Title
Kaczmarz's Method for Systems of Linear Equations: Coherence, Corruption, and Consensus

Author
Jarman, Benjamin Edward

Publication Date
2023

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA
Los Angeles

Kaczmarz's Method for Systems of Linear Equations: Coherence, Corruption, and Consensus

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Mathematics

by

Benjamin Edward Jarman

2023

ABSTRACT OF THE DISSERTATION

Kaczmarz's Method for Systems of Linear Equations: Coherence, Corruption, and Consensus

by

Benjamin Edward Jarman
Doctor of Philosophy in Mathematics
University of California, Los Angeles, 2023
Professor Deanna M. Hunter, Chair

In this thesis, we study variants of Kaczmarz's iterative method for solving systems of linear equations. We introduce and analyze a variant that incorporates heavy-ball momentum to accelerate convergence when the data is highly coherent. We provide theoretical and empirical results for both fixed systems and streamed data. Furthermore, we introduce variants of the method tailored to a specific sparse corruption model. These variants leverage information from the residual to avoid corrupted rows and estimate the solution's direction. We offer theoretical guarantees and empirical evidence showcasing the convergence of our methods to the uncorrupted system's solution. Additionally, we establish an equivalence between block variants of Kaczmarz's method and gossip protocols, which are widely used in distributed computing. We develop new convergence theory on both sides of this equivalence and extend our results to account for noisy models of network communication.

The dissertation of Benjamin Edward Jarman is approved.

Mason A. Porter

Guido F. Montufar Cuartas

Chenfanfu Jiang

Deanna M. Hunter, Committee Chair

University of California, Los Angeles

2023

For my Mum

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

It is difficult to express how grateful I am for my advisor, Deanna Needell. Without her personal guidance, enduring support, and endless patience, I would not have reached this point. Throughout numerous challenging times – not least the COVID-19 pandemic – she has been a pillar of strength for myself and other members of our research group. She has been the guiding force behind the work in this dissertation, both directly and through the wonderful collaborations she introduced me to.

I owe a great deal to the incredible collaborators I have had during my time at UCLA. In particular, I would like to thank Jamie Haddock for being a mentor, collaborator, and friend. I will treasure her invaluable advice on navigating graduate school, academia, and life, and well as the times we spent watching the Dodgers. I am grateful to Mason Porter for his unwavering support for myself and his other mentees, and for helping me improve as a scientific writer. I also extend my thanks to Liza Rebrova, Lara Kassab, Jacob Moorman, Yotam Yaniv, Nate Mankovich, Lu Cheng, Chen Yap, Abby Hickok, Jerry Luo, Michael Johnson, Longxiu Huang, Joyce Chew, and our 2021 REU cohort for the fruitful collaborations we shared.

I must express my gratitude to all the fantastic educators who helped me along the way. In approximate chronological order, I thank Mykola Postelnyak, Allan Graham, Peter Johnson, Henry Wilton, Imre Leader, Perla Sousi, and Tim Austin for lessons in both mathematics and life.

I am thankful to my father, Chris, for recognizing my aptitude for and enjoyment of mathematics at a young age, and for supporting me all throughout this journey. It has been a long road with many ups and downs, but I hope to have fulfilled the potential you saw in me and made you proud. I thank my brother Sam and his wife Nicole for their unwavering support. Their guidance has been invaluable in navigating my twenties, and I am deeply grateful to them for always providing me a place to stay whenever I returned to the UK. I also extend my gratitude to their children, Cara, Marlowe, and Avery, for making every

Christmas magical and ensuring I don't sleep in too late. I can't wait to be on the same side of the world as all of you once again. I thank my Gran Cynthia and her sisters Marcia and Eveon for all their love, support, and kindness.

Five years ago, I left home to move across the globe with no idea who I would find on the other side, but I have been fortunate to forge friendships that I hope will last a lifetime. I am grateful to everyone with whom I have shared adventures in LA and beyond, to the mathematics graduate student body for being a constant source of support, and especially to my roommates: Calvin, Matt, and Eilon. Moreover, I appreciate all the friendships that have endured since my undergraduate days. Special thanks to Dan for traversing this journey alongside me for all these years and always being there to share both frustrations and elations along the way.

I must pay special thanks to those in my family whose support on this journey has been immeasurable, but who are not here to see the end of it. I keep close to my heart all of those who we lost along the way. Grandad, Grandad George, Aunt Frances, Uncle George, Uncle John: I miss you all.

Lastly, and above all, I thank my Mum. I am so glad that you saw the start of my journey: I remember so clearly the day you dropped me off at Trinity's Great Gate for the first time, standing next to our old Ford Focus with the biggest smile. I wish more than anything else that you could be here for the end. I dedicate everything I have done and will do to your memory, which I keep close always.

Chapter 2 is a version of [40] and is joint work with Prof. Deanna Needell. Deanna Needell proposed and supervised the project. I contributed the convergence analysis, experiments, and writing.

Chapter 3 is a version of [11] and is joint with with Lu Cheng, Liza Rebrova, and Prof. Deanna Needell. Deanna Needell and Liza Rebrova proposed and co-supervised the project. Lu Cheng and I contributed the codebase and experiments. Liza Rebrova and I contributed the convergence analysis.

Chapter 4 is a version of [33] and is joint work with Chen Yap and Jamie Haddock.

Chen Yap contributed the codebase, and Chen Yap and I contributed the experiments. Jamie Haddock and I contributed the theory and wrote the paper together.

Chapter 5 is a version of [41] and is joint work with Yotam Yaniv and Prof. Deanna Needell. I proposed the project and Deanna Needell supervised. Yotam Yaniv and I contributed the convergence analysis and experiments.

I am grateful to the UCLA Graduate Division (Graduate Dean's Fellowship) and the National Science Foundation (NSF DMS-2011140 and NSF DMS-2108479) for financial support during the Ph.D. program.

VITA

2018	B.A., University of Cambridge.
2018	M.Math., University of Cambridge.
2018	Graduate Dean's Fellowship, UCLA.
2020	M.A., UCLA.
2021	M.A., University of Cambridge.
2021	Liggett Teaching Fellowship, UCLA.
2022	Quantitative Research Intern, G-Research.
2018–2023	Teaching Assistant, UCLA
2021–2023	Graduate Student Researcher, UCLA

PUBLICATIONS

B. Jarman, Y. Yaniv, D. Needell. "Online Signal Recovery via Heavy Ball Kaczmarz." *56th Asilomar Conference on Signals, Systems and Computers*, 2022.

L. Cheng, B. Jarman, D. Needell, L. Rebrova. "On Block Accelerations of Quantile Randomized Kaczmarz for Corrupted Systems of Linear Equations." *Inverse Problems*, **39**(2):024002, 2022.

P. Li, C. Tseng, Y. Zheng, J.A. Chew, L. Huang, B. Jarman, D. Needell. "Guided Semi-

Supervised Non-negative Matrix Factorization." *Algorithms*, **15**(2), 2022.

J. Haddock, B. Jarman, C. Yap. "Paving the Way for Consensus: Convergence of Block Gossip Algorithms." *IEEE Transactions on Information Theory*, **68**(11):7515–7527, 2022.

B. Jarman, D. Needell. "QuantileRK: Solving large-scale linear systems with corrupted, noisy measurement data." *55th Asilomar Conference on Signals, Systems and Computers*, 2021.

CHAPTER 1

Introduction

In our increasingly data-driven world, businesses and individuals alike rely on complex mathematical models to make critical decisions based on vast amounts of data. For instance, a bank may have a credit model for loan approvals, a hospital may have a model for treatment plans, or a space agency may have a model for flight paths. The lack of understanding of the algorithms underpinning such models can have negative impacts on individual lives and the world at large, particularly when the data used to make these decisions is imperfect. This thesis aims to deepen the understanding of a family of iterative methods for solving linear systems of equations, making them more accessible to people with varying mathematical backgrounds.

Solving a system of linear equations of the form $Ax = b$ is a fundamental component of many models in data science. As datasets grow larger, traditional methods for solving such systems may fail due to time or memory constraints. In recent times, iterative methods have become increasingly popular. These methods often require only a small amount of data to be loaded into memory at any particular time and require a relatively small number of computations per iteration. We study variants of one particular iterative method, the Kaczmarz method.

The Kaczmarz method takes an initial guess at the solution of the system, x^0, and sequentially projects it onto the hyperplanes defined by the rows of the system to produce a sequence of iterates. These iterates will get successively closer to the solution due to the geometry of the problem. The rate of convergence depends on the structure of the matrix A and the order in which rows are chosen for projection. While the convergence of the Kaczmarz method is well-understood when the system is consistent, in practice,

it is often applied in settings where imperfections in data (arising, for example, through transmission, storage, or collection) lead to inconsistent systems. In Chapters 2 and 3, we study variants of the Kaczmarz method designed to handle such imperfections and still return the desired solution.

The Kaczmarz method may also be used as an online method to recover a signal from a sequence of linear measurements. This scenario arises, for example, in medical scanning technologies, where a potential roadblock for the Kaczmarz method is the coherence of successive measurements. In Chapter 4, we show that the addition of a momentum term can significantly accelerate convergence in the coherent measurement setting.

