science & Business Media, 2013.
- [62] J. Nutini, B. Sepehry, I. Laradji, M. Schmidt, H. Koepke, and A. Virani. Convergence rates for greedy Kaczmarz algorithms, and faster randomized Kaczmarz rules using the orthogonality graph. In *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*, pages 547–556. AUAI Press, 2016.

- [63] A. Olshevsky. Linear time average consensus on fixed graphs and implications for decentralized optimization and multi-agent control. *arXiv preprint arXiv:1411.4186*, 2014.
- [64] A. Olshevsky and J.N. Tsitsiklis. Convergence speed in distributed consensus and averaging. *SIAM Journal on Control and Optimization*, 48(1):33–55, 2009.
- [65] B. N. Oreshkin, M. J. Coates, and M. G. Rabbat. Optimization and analysis of distributed averaging with short node memory. *IEEE Transactions on Signal Processing*, 58(5):2850–2865, 2010.
- [66] F. Pedroche Sánchez, M. Rebollo Pedruelo, C. Carrascosa Casamayor, and A. Palomares Chust. Convergence of weighted-average consensus for undirected graphs. *International Journal of Complex Systems in Science*, 4(1):13–16, 2014.
- [67] B.T. Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17, 1964.
- [68] C. Popa. Convergence rates for Kaczmarz-type algorithms. *arXiv preprint arXiv:1701.08002*, 2017.
- [69] Z. Qu, P. Richtárik, M. Takáč, and O. Fercoq. SDNA: Stochastic dual Newton ascent for empirical risk minimization. *ICML*, 2016.
- [70] Z. Qu, P. Richtárik, and T. Zhang. Quartz: Randomized dual coordinate ascent with arbitrary sampling. In *Advances in Neural Information Processing Systems*, pages 865–873, 2015.
- [71] M.G. Rabbat, R.D. Nowak, and J.A. Bucklew. Generalized consensus computation in networked systems with erasure links. In *IEEE 6th Workshop on Signal Processing Advances in Wireless Communications*, pages 1088–1092. IEEE, 2005.
- [72] P. Richtárik and M. Takáč. Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function. *Mathematical Programming*, 144(1-2):1–38, 2014.
- [73] P. Richtárik and M. Takáč. Stochastic reformulations of linear systems: algorithms and convergence theory. *arXiv:1706.01108*, 2017.
- [74] F. Schöpfer and D.A. Lorenz. Linear convergence of the randomized sparse Kaczmarz method. *arXiv preprint arXiv:1610.02889*, 2016.
- [75] Sh. Shalev-Shwartz and T. Zhang. Stochastic dual coordinate ascent methods for regularized loss. *Journal of Machine Learning Research*, 14(1):567–599, 2013.
- [76] W. Shi, Q. Ling, G. Wu, and W. Yin. Extra: An exact first-order algorithm for decentralized consensus optimization. *SIAM Journal on Optimization*, 25(2):944–966, 2015.
- [77] T. Strohmer and R. Vershynin. A randomized Kaczmarz algorithm with exponential convergence. *J. Fourier Anal. Appl.*, 15(2):262–278, 2009.
- [78] I. Sutskever, J. Martens, G.E. Dahl, and G.E. Hinton. On the importance of initialization and momentum in deep learning. *ICML (3)*, 28:1139–1147, 2013.
- [79] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *CVPR*, pages 1–9, 2015.
- [80] K. Tsianos, S. Lawlor, and M. G. Rabbat. Communication/computation tradeoffs in consensus-based distributed optimization. In *Conference on Neural Information Processing Systems*, 2012.
- [81] John Tsitsiklis, Dimitri Bertsekas, and Michael Athans. Distributed asynchronous deterministic and stochastic gradient optimization algorithms. *IEEE transactions on automatic control*, 31(9):803–812, 1986.
- [82] S. Tu, S. Venkataraman, A.C. Wilson, A. Gittens, M.I. Jordan, and B. Recht. Breaking locality accelerates block Gauss-Seidel. In *ICML*, 2017.
- [83] D. Ustebay, M. Coates, and M. Rabbat. Greedy gossip with eavesdropping. In *Wireless Pervasive Computing, 2008. ISWPC 2008. 3rd International Symposium on*, pages 759–763. IEEE, 2008.
- [84] E. Wei and A. Ozdaglar. On the $\mathcal{O}(1/k)$ convergence of asynchronous distributed alternating direction method of multipliers. In *2013 IEEE Global Conference on Signal and Information Processing*, pages 551–554. IEEE, 2013.
- [85] A.C. Wilson, R. Roelofs, M. Stern, N. Srebro, and B. Recht. The marginal value of adaptive gradient methods in machine learning. *arXiv preprint arXiv:1705.08292*, 2017.
- [86] S.J. Wright. Coordinate descent algorithms. *Mathematical Programming*, 151(1):3–34, 2015.
- [87] L. Xiao and S. Boyd. Fast linear iterations for distributed averaging. *Systems & Control Letters*, 53(1):65–78, 2004.
- [88] L. Xiao, S. Boyd, and S. Lall. A scheme for robust distributed sensor fusion based on average consensus. In *Information Processing in Sensor Networks, 2005. IPSN 2005. Fourth International Symposium on*, pages 63–70. IEEE, 2005.
- [89] C. B. Yu, B.D.O. Anderson, S. Mou, J. Liu, F. He, and A. S. Morse. Distributed averaging using periodic gossiping. *IEEE Transactions on Automatic Control*, 2017.