A common problem in distributed computing is the *average consensus* problem. Given a network consisting of nodes and edges, where each node stores a value, the objective is for each node to learn the average of all of these values through communicating across edges in the network. This scenario arises in clock synchronization, blockchain technology, GPS-free localization, and elsewhere. Efficient iterative methods are preferred when networks are large, as is common in practice. Moreover, communications often come with some level of noise, which methods should be robust to.

In Chapter 5, we show that a popular class of iterative methods for solving the average consensus problem, called gossip protocols, may be identified with a certain multi-row variant of the Kaczmarz method for solving linear systems. We demonstrate that this equivalence brings forth a new depth of understanding for gossip protocols, including new convergence guarantees and empirical results on the role played by noise.

We would also like to draw attention to work completed by the author in the course of their studies that has not been included here, due to the topics of said work being distinct from the themes of this dissertation. In [38], we study the problem of quantifying the equity of a geographical distribution of social resource sites (in our study, polling sites). We employ persistent homology, a tool from topological data analysis, to examine the effective availability and coverage of polling sites across five cities and one county in the United States. In [50], we address the problem of performing classification and topic modelling

on large-scale datasets of documents, accompanied by pre-assigned document class labels. We propose a new method, Guided Semi-Supervised Non-negative Matrix Factorization (GSSNMF), that incorporates these class labels as well as user-designed topic seed words. GSSNMF produces an assignment of documents into classes, along with a set of topics. We demonstrate the effectiveness of our method on legal documents provided by the California Innocence Project.

CHAPTER 2

QuantileRK: Solving Large-Scale Linear Systems with Corrupted, Noisy Data

This chapter is a version of [40] and is joint work with Prof. Deanna Needell. Deanna Needell proposed and supervised the project. I contributed the convergence analysis, experiments, and writing. We discuss the use of a quantile-based variant of the Kaczmarz method for solving linear systems that have been affected by (potentially arbitrarily large) sparse corruptions, as well as small-scale noise that could affect every equation. These phenomena are prevalent in real-world linear systems, hence both introducing and understanding efficient methods to solve such imperfect systems is of much importance. We give both theoretical guarantees and experimental results to demonstrate the effectiveness of the method.

2.1 Introduction

From medical imaging [39], to image reconstruction and signal processing [37, 22], to modern data science and statistical analysis [49], solving systems of linear equations, has long been a central problem in applied mathematics. Such systems will often be large, overdetermined, and consistent: we consider the system $\mathbf{A}\mathbf{x} = \mathbf{b}$, where $\mathbf{A} \in \mathbb{R}^{m\times n}$, $\mathbf{b} \in \mathbb{R}^m$, and $m \geq n$, with solution $\mathbf{x}^\star$.

A practical challenge is that measurement data often becomes damaged during collection, transmission, or storage, violating consistency. Two important types of damage are

- *corruption*; large errors due to faulty software, hardware, or mismeasurement, affecting a small fraction of data, and
- *noise*; small errors due to imprecision or processing that may affect every measurement.

The Randomized Kaczmarz (RK) method [43, 75] is a popular iterative projective method for large, overdetermined, consistent systems due to its exponential convergence and low memory requirements. An initial guess $\mathbf{x}_0$ is iteratively projected onto randomly chosen hyperplanes corresponding to solution spaces to rows of the system. More precisely, letting $\mathbf{a}_1, \cdots, \mathbf{a}_m$ be the rows of $\mathbf{A}$, the kth iterate is computed as

$$\mathbf{x}_k = \mathbf{x}_{k-1} + \frac{b_i - \mathbf{a}_i \mathbf{x}_{k-1}}{\|\mathbf{a}_i\|^2} \mathbf{a}_i^\top,$$

where row i has been chosen with probability proportional to its Euclidean norm (denoted $\|\cdot\|$).

Strohmer and Vershynin [75] showed that RK converges exponentially in expectation. This was extended to the noisy case in [63] where a vector of noise $\mathbf{r}$ is added to the measurement data $\mathbf{b}$. In this case, exponential convergence is still achieved up to an error horizon depending on the size of the noise. Namely, letting $\mathbf{e}_k := \mathbf{x}_k - \mathbf{x}^\star$ be the error at the kth iteration,

$$\mathbb{E}\,\|\mathbf{e}_k\|^2 \leq \left(1 - \frac{\sigma_{\min}^2(\mathbf{A})}{\|\mathbf{A}\|_F^2}\right) \|\mathbf{e}_0\|^2 + \frac{\|\mathbf{A}\|_F^2}{\sigma_{\min}^2(\mathbf{A})} \|\mathbf{r}\|^2,$$

where $\sigma_{\min}(\mathbf{A})$ is the smallest singular value of $\mathbf{A}$, and $\|\cdot\|_F$ is the Fröbenius norm.

Variants of RK, including those involving multi-row projections or greedy row selection, have been shown to exhibit similar robustness to noise [64, 31].

Corrupted data proves more of a challenge for projection-based methods: projecting onto a row with large corruption can cause the iterate to move far from the solution and severely disrupt convergence. Recent modifications have been designed to handle this

issue, see [32, 18, 34]. In this paper we focus on the method introduced in [34] and analyzed further in [74], where the authors constructed a quantile-based modification of RK, QuantileRK, in which the quantile of the absolute values of a subresidual is used to detect and avoid projecting onto corrupted rows.

Here, we extend the theory and show that QuantileRK is robust to both corruptions and noise in the measurement data. We give a theoretical result showing exponential convergence down to an error horizon, and provide experiments demonstrating the strength of the method in identifying and solving the underlying system beneath highly damaged measurement data.

2.2 Proposed Method

2.2.1 Preliminaries & Notation

We aim to solve the consistent system $\mathbf{A}\mathbf{x} = \tilde{\mathbf{b}}$ with access only to the observed measurement vector $\mathbf{b} = \tilde{\mathbf{b}} + \mathbf{b}^C + \mathbf{r}$, where $\mathbf{b}^C$ is a sparse vector of corruptions, and $\mathbf{r}$ is a vector of noise. In practice, $\mathbf{b}^C$ will contain large entries, and $\mathbf{r}$ small, but we make no such assumption for our theory. We define β to be the fraction of data that is corrupted, i.e. $\beta = |\{i : b_i^C > 0\}|/m$.

We build on the foundations established in [34]. To utilize results from random matrix theory, we view $\mathbf{A}$ as a random matrix and make the following assumptions, that will for example hold if $\mathbf{A}$ is Gaussian with normalized rows:

Assumption 1. All rows $\mathbf{a}_i$ of $\mathbf{A}$ are independent, and $\sqrt{n}\mathbf{a}_i$ is mean zero isotropic with uniformly bounded subgaussian norm, $\|\sqrt{n}\mathbf{a}_i\|_{\psi_2} \leq K$.

Assumption 2. Each entry a_{ij} of $\mathbf{A}$ has probability density function ϕ_{ij} satisfying $\phi_{ij}(t) \leq D\sqrt{n}$ for all $t \in \mathbb{R}$.

We define the q-quantile of the absolute values of the residual, or sub-residual formed

by rows in an index set S:

$$Q_q(\mathbf{x}) = q - \text{quantile}\{|b_i - \langle \mathbf{a}_i, \mathbf{x}\rangle| : i \in [m]\}$$
$$Q_q(\mathbf{x}, S) = q - \text{quantile}\{|b_i - \langle \mathbf{a}_i, \mathbf{x}\rangle| : i \in S\}.$$

Throughout, $C, c, c_1, c_2, \cdots$ refer to absolute constants whose values may vary line by line.

2.2.2 QuantileRK

Projecting iterates onto corrupted hyperplanes will often cause abnormally large movements. Our method detects this by taking a quantile of the residual entries of a collection of rows at each iteration, and deeming a row acceptable for projection if its residual entry is less than said quantile. Whilst the method may still project onto corrupted rows, the movement away from the solution caused by these 'bad' projections will on average be outweighed by projections onto uncorrupted rows. We present pseudocode for the method in Algorithm 1, under the assumption that $\mathbf{A}$ has been standardized to have normalized rows for simplicity.

Algorithm 1 QuantileRK(q)

procedure QUANTILERK($\mathbf{A}$, b, q, t, N)
  $\mathbf{x}_0 = 0$
  for j = 1, …, N **do**
    sample $i_1, \dots i_t \sim$ Uniform$(1, \dots, m)$
    sample $k \sim$ Uniform$(1, \dots, m)$
    compute $q_k = Q_q(\mathbf{x}_{j-1}, \{i_l : l \in [t]\})$
    if $|\langle \mathbf{a}_k, \mathbf{x}_{j-1}\rangle - b_k| \leq q_k$ **then**
      $\mathbf{x}_j = \mathbf{x}_{j-1} - (\langle \mathbf{x}_{j-1}, \mathbf{a}_k\rangle - b_k)\, \mathbf{a}_k$
    else
      $\mathbf{x}_j = \mathbf{x}_{j-1}$
    end if
  end for
  return $\mathbf{x}_N$
end procedure

In [34], the authors proved that for $\mathbf{A}$ sufficiently tall and β sufficiently small, Quantil-

eRK convergences exponentially, with

$$\mathbb{E}(\|\mathbf{e}_k\|^2) \leq \left(1 - \frac{C_q}{n}\right)^k \|\mathbf{e}_0\|^2,$$

Our main result, Theorem 2.1, builds on this and shows that the addition of noise does not harm the convergence rate, and exponential convergence is still achieved up to a horizon proportional to the size of the noise.

Theorem 2.1. *Let the linear system be defined by the standardized random matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ satisfying Assumptions 1 and 2. Assume that $\beta \leq \min(cq, 1-q)$, and that $m \geq Cn$. Then with high probability, the iterates produced by QuantileRK, with $q \in (0,1)$, where in each iteration the quantile is computed using the full residual, and initialized with arbitrary $\mathbf{x}_0 \in \mathbb{R}^n$, satisfy*

$$\mathbb{E}(\|\mathbf{e}_k\|^2) \leq \left(1 - \frac{C_q}{n}\right)^k \|\mathbf{e}_0\|^2 + \frac{2n}{c_1} \|\mathbf{r}\|_\infty^2 . \tag{2.1}$$

Remark 2.2. It is natural to ask whether one may consider some of the larger entries in $\mathbf{r}$ as corruptions, by increasing β, leading to a smaller error horizon. This is possible, but there is a tradeoff: increasing β forces a decrease in q, which slows convergence. The effectiveness will be application dependent: if the distribution of noise is concentrated, it would take a significant increase in β to see a decrease in the error horizon, leading to substantially slower convergence; however, if the noise has large spikes, increasing β may be worthwhile.

2.2.3 Proof of Main Result

We follow the proof of the main QuantileRK convergence result from [34] closely, making necessary alterations for the presence of noise throughout. We firstly present a modified version of Remark 3 from said paper:

Lemma 2.3. *Let $\alpha \in (0,1]$, let the random matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ satisfy Assumption 1, and let $\mathbf{x}^\star$ be the solution to the consistent system $\mathbf{A}\mathbf{x} = \tilde{\mathbf{b}}$. Then if $m \geq n$, there exists a constant $C_K > 0$ so*

that with probability at least $1 - 2e^{-m}$*, for every* $\mathbf{x} \in \mathbb{R}^n$ *the bound*

$$|\langle \mathbf{a}_i, \mathbf{x} \rangle - b_i| \leq \frac{C_K}{\alpha\sqrt{n}} \|\mathbf{x} - \mathbf{x}^\star\| + \|r\|_\infty$$

holds for all but at most $(\alpha + \beta)m$ *indices* i*.*

Proof. Applying ([34], Proposition 2) with the unit vector $(\mathbf{x} - \mathbf{x}^\star)/\|\mathbf{x} - \mathbf{x}^\star\|$, excluding the βm corrupted rows, yields

$$|\langle \mathbf{a}_i, \mathbf{x} \rangle - \langle \mathbf{a}_i, \mathbf{x}^\star \rangle| \leq \frac{C_K}{\alpha\sqrt{n}} \|\mathbf{x} - \mathbf{x}^\star\|$$

for at most $(\alpha + \beta)m$ indices i. For each i for which the above holds, we have $\langle \mathbf{a}_i, \mathbf{x}^\star \rangle = \tilde{b}_i = b_i - r_i$ (i.e., $b_i^C = 0$). Then the right hand side can be written as

$$\begin{aligned} |\langle \mathbf{a}_i, \mathbf{x} \rangle - \langle \mathbf{a}_i, \mathbf{x}^\star \rangle| &= |\langle \mathbf{a}_i, \mathbf{x} \rangle - b_i + r_i| \\ &\geq |\langle \mathbf{a}_i, \mathbf{x} \rangle - b_i| - |r_i| \\ &\geq |\langle \mathbf{a}_i, \mathbf{x} \rangle - b_i| - \|\mathbf{r}\|_\infty . \end{aligned}$$

Combining the inequalities yields the result. □

Taking $\alpha \leq 1 - q - \beta$ immediately gives the following corollary, showing that the quantiles are well-concentrated:

Corollary 2.4. *Under the same assumptions as Lemma 2.3, and taking* $\alpha \leq 1 - q - \beta$*, we have*

$$\mathbb{P}\left(Q_q(\mathbf{x}) \leq \frac{C_\alpha \|\mathbf{x} - \mathbf{x}^*\|}{\sqrt{n}} + \|\mathbf{r}\|_\infty \right) \geq 1 - 2e^{-m}.$$

We are now ready to prove Theorem 2.1.

Proof of Theorem 2.1. Let $\mathcal{E}_{Accept}(k)$ denote the event that we sample a row that with residual less than the computed quantile at the kth iteration. It is clear that we have $\mathbb{P}(\mathcal{E}_{Accept}(k)) = q$.

Let J be a collection of indices of size $2\beta m$, containing all corrupted indices and at least βm acceptable indices. Then split all acceptable indices into two subsets: those inside J, denoted by I_1, and those outside of J, denoted by I_2. Let $\mathcal{E}_L^k$ denote the event that at the k-th iteration an index in sampled from $L \subset [m]$. We argue that the possible damage to convergence caused by projecting onto a corrupted row in I_1 is outweighed by the movement towards the solution caused by projecting onto a row in I_2.

Observe firstly that

$$\mathbb{E}_k(\|\mathbf{e}_{k+1}\|^2) = q\mathbb{E}_k(\|\mathbf{e}_{k+1}\|^2 \,|\mathcal{E}_{Accept}(k+1)) + (1-q)\,\|\mathbf{e}_k\|^2\,, \tag{2.2}$$

since we have no update to our iterate if the sampled row was not acceptable.

We now deal with $\mathbb{E}_k(\|\mathbf{e}_{k+1}\|^2 \,|\mathcal{E}_{Accept}(k+1))$ by splitting into two cases; sampling a row from I_1 or from I_2. Note that the probability of sampling an index from I_1, conditioned on $\mathcal{E}_{Accept}(k+1)$, p_J, satisfies $p_J \leq 2\beta m/qm \leq 2\beta/q$.

Firstly, if we sample from I_2, the iterate $\mathbf{x}_{k+1}$ is obtained by performing an iteration of standard RK on the noisy system $\mathbf{A}_{I_2}\mathbf{x} = \tilde{\mathbf{b}}_{I_2} + \mathbf{r}_{I_2}$. Noting that I_2 has size at least $(q-\beta)m$, Proposition 2 from [34] (with $\alpha = q-\beta$) yields that $\sigma_{min}(\mathbf{A}_{I_2}) \geq C_{\alpha,D}\sqrt{m/n}$ with high probability, provided that $\mathbf{A}$ is tall enough. Furthermore since $\mathbf{A}$ has normalized rows, we have $\|\mathbf{A}_{I_2}\|_F \geq \sqrt{(q-\beta)m}$. Thus

$$\kappa(\mathbf{A}_{I_2}) \geq C_{q,D}\sqrt{n}.$$

Then by the analysis of RK with noise in [63], we have that

$$\begin{aligned}\mathbb{E}_k(\|\mathbf{e}_{k+1}\|^2 \,|\mathcal{E}_{I_2}^{k+1}) &\leq \left(1 - \frac{c_1}{n}\right)\|\mathbf{e}_k\|^2 + \|\mathbf{r}_{I_2}\|_\infty^2 \\ &\leq \left(1 - \frac{c_1}{n}\right)\|\mathbf{e}_k\|^2 + \|\mathbf{r}\|_\infty^2\,.\end{aligned}$$

The $\beta = 0$ case (i.e., when we have no corruptions) follows immediately from this and Equation (2.2). In the case where $\beta > 0$, i.e., when I_1 is not empty, we consider the

possibility that we sample from I_1. Our update will take the form $\mathbf{x}_{k+1} = \mathbf{x}_k - h_i \mathbf{a}_i$, where $|h_i| \leq Q_q(\mathbf{x}_k)$, and so we have

$$\mathbb{E}_k(\|\mathbf{e}_{k+1}\|^2 | \mathcal{E}_{I_1}^{k+1}) \leq \|\mathbf{e}_k\|^2 + Q_q(\mathbf{x}_k)^2 + 2Q_q(\mathbf{x}_k)\mathbb{E}_k(|\langle \mathbf{e}_k, \mathbf{a}_i \rangle| i \sim \text{Unif}(I_1)).$$

To continue estimating, note that we have by ([34], Lemma 4), with probability $1 - 2e^{-cm}$,

$$\begin{aligned}\mathbb{E}_k(|\langle \mathbf{e}_k, \mathbf{a}_i \rangle| | i \sim \text{Unif}(I_1)) &= \frac{1}{|I_1|} \sum_{i \in I_1} |\langle \mathbf{e}_k, \mathbf{a}_i \rangle| \\ &\leq \frac{C \|\mathbf{e}_k\|}{\sqrt{\beta n}}.\end{aligned}$$

Then using this and the result of Corollary 2.4:

$$\mathbb{E}_k(\|\mathbf{e}_{k+1}\|^2 | \mathcal{E}_{I_1}^{k+1}) \leq \left(1 + \frac{\sqrt{\beta} c_2 + c_3}{n\sqrt{\beta}}\right) \|\mathbf{e}_k\|^2 + \left(\frac{c_4\sqrt{\beta} + c_5}{\sqrt{n\beta}}\right) \|\mathbf{r}\|_\infty \|\mathbf{e}_k\| + \|\mathbf{r}\|_\infty^2 .$$