- [90] K. Yuan, Q. Ling, and W. Yin. On the convergence of decentralized gradient descent. *SIAM Journal on Optimization*, 26(3):1835–1854, 2016.
- [91] W. Zhang, Y. Guo, H. Liu, Y. J. Chen, Z. Wang, and J. Mitola III. Distributed consensus-based weight design for cooperative spectrum sensing. *IEEE Transactions on Parallel and Distributed Systems*, 26(1):54–64, 2015.
- [92] W. Zhang, Z. Wang, Y. Guo, H. Liu, Y. Chen, and J. Mitola III. Distributed cooperative spectrum sensing based on weighted average consensus. In *2011 IEEE Global Telecommunications Conference-GLOBECOM 2011*, pages 1–6. IEEE, 2011.
- [93] A. Zouzias and N.M. Freris. Randomized extended Kaczmarz for solving least squares. *SIAM Journal on Matrix Analysis and Applications*, 34(2):773–793, 2013.
- [94] A. Zouzias and N.M. Freris. Randomized gossip algorithms for solving Laplacian systems. In *Control Conference (ECC), 2015 European*, pages 1920–1925. IEEE, 2015.

A Missing Proofs

A.1 Proof of Theorem 4

Proof. Let $z^k := \|x^k - x^*\|$, $x^0 = c$ is the starting point and ρ is as defined in (10). From Theorem 1 we know that sketch and project method converges with

$$\mathbb{E}[\|x^k - x^*\|_{\mathbf{B}}^2] \leq \rho^k \|x^0 - x^*\|_{\mathbf{B}}^2, \quad (55)$$

where x^* is the solution of (7). Inequality (55), together with Markov inequality can be used to give the following bound

$$\mathbb{P}(z^k/z^0 \geq \varepsilon^2) \leq \frac{\mathbb{E}(z^k/z^0)}{\varepsilon^2} \leq \frac{\rho^k}{\varepsilon^2}. \quad (56)$$

Therefore, as long as k is large enough so that $\rho^k \leq \varepsilon^3$, we have $\mathbb{P}(z^k/z^0 \geq \varepsilon^2) \leq \varepsilon$. That is, if

$$\rho^k \leq \varepsilon^3 \Leftrightarrow k \geq \frac{3 \log \varepsilon}{\log \rho} \Leftrightarrow k \geq \frac{3 \log(1/\varepsilon)}{\log(1/\rho)},$$

then:

$$\mathbb{P}\left(\frac{\|x^k - \bar{c}1\|}{\|x^0 - \bar{c}1\|} \geq \varepsilon\right) \leq \varepsilon.$$

Hence, an upper bound for value $T_{ave}(\varepsilon)$ can be obtained as follows,

$$\begin{aligned} T_{ave}(\varepsilon) &= \sup_{c \in \mathbb{R}^n} \inf \left\{ k : \mathbb{P}\left(z^k > \varepsilon z^0\right) \leq \varepsilon \right\} \leq \sup_{c \in \mathbb{R}^n} \inf \left\{ k : k \geq \frac{3 \log(1/\varepsilon)}{\log(1/\rho)} \right\} \\ &= \sup_{c \in \mathbb{R}^n} \frac{3 \log(1/\varepsilon)}{\log(1/\rho)} = \frac{3 \log(1/\varepsilon)}{\log(1/\rho)} \leq \frac{3 \log(1/\varepsilon)}{1 - \rho}, \end{aligned} \quad (57)$$

where in last inequality we use $1/\log(1/\rho) \leq 1/1 - \rho$ which is true because $\rho \in (0, 1)$. \square

A.2 Proof of Theorem 5

Proof. The following notation conventions are used in this proof. With q_k we indicate the number of connected components of subgraph \mathcal{G}_k , while with \mathcal{V}_r we denote the set of nodes of each connected component q_k ($r \in \{1, 2, \dots, q_k\}$). Finally, $|\mathcal{V}_r|$ shows the cardinality of set \mathcal{V}_r . Notice that, if \mathcal{V} is the set of all nodes of the graph then $\mathcal{V} = \bigcup_{r=\{1,2,\dots,q\}} \mathcal{V}_r$ and $|\mathcal{V}| = \sum_{r=1}^q |\mathcal{V}_r|$.

Note that from equation (29), the update of RBK for $\mathbf{A} = \mathbf{Q}$ (Incidence matrix) can be expressed as follows:

$$\begin{aligned} &\underset{x}{\text{minimize}} \quad \phi^k(x) := \|x - x^k\|^2 \\ &\text{subject to} \quad \mathbf{I}_{\mathcal{C}}^\top \mathbf{Q} x = 0 \end{aligned} \quad (58)$$

Notice that $\mathbf{I}_{\mathcal{C}}^\top \mathbf{Q}$ is a row submatrix of matrix \mathbf{Q} with rows those that correspond to the random set $C \subseteq \mathcal{E}$ of the edges. From the expression of matrix \mathbf{Q} we have that

$$(\mathbf{I}_{\mathcal{C}}^\top \mathbf{Q})_e^\top = f_i - f_j, \quad \forall e = (i, j) \in C \subseteq \mathcal{E}.$$

Now, using this, it can be seen that the constraint $\mathbf{I}_C^\top \mathbf{Q}x = 0$ of problem (7) is equivalent to q equations (number of connected components) where each one of them forces the values x_i^{k+1} of the nodes $i \in \mathcal{V}_r$ to be equal. That is, if we use z_r to represent the value of all nodes that belong in the connected component r then:

$$x_i^{k+1} = z_r \quad \forall i \in \mathcal{V}_r, \quad (59)$$

and the constrained optimization problem (7) can be expressed as unconstrained as follows:

$$\underset{z}{\text{minimize}} \quad \phi^k(z) = \sum_{i \in \mathcal{V}_1} (z_1 - x_i^k)^2 + \dots + \sum_{i \in \mathcal{V}_q} (z_q - x_i^k)^2, \quad (60)$$

where $z = (z_1, z_2, \dots, z_q) \in \mathbb{R}^q$ is the vector of all values z_r when $r \in \{1, 2, \dots, q\}$. Since our problem is unconstrained the minimum of equation (60) is obtained when $\nabla \phi^k(z) = 0$.

By evaluating partial derivatives of (60) we obtain:

$$\frac{\partial \phi^k(z)}{\partial z_r} = 0 \iff \sum_{i \in \mathcal{V}_r} 2(z_r - x_i^k) = 0.$$

As a result,

$$z_r = \frac{\sum_{i \in \mathcal{V}_r} x_i^k}{|\mathcal{V}_r|}, \quad \forall r \in \{1, 2, \dots, q\}.$$