We can now estimate $\mathbb{E}_k(\|\mathbf{e}_{k+1}\|^2 | \mathcal{E}_{Accept}(k+1))$ as follows:

$$\begin{aligned}\mathbb{E}_k(\|\mathbf{e}_{k+1}\|^2 | \mathcal{E}_{Accept}(k+1)) &= p_J \mathbb{E}_k(\|\mathbf{e}_{k+1}\|^2 | \mathcal{E}_{I_1}^{k+1}) + (1 - p_J)\mathbb{E}_k(\|\mathbf{e}_{k+1}\|^2 | \mathcal{E}_{I_2}^{k+1}) \\ &\leq \left(1 - \frac{c_1}{n} + p_J \left(\frac{\sqrt{\beta}(c_1 + c_2) + c_3}{n\sqrt{\beta}}\right)\right) \|\mathbf{e}_k\|^2 + \\ &\quad p_J \left(\frac{c_4\sqrt{\beta} + c_5}{\sqrt{n\beta}}\right) \|\mathbf{r}\|_\infty \|\mathbf{e}_k\| + \|\mathbf{r}\|_\infty^2 .\end{aligned}$$

To handle the $\|\mathbf{r}\|_\infty \|\mathbf{e}_k\|$ term, we split into two cases. The motivation is that when our error is large relative to the noise, the quantile can detect corruptions well, whereas when the error is small relative to the noise, our movement will be small. Firstly, if $\sqrt{n}\|\mathbf{r}\|_\infty \leq \|\mathbf{e}_k\|$ (i.e. when our error is large), we have

$$\begin{aligned}\mathbb{E}_k(\|\mathbf{e}_{k+1}\|^2 | \mathcal{E}_{Accept}(k+1)) &\leq \left(1 - \frac{c_1}{n} + p_J \left(\frac{\sqrt{\beta}(c_1 + c_2 + c_4) + c_3 + c_5}{n\sqrt{\beta}}\right)\right) \|\mathbf{e}_k\|^2 + \|\mathbf{r}\|_\infty^2 \\ &\leq \left(1 - \frac{0.5c_1}{n}\right) \|\mathbf{e}_k\|^2 + \|\mathbf{r}\|_\infty^2\end{aligned}$$

for small enough β (we need $\sqrt{\beta} \leq cq$). On the other hand, if $\sqrt{n}\,\|\mathbf{r}\|_\infty \geq \|\mathbf{e}_k\|$, we have

$$\begin{aligned}\mathbb{E}_k(\|\mathbf{e}_{k+1}\|^2 \,|\mathcal{E}_{Accept}(k+1)) &\leq \left(1 - \frac{c_1}{n} + p_J \frac{\sqrt{\beta}(c_1+c_2)+c_3}{n\sqrt{\beta}}\right)\|\mathbf{e}_k\|^2 + p_J\left(\frac{c_4\sqrt{\beta}+c_5}{\sqrt{\beta}}\right)\|\mathbf{r}\|_\infty^2 \\ &\leq \left(1 - \frac{0.5c_1}{n}\right)\|\mathbf{e}_k\|^2 + \|\mathbf{r}\|_\infty^2,\end{aligned}$$

again for $\sqrt{\beta} \leq cq$ sufficiently small.

We may now substitute our expressions into Equation (2.2) to obtain our per-iteration guarantee:

$$\mathbb{E}_k(\|\mathbf{e}_{k+1}\|^2) \leq \left(1 - \frac{0.5qc_1}{n}\right)\|\mathbf{e}_k\|^2 + q\,\|\mathbf{r}\|_\infty^2.$$

By induction, we obtain our overall guarantee:

$$\begin{aligned}\mathbb{E}(\|\mathbf{e}_k\|^2) &\leq \left(1 - \frac{0.5qc_1}{n}\right)^k \|\mathbf{e}_0\|^2 + \sum_{j=0}^{k-1}\left(1 - \frac{0.5qc_1}{n}\right)^j q\,\|\mathbf{r}\|_\infty^2 \\ &\leq \left(1 - \frac{0.5qc_1}{n}\right)\|\mathbf{e}_0\|^2 + \frac{2n}{c_1}\|\mathbf{r}\|_\infty^2.\end{aligned}$$

□

2.2.4 Experimental Results

Experiments are performed on 2000×100 standardized Gaussian matrices $\mathbf{A}$. We sample a Gaussian $\mathbf{x}^* \in \mathbb{R}^{100\times 1}$, compute $\mathbf{b} = \mathbf{A}\mathbf{x}$, and then corrupt a fraction β of the rows of $\mathbf{b}$ by adding corruptions of size to be specified. We add noise $\mathbf{r} \in \mathbb{R}^{2000\times 100}$ with Uniform$(-0.02, 0.02)$ entries, and apply QuantileRK to the resulting system. At each iteration 400 rows are sampled, from which the subresidual is computed.

In Figure 2.1 we take $q = 0.7$, $\beta = 0.2$, and corrupt the already noisy system with corruptions taken from Uniform$(-k, k)$ for a range of k. We see that when corruptions are large relative to the noise, they are better detected by the quantile, faster convergence is

achieved. When corruptions are small, they do not disrupt convergence enough to break the method, and convergence is achieved down to the error horizon.

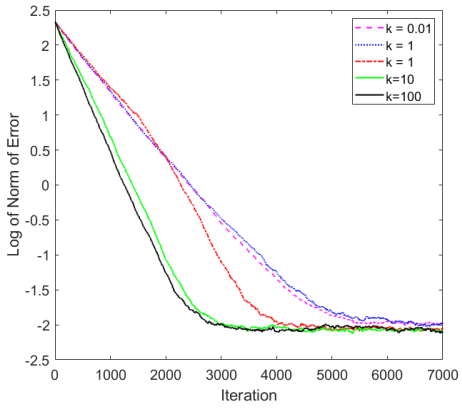


Figure 2.1: Convergence of QuantileRK(0.7) applied to normalized 2000×100 Gaussian systems with Uniform$(-k, k)$ corruptions, for a range of k, and Uniform$(-0.02, 0.02)$ noise. The logarithm of the error is plotted for each iteration to show the linearity of the convergence.

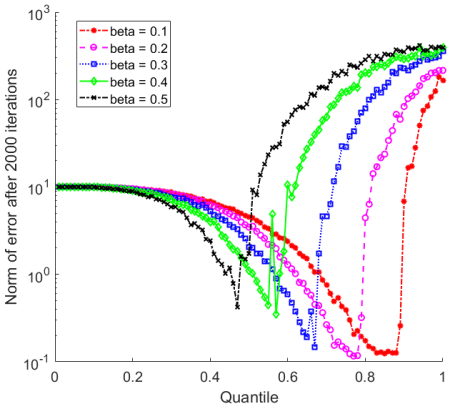


Figure 2.2: Norm of the error after 2000 iterations, $\|\mathbf{x}_{2000} - \mathbf{x}^\star\|$, of QuantileRK$(q)$ applied to normalized 2000×100 Gaussian systems, for a range of corruption rates β and quantile choices q. Corruptions are Uniform$(-100, 100)$ and are added to uniformly random rows. Uniform$(-0.02, 0.02)$ noise is applied to all rows.

We would like to take q as large as possible so that we may sample rows yielding large movement, but we must take $q < 1 - \beta$ to avoid corrupted rows. In Figure 2.2 we plot the normed error after 2000 iterations for a range of q and β, and we see that we can be very aggressive with our choice of q: we are able to take it very close to $1 - \beta$, and should do so

to accelerate convergence.

In Figure 2.3 we simulate 100 trials, and compare the error (after 5000 and 10000 iterations respectively) to the predicted horizon. Indeed, our results show that the predicted horizon is closely respected.

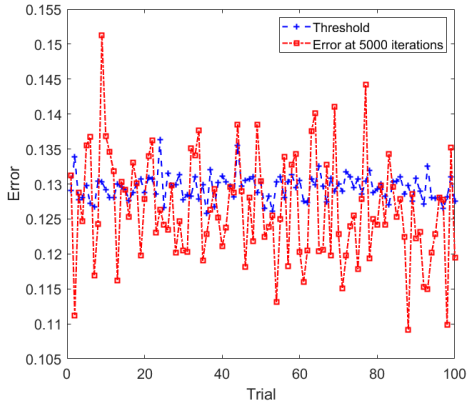


Figure 2.3: Comparing the error after 5000 iterations of QuantileRK(0.7), $\|\mathbf{x}_{5000} - \mathbf{x}^\star\|$, with the error horizon predicted by Theorem 2.1. One 2000×100 normalized Gaussian system is generated for each trial, we take $\beta = 0.2$ and add Uniform$(-100, 100)$ corruptions to rows selected uniformly at random.

2.2.5 Conclusion and Future Work

We have shown, both theoretically and empirically, that QuantileRK is a powerful method for solving linear systems where measurement data has been damaged by both corruptions and noise. We believe that this method will prove tractable in practice, as corruption and noise are ubiquitous in real-world data.

We are interested in pursuing quantile-based modifications to other projection-based iterative methods, see [28] for a general framework, and also in relaxing the conditions placed on our system: see [74] for some work in this direction.

CHAPTER 3

Block Accelerations of QuantileRK

This chapter is a version of [11] and is joint with with Lu Cheng, Liza Rebrova, and Prof. Deanna Needell. Deanna Needell and Liza Rebrova proposed and co-supervised the project. Lu Cheng and I contributed the codebase and experiments. Liza Rebrova and I contributed the convergence analysis. We continue on the topic of solving systems of linear equations affected by sparse corruptions by analyzing variants of QuantileRK that make use of multiple equations at each iteration. These variants require very little extra computation time (and in fact, much of the computation may be performed in parallel), and lead to significantly faster convergence. We show this both in theory and empirically, as well as including experiments designed to aid practitioners in parameter selection.

3.1 Introduction

Let $\mathbf{A} \in \mathbb{R}^{m \times n}$, $\mathbf{b} \in \mathbb{R}^m$, and suppose we wish to find $\mathbf{x} \in \mathbb{R}^n$ such that $\mathbf{A}\mathbf{x} = \mathbf{b}$. Such linear systems are ubiquitous across applied mathematics and the sciences, arising in contexts ranging from medical imaging [60, 39] to machine learning [3], sensor networks [72], and more. A common and widely studied approach is to seek the least squares solution $\mathbf{x}_{\mathrm{LS}} = \operatorname{argmin} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|$, for which many methods have been devised.

In this paper, we consider the related problem of trying to solve a consistent system $\mathbf{A}\mathbf{x} = \mathbf{b}^t$, where $\mathbf{A}$ is of full rank, and whose solution is $\mathbf{x}^\star$. Here, however, instead of observing the *true* right hand side $\mathbf{b}^t$ one observes a *corrupted* version, $\mathbf{b} = \mathbf{b}^t + \mathbf{b}^c$, where $\mathbf{b}^c$ represents a vector of corruptions. In this setting, $\mathbf{x}_{\mathrm{LS}}$ may be far from $\mathbf{x}^\star$, rendering least squares solvers unsuitable. Frequently, such systems are highly overdetermined, with

$m \gg n$, for example in settings where one has many more measurements than covariates. In this case, it is reasonable to hope to recover $\mathbf{x}^\star$ as long as a sufficiently small fraction of rows are corrupted. Indeed, we assume that corruptions may be of arbitrary size and location, but affect only some fraction $\|\mathbf{b}^c\|_0 / m := \beta \in [0, 1)$ of data points. We refer to a row with a corrupted right hand side entry as a *corrupted row*.

This model covers a wide variety of scenarios in which data may suffer corruptions during collection, transmission, storage, or otherwise. As one example, a frequent setting in which overdetermined linear systems appear is that of computerized tomography: in this case, each row of the system represents the absorption of a single X-ray beam through a medium, and solving the system recovers an image of said medium. A small number of beams malfunctioning may lead to catastrophic errors of arbitrary size in the resulting data, but as long as the number of such errors is a small one may still hope to recover the underlying solution to the uncorrupted system. Similar situations may arise in sensor networks from malfunctioning sensors, or error correcting codes from transmission errors. Note that typical methods for the least squares problem are unsuitable in this setting, as with arbitrarily large corruptions the least squares solution $\mathbf{x}_{\mathrm{LS}}$ may be far from $\mathbf{x}^\star$ (even if β is small); this is contrary to the widely-studied *noisy* setting, in which one assumes that every data point may be damaged by some small amount of noise, but that the least squares solution is still an accurate estimation of the solution.

This sparse corruption model is well-studied within the error-correction and compressed sensing literature: see [8, 20, 23]. However, such methods often require loading the entire system into memory; a requirement that is frequently impractical or impossible in settings where the system is large-scale, such as those systems arising in medical imaging applications [39]. Recent works [32, 34, 74] have introduced novel approaches that in fact require only loading small portions (even single rows) of the system into memory at any time, whilst achieving linear convergence even in the presence of large – or adversarially located – corruptions.

In this work, we introduce a new iterative solver for corrupted linear systems, QuantileABK, building upon the averaged block Kaczmarz method introduced in [61] and the

quantile-based variant of randomized Kaczmarz, QuantileRK, introduced in [30]. As with many iterative methods in the Kaczmarz family, QuantileABK relies on *residual* information to determine the step size. The residual at the iterate $\mathbf{x}_k$ is the vector of distances from $\mathbf{x}_k$ to the hyperplanes defined by the rows of the matrix $\mathbf{A}$, that is, $\mathbf{b} - \mathbf{A}\mathbf{x}_k$. The standard randomized Kaczmarz method, on a consistent uncorrupted system, makes steps in the directions of projections to the individual hyperplanes of length equal to the corresponding residual entry. The underlying idea of the QuantileRK method is that large residual components suggest (a) potential corruptions and (b) large and potentially unstable next iteration steps. So, statistics of the absolute values of the residual entries are used to select trustworthy directions and only use them. We give more detailed backgrounds to each of the aforementioned prior methods in Section 3.1.3. An important inefficiency of QuantileRK is that despite the entire residual being computed to detect corruptions, only a single row is used to compute the next iterate. Our method instead leverages the information gained from the residual with a more complex projection step to take a highly over-relaxed step size, leading to a huge acceleration in convergence over the single-row method QuantileRK [30].

We prove several convergence results for the proposed method. For an example of the acceleration our method brings, here is a simplified restatement of one of the results that holds for a particular class of random matrices:

Theorem 3.1 (Informal restatement of Theorem 3.9)**.** *Assume that $\mathbf{A} \in \mathbb{R}^{m\times n}$ satisfies a certain random matrix model (see Section 3.1.3.4) and has sufficiently large aspect ratio m/n. Suppose then that the system $\mathbf{A}\mathbf{x} = \mathbf{b}$ has a fraction β of corrupted rows, with β sufficiently small. Then with high probability, the iterates produced by applying QuantileABK (see Algorithm 2) to this system satisfy*

$$\|\mathbf{x}_k - \mathbf{x}^\star\|^2 \leq (1-c)^k \|\mathbf{x}_0 - \mathbf{x}^\star\|^2,$$

where c depends only on a user-chosen quantile parameter (in particular, c is independent of m and n).

This result may be compared to ([34], Theorem 1), to see that our method converges

faster than QuantileRK in this setting by a factor linear in n, the number of columns of the system. Moreover, our method has a computational cost of the same order, with the most significant cost in both methods being the computation of the residual. We note that this acceleration in convergence occurs also in the uncorrupted case (i.e., when $\beta = 0$). See Section 3.2 for the formal description of the algorithm and further discussion, and Section 3.1.4 for all theorem statements. Notably, we do not restrict ourselves to the random matrix setting: as in [74], we show a general guarantee of linear convergence, with a rate depending on the spectral properties of $\mathbf{A}$ and its row submatrices (Theorem 3.7).

The idea to leverage several equations to speed up Kaczmarz methods is not new, it is in the core of a sequence of Block Kaczmarz methods, including [21, 64, 66, 61]. However, not all of them are equally extendable to the corrupted framework. The focus of this work is to discriminate between block Kaczmarz accelerations in terms of their provable robustness to adversarial corruptions: see additional discussion in Sections 3.1.3.2 and 3.4.3.

3.1.1 Organization

The remainder of the paper is organized as follows. In Section 3.1.2 we introduce notation used throughout the paper. In Section 3.1.3 we give a detailed background for previous methods upon which our method is built, and in Section 3.1.4 we give a summary of our main results. Section 3.2 contains a description of our proposed method, and Section 3.3 contains our theoretical results. In Section 3.4 we demonstrate our method in a range of experiments, and finally in Section 3.5 we conclude and offer ideas for future directions.

3.1.2 Notation

For a matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$, we denote its rows by $\mathbf{a}_i \in \mathbb{R}^n$, $i \in [m]$. For a collection of indices $\tau \subseteq [m]$, we let $\mathbf{A}_\tau$ denote the matrix obtained from $\mathbf{A}$ by restricting to rows indexed by τ. We denote the operator norm of $\mathbf{A}$ by $\|\mathbf{A}\|$, and the Fröbenius norm by $\|\mathbf{A}\|_F$. For a vector $\mathbf{v}$ we denote its Euclidean norm by $\|\mathbf{v}\|$. For a matrix $\mathbf{A}$ we denote its largest singular value by $\sigma_{\max}(\mathbf{A})$, and smallest by $\sigma_{\min}(\mathbf{A})$. When the matrix at hand is clear, we abbreviate these

to $\sigma_{\max}$ and $\sigma_{\min}$.

In sections where we view $\mathbf{A}$ as an instance of a certain family of random matrices, we use some definitions from probability. Namely, for a real-valued random variable X, we denote its subgaussian norm by $\|X\|_{\Psi_2} := \inf\{t > 0 : \mathbb{E}(\exp(X^2/t^2)) \leq 2\}$. For a random vector $\mathbf{v} \in \mathbb{R}^n$, its subgaussian norm is defined as $\|\mathbf{v}\|_{\Psi_2} := \sup_{\mathbf{x} \in S^{n-1}} \|\langle \mathbf{v}, \mathbf{x} \rangle\|_{\Psi_2}$. A random variable is said to be subgaussian if it has finite subgaussian norm. Lastly, a random vector $\mathbf{v} \in \mathbb{R}^n$ is said to be isotropic if $\mathbb{E}(\mathbf{v}\mathbf{v}^\top) = \mathbf{I}$, where $\mathbf{I}$ denotes an appropriately-sized identity matrix.