Thus from (59), the value of each node $i \in \mathcal{V}_r$ will be updated to

$$x_i^{k+1} = z_r = \frac{\sum_{i \in \mathcal{V}_r} x_i^k}{|\mathcal{V}_r|}.$$

□

B Notation Glossary

The Basics	
\mathbf{A}, b	$m \times n$ matrix and $m \times 1$ vector defining the system $\mathbf{A}x = b$
\mathcal{L}	$\{x : \mathbf{A}x = b\}$ (solution set of the linear system)
\mathbf{B}	$n \times n$ symmetric positive definite matrix
$\langle x, y \rangle_{\mathbf{B}}$	$x^{\top} \mathbf{B} y$ (\mathbf{B} -inner product)
$\ x\ _{\mathbf{B}}$	$\sqrt{\langle x, x \rangle_{\mathbf{B}}}$ (\mathbf{B} -norm)
\mathbf{M}^{\dagger}	Moore-Penrose pseudoinverse of matrix \mathbf{M}
\mathbf{S}	a random real matrix with m rows
\mathcal{D}	distribution from which matrix \mathbf{S} is drawn ($\mathbf{S} \sim \mathcal{D}$)
\mathbf{H}	$\mathbf{S}(\mathbf{S}^{\top} \mathbf{A} \mathbf{B}^{-1} \mathbf{A}^{\top} \mathbf{S})^{\dagger} \mathbf{S}^{\top}$
\mathbf{Z}	$\mathbf{A}^{\top} \mathbf{H} \mathbf{A}$
$\text{Range}(\mathbf{M})$	range space of matrix \mathbf{M}
$\text{Null}(\mathbf{M})$	null space of matrix \mathbf{M}
$\mathbb{P}(\cdot)$	probability of an event
$\mathbb{E}[\cdot]$	expectation
Projections	
$\Pi_{\mathcal{L}, \mathbf{B}}(x)$	projection of x onto \mathcal{L} in the \mathbf{B} -norm
$\mathbf{B}^{-1} \mathbf{Z}$	projection matrix, in the \mathbf{B} -norm, onto $\text{Range}(\mathbf{B}^{-1} \mathbf{A}^{\top} \mathbf{S})$
Graphs	
$\mathcal{G} = (\mathcal{V}, \mathcal{E})$	an undirected graph with vertices \mathcal{V} and edges \mathcal{E}
n	$= \mathcal{V} $ (number of vertices)
m	$= \mathcal{E} $ (number of edges)
$e = (i, j) \in \mathcal{E}$	edge of \mathcal{G} connecting nodes $i, j \in \mathcal{V}$
d_i	degree of node i
$c \in \mathbb{R}^n$	$= (c_1, \dots, c_n)$; a vector of private values stored at the nodes of \mathcal{G}
\bar{c}	$\bar{c} = \frac{\sum_i^n \mathbf{B}_{ii} c_i}{\sum_i^n \mathbf{B}_{ii}}$ (the weighted average of the private values)
$\mathbf{Q} \in \mathbb{R}^{m \times m}$	Incidence matrix of \mathcal{G}
$\mathbf{L} \in \mathbb{R}^{n \times n}$	$= \mathbf{Q}^{\top} \mathbf{Q}$ (Laplacian matrix of \mathcal{G})
$\mathbf{D} \in \mathbb{R}^{n \times n}$	$= \text{Diag}(d_1, d_2, \dots, d_n)$ (Degree matrix of \mathcal{G})
$\mathbf{L}^{rw} \in \mathbb{R}^{n \times n}$	$= \mathbf{D}^{-1} \mathbf{L}$ (random walk normalized Laplacian matrix of \mathcal{G})
$\mathbf{L}^{sym} \in \mathbb{R}^{n \times n}$	$= \mathbf{D}^{-1/2} \mathbf{L} \mathbf{D}^{-1/2}$ (symmetric normalized Laplacian matrix of \mathcal{G})
$\alpha(\mathcal{G})$	$= \lambda_{\min}^+(\mathbf{L})$ (algebraic connectivity of \mathcal{G})
Eigenvalues	
\mathbf{W}	$\mathbf{B}^{-1/2} \mathbb{E}[\mathbf{Z}] \mathbf{B}^{-1/2}$ (psd matrix)
$\lambda_1, \dots, \lambda_n$	eigenvalues of \mathbf{W}
$\lambda_{\max}, \lambda_{\min}^+$	largest and smallest nonzero eigenvalues of \mathbf{W}
Algorithms	
ω	relaxation parameter / stepsize
β	heavy ball momentum parameter
ρ	$1 - \omega(2 - \omega)\lambda_{\min}^+$

Table 1: Frequently used notation.