We will frequently make use of a quantile of the absolute residual. For $q \in [0, 1]$ and $\mathbf{x} \in \mathbb{R}^n$, we denote the q^{th} quantile of the (corrupted) absolute residual $|\mathbf{A}\mathbf{x} - \mathbf{b}|$ by

$$Q_q(\mathbf{x}) := \text{q}^{\text{th}} \text{ quantile of } \{|\langle \mathbf{a}_i, \mathbf{x} \rangle - b_i| : i \in [m]\},$$

recalling that the q^{th} quantile of a multiset S is the $\lceil qS \rceil^{\text{th}}$ smallest element of S.

Lastly, we use $C, c, c_1, \cdots$ to denote absolute constants that may vary from line to line. Subscripts are used to denote dependence on particular quantities, e.g. C_q denotes an absolute constant depending on q.

3.1.3 Background & Related Work

3.1.3.1 Randomized Kaczmarz

The Kaczmarz method [43] (later rediscovered for use in computerized tomography as the Algebraic Reconstruction Technique [39]) is a popular iterative method for solving overdetermined consistent linear systems. An arbitrary initial iterate $\mathbf{x}_0$ is projected sequentially onto the hyperplanes corresponding to rows of the system $\mathbf{A}\mathbf{x} = \mathbf{b}$, so that at the k^{th} iteration the update has the form

$$\mathbf{x}_k = \mathbf{x}_{k-1} - \frac{\mathbf{a}_i^\top \mathbf{x}_{k-1} - b_i}{\|\mathbf{a}_i\|^2} \mathbf{a}_i,$$

where $i = k \bmod m$. Whilst convergence to $\mathbf{x}_0$ is guaranteed via a simple application of Pythagoras's theorem, quantitative convergence guarantees proved elusive. In the landmark paper [75], the authors proved a linear convergence guarantee when rows are selected at random according to a particular distribution. Namely, in their randomized Kaczmarz method, at iteration k row i is selected with probability $\|\mathbf{a}_i\|^2 / \|\mathbf{A}\|_F^2$, and the update takes the same form as above. This row selection scheme gave rise to Theorem 3.2.

Theorem 3.2 (Strohmer & Vershynin, 2009)**.** *Suppose that* $\mathbf{Ax} = \mathbf{b}$ *is consistent with solution* $\mathbf{x}^*$*. Then the iterates produced by applying randomized Kaczmarz to this system satisfy:*

$$\mathbb{E}\left(\|\mathbf{x}_k - \mathbf{x}^\star\|^2\right) \leq \left(1 - \frac{\sigma_{\min}^2}{\|\mathbf{A}\|_F^2}\right)^k \|\mathbf{x}_0 - \mathbf{x}^\star\|^2 .$$

This result spurred a boom in related research, including Kaczmarz variants with differing row selection protocols [73, 31, 2], block update methods [21, 71, 64], and adaptive methods [29]. Our method is motivated by block methods in particular, which we proceed to discuss in more detail.

3.1.3.2 Block Kaczmarz Methods

Variants of the Kaczmarz method that make use of more than a single row at each iteration, often referred to as block methods, have been extensively studied. Two particular methodologies have proven popular:

- *projective* block methods, in which at each iteration the iterate is projected onto the subspace defined by an entire block of rows [64, 71, 21], and
- *averaged* block methods, in which at each iteration the projections of the previous iterate onto each individual row in a block are computed and then averaged [61, 59].

Consider first the projective methodology. It has been shown the projective block Kaczmarz significantly outperforms randomized Kaczmarz [64], particularly in the case when the system has coherent rows [65]. Each iteration of the projective algorithm computes the

best possible update given the information from the considered block, however, block projections are known to be significantly less stable for more sophisticated tasks, for example linear feasibility problems [5].

The presence of corruptions may also significantly disrupt projective block variants. Whilst in the single-row setting the quantile statistic is able to control the potential harm caused by projecting onto a corrupted row, a block containing a corrupted row may yield a projection that is arbitrarily far from the true solution. To some extent, this issue can be alleviated by posing an assumption of row *incoherence*: that every two rows are not nearly parallel, i.e., their normal vectors have small scalar products. Informally, this results in the intersection subspaces being "close enough" to individual projection points due to non-trivial angles between the solution hyperplanes for individual equations. The incoherence condition is implicitly needed in the existing non-block QuantileRK results [34, 74] to ensure that the quantile statistic is representative. In this work, it also appears in the form of a restricted smallest singular value, discussed below.

However, the incoherence assumption does not resolve the second deficiency of projective block methods applied to corrupted systems. Namely, a residual-based criterion for deciding if a certain equation is trustworthy or corrupted cannot guarantee to identify all corrupted equations: for example, a current iterate might satisfy some corrupted equation exactly. Projecting onto a block containing a corrupted equation keeps the iterate inside its corrupted (shifted) hyperplane. Finally, when increasing block size, one rapidly increases the chance of an adversarial setting in which the majority of the blocks contain at least one corrupted row. A concrete adversarial construction for projective block methods is discussed in Section 3.4.3.

Given the lack of robustness of projective block methods, we focus in this work on modifying an averaged block Kaczmarz method introduced by Necoara [61] and also considered in [59]. For a consistent system $\mathbf{A}\mathbf{x} = \mathbf{b}$ with solution $\mathbf{x}^\star$, at the k^{th} iteration a block of row indices τ_k is selected from a distribution $\mathcal{D}$ on $[m]$. Then, the projections of $\mathbf{x}_{k-1}$ onto each row in τ are computed and averaged, possibly in a weighted fashion. A step of size α_k – potentially dependent on the iteration – is then taken in this averaged direction.

The update is thus given by

$$\mathbf{x}_k = \mathbf{x}_{k-1} - \alpha_k \sum_{i \in \tau_k} w_i^k \frac{\mathbf{a}_i^\top \mathbf{x}_{k-1} - b_i}{\|\mathbf{a}_i\|^2} \mathbf{a}_i, \quad \text{with weights } w_i^k \text{ such that } \sum_{i \in \tau_k} w_i^k = 1.$$

The method may be found in full as Algorithm 4.1 in [61], and we refer to it as AveragedRBK. The convergence of AveragedRBK depends on the spectra of the row submatrices formed by sampled blocks. Indeed, the key quantity

$$\sigma^2_{\mathcal{D},\max} := \max_{\tau \sim \mathcal{D}} \sigma^2_{\max}(\mathbf{A}_\tau),$$

the largest singular value of any row-submatrix with rows sampled from $\mathcal{D}$.

Necoara's framework allows many freedoms: in row selection strategy, weighting scheme, and step size. Specializing to the particular case of uniformly weighted rows, a constant (optimized) step size, and fixed block size (but without restraint on other aspects of $\mathcal{D}$), the following convergence result holds.

Theorem 3.3 (Necoara, 2019)**.** *Suppose that the system $\mathbf{Ax} = \mathbf{b}$ is consistent with solution $\mathbf{x}^\star$, and that $\mathbf{A}$ has been normalized such that each row has unit norm. Then the iterates produced by applying AveragedRBK with block size $|\tau|$, step size $\frac{|\tau|}{\sigma^2_{\mathcal{D},\max}}$, and row weights $1/|\tau|$, satisfy*

$$\mathbb{E}\left(\|\mathbf{x}_k - \mathbf{x}^\star\|^2\right) \leq \left(1 - \frac{|\tau|\sigma^2_{\min}}{m\sigma^2_{\mathcal{D},\max}}\right)^k \|\mathbf{x}_0 - \mathbf{x}^\star\|^2 .$$

We include this particular result as it allows for easier comparison with other methods, but we refer the reader to [61] for more general results. In particular, we see that under the setup of Theorem 3.3, AveragedRBK achieves an improvement in convergence rate by a factor of $|\tau|/\sigma^2_{\mathcal{D},\max}$ compared to RK (recall Theorem 3.2). This is greater than one in most sensible cases, for instance if $\operatorname{rank}(\mathbf{A}_\tau) \geq 2$ for all τ, and will represent a significant speedup in cases where the sampled blocks are well-conditioned. We refer to Section 4.3 of [61] for further details. Furthermore, we note that the accelerated convergence rate does not necessarily come with greater computation time as the individual row projections may

be performed in parallel: see [59].

3.1.3.3 Kaczmarz Variants for Least Squares

Research on randomized Kaczmarz and its variants originated in the setting of a consistent, full rank system. Since then, convergence results have been extended to the rank-deficient (but still consistent) case for randomized Kaczmarz [86] and projective block Kaczmarz [33]. Generalizing results and methods to the inconsistent setting has also been an area of interest, for example in [63] the author shows the following result in the setting of a *noisy* system.

Theorem 3.4 (Needell, 2010)**.** *Suppose that* $\mathbf{Ax} = \mathbf{b}$ *is a consistent system with solution* $\mathbf{x}^\star$*, and that* $\mathbf{r}$ *is some vector of noise. Then the iterates produced by applying randomized Kaczmarz to the system* $\mathbf{Ax} = \mathbf{b} + \mathbf{r}$ *satisfy*

$$\mathbb{E}\left(\|\mathbf{x}_k - \mathbf{x}^\star\|^2\right) \leq \left(1 - \frac{\sigma_{\min}^2}{\|A\|_F^2}\right)^k \|\mathbf{x}_0 - \mathbf{x}^\star\|^2 + \frac{n}{\sigma_{\min}^2} \|\mathbf{r}\|_\infty^2 .$$

This result shows that randomized Kaczmarz is guaranteed to converge at the same rate as for a consistent system, but only up to some error horizon. Similar results, of convergence to a horizon, have been shown for projective block Kaczmarz [64], averaged block Kaczmarz [59], and other variants [31].

Other works have developed methods that converge all the way to the least squares solution [9, 86]. For example, in randomized extended Kaczmarz [86], randomized Kaczmarz is applied simultaneously to the systems $\mathbf{A}^\top \mathbf{z} = 0$ and $\mathbf{Ax} = \mathbf{b} - \mathbf{z}$, with the $\mathbf{z}$ and $\mathbf{x}$ iterates converging to $\mathbf{b}_{\mathrm{Im}(\mathbf{A})^\perp}$ and $\mathbf{x}_{\mathrm{LS}}$ respectively. Recent works have expanded this idea to both projective and averaged block variants [66, 16]. However, as noted previously, such methods are unsuitable in the sparse corruption model as $\mathbf{x}_{\mathrm{LS}}$ may be a poor approximation of the true solution $\mathbf{x}^*$. We discuss previous works in this direction next.

3.1.3.4 Quantile Randomized Kaczmarz

The first study of Kaczmarz methods for the sparse corruption model may be found in [32], in which the authors make use of the notion that corrupted rows are likely to have larger residual entries, as their corresponding hyperplane is displaced far from both the current iterate and true solution. Through applying several rounds of Kaczmarz-type iterations, such corrupted rows may be detected with high probability. However, the method requires severe restrictions on the number of corrupted rows. In particular, the method does not support the sparse corruption model we consider here, in which the number of corruptions scales linearly with the number of rows.

In [34], the authors expand on this residual-based heuristic and introduce a quantile-based modification of randomized Kaczmarz, QuantileRK, which also attempts to detect and avoid projecting onto corrupted rows. A sample of rows is taken and a quantile of the resultant (absolute) subresidual is computed, and then one further row is sampled. If this sampled row has absolute residual entry below the quantile, it is deemed acceptable for projection, otherwise the iterate remains unchanged. The algorithm is given in full in [34] as Method 1.

Whilst extensive experiments in [34] indicate the effectiveness of QuantileRK for a variety of systems, corruption models, and very high corruption rates (values of β up to 0.5), the authors require significant restrictions on the matrix $\mathbf{A}$ for their theoretical results. In particular, they assume a random matrix heuristic, captured in the following definition.

Definition. (Subgaussian-type systems) Let $\mathbf{A} \in \mathbb{R}^{m \times n}$ be a random matrix. We say that $\mathbf{A}$ is of subgaussian-type if all of the following hold:

(1) $\|\mathbf{a}_i\| = 1$ for all $i \in [m]$.

(2) $\sqrt{n}\mathbf{a}_i$ is mean-zero and isotropic for all $i \in [m]$.

(3) For some $K > 0$, $\|\sqrt{n}\mathbf{a}_i\|_{\Psi_2} \leq K$ for all $i \in [m]$.

(4) For some $D > 0$, every entry a_{ij} of $\mathbf{A}$ has density function uniformly bounded by $D\sqrt{n}$.

Here, K and D are absolute constants independent form the size of the matrix.

These conditions are satisfied, for example, by a matrix whose rows are sampled uniformly from the unit sphere in $\mathbb{R}^n$. With these constraints, in [34] the authors prove the following high-probability linear convergence guarantee, without placing any restriction on the size or placement of corruptions but with an additional requirement that $\mathbf{A}$ be sufficiently tall.

Theorem 3.5 (Haddock et al., 2022)**.** *Assume that* $\mathbf{A}$ *is of subgaussian-type, and that the system* $\mathbf{Ax} = \mathbf{b}$ *has a fraction* β *of corrupted rows. Then with high probability, the iterates produced by QuantileRK with* $t = m$ (*i.e., the full residual is computed at each iteration*) *applied to this system satisfy*

$$\mathbb{E}\left(\|\mathbf{x}_k - \mathbf{x}^\star\|^2\right) \leq \left(1 - \frac{C_q}{n}\right)\|\mathbf{x}_0 - \mathbf{x}^\star\|^2,$$

for some constant C_q*, so long as* $\beta \leq \min(c_1 q^2, 1-q)$ *and* $m \geq Cn$*.*

In [74], Steinerberger sought to generalize the theory behind QuantileRK beyond the random matrix setting. Indeed, he distilled the critical controls that the random matrix heuristic provides to conditions on the quantity

$$\sigma^2_{q-\beta,\min} := \inf_{\tau \subset [m], |\tau| = (q-\beta)m} \sigma^2_{\min}(\mathbf{A}_\tau). \tag{3.1}$$

Whilst assuming that $\mathbf{A}$ is of subgaussian-type allows for estimations of $\sigma^2_{q-\beta,\min}$ (see [34], Proposition 1), one may also give a much more general convergence result in terms of this quantity, albeit with stricter relative conditions on q and β.

Theorem 3.6 (Steinerberger, 2022)**.** *Suppose that* $\mathbf{Ax} = \mathbf{b}$ *has a fraction* β *of corrupted rows. Then for* $\beta < q < 1 - \beta$*, if*

$$\frac{q}{q-\beta}\left(\frac{2\sqrt{\beta}}{\sqrt{1-q-\beta}} + \frac{\beta}{1-q-\beta}\right) < \frac{\sigma^2_{q-\beta,\min}}{\sigma^2_{\max}},$$

then the iterates of QuantileRK(q) with $t = m$ applied to this system satisfy

$$\mathbb{E}\left(\|\mathbf{x}_k - \mathbf{x}^\star\|^2\right) \leq \left(1 - c_{\mathbf{A},\beta,q}\right)^k \|\mathbf{x}_0 - \mathbf{x}^\star\|^2,$$

where

$$c_{\mathbf{A},\beta,q} = (q-\beta)\frac{\sigma^2_{q-\beta,\min}}{q^2 m} - \frac{\sigma^2_{\max}}{qm}\left(\frac{2\sqrt{\beta}}{\sqrt{1-q-\beta}} + \frac{\beta}{1-q-\beta}\right) > 0.$$

Informally, the convergence rate is good if the uniform restricted smallest singular value $\sigma^2_{q-\beta,\min}$ is well-separated from zero, which itself may be viewed as a version of the incoherence assumption mentioned above in Section 3.1.3.2. Indeed, a row subsystem $\mathbf{A}_\tau$ with nearly parallel rows is nearly degenerate and $\sigma_{\min}(\mathbf{A}_\tau)$ is very small. On the other hand, independent subgaussian rows are nearly mutually orthogonal with high probability (see, e.g., [79]) and have $\sigma_{\min}(\mathbf{A}_\tau) = O(\tau/n)$ when $\tau \gg n$. Further discussion in [74] aids in understanding the relative condition on q, β, and $\sigma^2_{q-\beta,\min}$ of Theorem 3.6, as well as drawing connections to the random matrix case studied in [34].

3.1.4 Summary of Main Results

We introduce a new method, quantile averaged block Kaczmarz (QuantileABK), that applies the quantile-based techniques of QuantileRK to the averaged block Kaczmarz method. Namely, at each iteration a sample of rows is taken, the quantile of the corresponding subresidual is computed, and then an iteration of averaged block Kaczmarz is performed using *every* row with residual entry below the quantile. We defer a full explanation of the method to Section 3.2, including discussions on appropriate weights and step sizes.

Theorem 3.7 shows that our method is guaranteed to converge at least linearly as long as q, β and $\sigma^2_{q-\beta,\min}$ satisfy a similar constraint to that in Theorem 3.6, without any assumption of randomness on $\mathbf{A}$ (but still upholding the assumptions of full rank and unit norm rows). The proof of Theorem 3.7 can be found in Section 3.3.2.

Theorem 3.7. *Let $A \in \mathbb{R}^{m\times n}$ be of full rank with unit-norm rows. Suppose that the system $\mathbf{Ax} = \mathbf{b}$*

has a fraction β of corrupted entries, and that $\beta < q < 1 - \beta$. If

$$\frac{\sqrt{\beta}}{\sqrt{1-q-\beta}} < \frac{\sigma^2_{q-\beta,\min}}{\sigma^2_{\max}},$$

then the iterates of QuantileABK(q), using a theoretically optimal step size, applied to this system satisfy

$$\|\mathbf{x}_k - \mathbf{x}^\star\|^2 \leq \left(1 - \frac{c_1^2}{4c_2}\right)^k \|\mathbf{x}_0 - \mathbf{x}^\star\|^2,$$

where

$$c_1 = \frac{2\sigma^2_{q-\beta,\min}}{qm} - \frac{2\sqrt{\beta}\sigma^2_{\max}}{qm\sqrt{1-q-\beta}}, \qquad c_2 = \frac{\sigma^2_{\max}\sigma^2_{q-\beta,\min}}{q^2m^2} - \frac{2\sqrt{\beta}\sigma^2_{\max}\sigma^2_{q-\beta,\min}}{q^2m^2\sqrt{1-q-\beta}} + \frac{\beta\sigma^4_{\max}}{q^2m(1-q-\beta)}.$$

The constants c_1, c_2 are difficult to interpret, so we include a different viewpoint in Corollary 3.8 (also proved in Section 3.3.2) to give a better idea of scaling.

Corollary 3.8. *If in Theorem 3.7 we choose q such that for some $\epsilon \in [0, 1)$*

$$\frac{\sqrt{\beta}}{\sqrt{1-q-\beta}} = \epsilon\frac{\sigma^2_{q-\beta,\min}}{\sigma^2_{\max}},$$

then the optimal step size may be expressed as

$$\alpha_{opt} = \frac{qm(1-\epsilon)}{\sigma^2_{\max} - \epsilon(2-\epsilon)\sigma^2_{q-\beta,\min}}, \tag{3.2}$$

and we have the following convergence guarantee:

$$\|\mathbf{x}_k - \mathbf{x}^\star\|^2 \leq \left(1 - \frac{(1-\epsilon)^2\sigma^2_{q-\beta,\min}}{\sigma^2_{\max} - \epsilon(2-\epsilon)\sigma^2_{q-\beta,\min}}\right)^k \|\mathbf{x}_0 - \mathbf{x}^\star\|^2.$$

In general, the quantity $\sigma^2_{q-\beta,\min}$ is hard to estimate - both theoretically and empirically, particularly for very tall $\mathbf{A}$. By specializing to the case of $\mathbf{A}$ being of subgaussian-type (recall Section 3.1.3.4) we utilize results from [34] to estimate $\sigma^2_{q-\beta,\min}$ and obtain Theorem 3.9,

a formal statement of the earlier Theorem 3.1. The proof of Theorem 3.9 can be found in Section 3.3.3.

Theorem 3.9. *Let A be a random matrix satisfying Section 3.1.3.4 with constants K and D. Suppose then that the system $\mathbf{Ax} = \mathbf{b}$ has a fraction β of corrupted entries, with $\beta < q < 1 - \beta$ and we have*

$$0 \leq \epsilon < 1, \quad \text{where } \epsilon := \frac{\beta}{C_3(1-q+\beta)(q-\beta)^6}, \tag{3.3}$$

and C_3 is an absolute constant depending only on the distribution of the rows of $\mathbf{A}$. Suppose furthermore that $\mathbf{A}$ has sufficiently large aspect ratio,

$$\frac{m}{n} > C_4 \frac{1}{q-\beta} \log \frac{DK}{q-\beta}.$$

Then the optimal step size for QuantileABK(q) is

$$\alpha_{opt} = c_{\epsilon,q,\beta} n, \tag{3.4}$$

where $c_{\epsilon,q,\beta}$ is a constant depending on ϵ, q, β. Moreover, with probability at least $1 - c\exp(-c_q m)$ the iterates of QuantileABK(q), using the step size given in Equation (3.4)*, satisfy*

$$\|\mathbf{x}_k - \mathbf{x}^\star\|^2 \leq (1 - C_q)^k \|\mathbf{x}_0 - \mathbf{x}^\star\|^2,$$

where c_q, C_q depend only on q (in particular, they are independent of m and n).

As a concrete example, if $\beta = 0.012$ and $\mathbf{A}$ is a sufficiently tall normalized Gaussian matrix, the conditions of Theorem 3.9 allow taking q as large as 0.8486. With $\beta = 0.012$, $q = 0.8486$, we obtain a convergence rate of $C_q = C_{0.8486} \geq 0.0287$ (note that this is independent of the size of the system and decreases initial distance to the solution 10 times in 80 iterations).

Finally, in all the theorems, one does not have to compute the optimal step size precisely to get convergence rate of the optimal order. In particular, Remark 3.15 shows that with the step size $\tilde{\alpha} = \xi \alpha_{opt}$ with $\xi \in (0, 2)$, the convergence rate is $(\xi - \xi^2/2)$ times the "optimal"

convergence rate.

3.2 Proposed Method

Here we provide a formal description of our algorithm. Under the same heuristic as in [34, 74], we use the q-quantile of the absolute residual $|\mathbf{A}\mathbf{x} - \mathbf{b}|$ as a threshold to detect and avoid projecting onto rows that are too far from the current iterate (and thus, are likely to be corrupted). Then, an iteration of averaged block Kaczmarz is performed using the qm rows with residual entries less than the computed quantile, using a fixed step size α.

Algorithm 2 Quantile Averaged Block Kaczmarz

procedure QUANTILEABK($\mathbf{A}, \mathbf{b}$, N, q, α, $\mathbf{x}_0$)
  for $k = 1, 2, \ldots, N-1$ **do**
    Compute $Q_q(\mathbf{x}_{k-1}) = q^{\text{th}}$ quantile of $\{|\mathbf{a}_i^T\mathbf{x}_{k-1} - b_i| : i \in [m]\}$
    Set $\tau = \{i \in [m] : |\mathbf{a}_i^T\mathbf{x}_{k-1} - b_i| < Q_q(\mathbf{x}_{k-1})\}$
    Update $\mathbf{x}_k = \mathbf{x}_{k-1} - \frac{\alpha}{|\tau|}\sum_{i\in\tau}(\mathbf{a}_i^T\mathbf{x}_{k-1} - b_i)\mathbf{a}_i$
  end for
  return $\mathbf{x}_N$
end procedure

Note here that we show the algorithm as running for a prespecified number of iterations N, but in practice one may use any desired stopping criterion.

We note that the iterates of both QuantileRK and QuantileABK are at least $O(\beta m)$-times more computationally intensive than the standard RK method. This is because one needs to compute the residual entries of that many rows to obtain a quantile statistic that is able to accurately detect corrupted rows.

We note that the performance of the method depends heavily on the parameters q and α. In Section 3.3, we prove our main convergence result, including a derivation of an optimal value of α and constraints on q to ensure convergence. We follow this with experiments

in Section 3.4 to examine the optimal choice of α in practice, and to show the effects of varying q.

Remark 3.10. Note that in [61], a *weighted* average is taken at each iteration, whilst we take an unweighted average. We reason that in our method, there is no particular reason to weight some rows more heavily than others: whilst one may be inclined to weight rows, say, proportionally to their residual entry, this has the knock-on effect of weighting potentially corrupted rows more heavily. However, we believe that our analysis may be extended to include additional weight parameters.

Remark 3.11. We choose to use a fixed step size at each iteration, but it is possible to extend the method to have varying step size. In particular, the theoretically optimal step size derived in Theorem 3.7 is difficult to estimate *a priori*, and may be substituted with an adaptive step size calculated only with information available at runtime as analyzed in [61]. Note that the QuantileSGD method proposed in [34], like QuantileRK, also utilizes the idea of varying the step size. While QuantileRK uses the quantile of the residual to decide whether to update the current iterate, QuantileSGD always does the weighted update, with the step size determined by the quantile size (and thus changing with the iterations).

3.3 Theoretical Results

3.3.1 Preliminaries

We begin our theory by introducing requisite preliminary results from [74, 34], and we include their proofs for completeness. Firstly, we provide an estimate on the residual quantiles computed at each iteration. Such an estimate is necessary to bound the impact that corrupted rows passing under this threshold can have on convergence. This is ([74], Lemma 1) and is a deterministic version of ([34], Corollary 1).

Lemma 3.12. *Consider applying QuantileABK to the system* $\mathbf{Ax} = \mathbf{b}$. *If* $0 < q < 1 - \beta$, *then the*

quantile computed at the k-th iteration satisfies

$$Q_q(\mathbf{x}_k) \leq \frac{\sigma_{\max}}{\sqrt{m}\sqrt{1-q-\beta}} \|\mathbf{x}_k - \mathbf{x}^\star\|.$$

Proof. We follow [74]. Let $\tau_1, \tau_2 \subset [m]$ denote the sets of indices of uncorrupted and corrupted rows respectively. Note that $|\tau_2| \leq \beta m$. We then have that

$$\begin{aligned}
\sum_{i\in\tau_1} |\langle \mathbf{a}_i, \mathbf{x}_k\rangle - b_i|^2 &= \|\mathbf{A}_{\tau_1}\mathbf{x}_k - \mathbf{b}_{\tau_1}\|^2 \\
&= \|\mathbf{A}_{\tau_1}\mathbf{x}_k - \mathbf{A}_{\tau_1}\mathbf{x}^\star\|^2 \\
&\leq \|\mathbf{A}_{\tau_1}\|^2 \|\mathbf{x}_k - \mathbf{x}^\star\|^2 \\
&\leq \|\mathbf{A}\|^2 \|\mathbf{x}_k - \mathbf{x}^\star\|^2 \\
&= \sigma_{\max}^2 \|\mathbf{x}_k - \mathbf{x}^\star\|^2.
\end{aligned}$$

Next, note that by the definition of $Q_q(\mathbf{x}_k)$, at least $(1-q)m$ rows have absolute residual entry greater than $Q_q(\mathbf{x}_k)$, and at least $(1-q-\beta)m$ of those are uncorrupted. Therefore,

$$m(1-q-\beta)Q_q(\mathbf{x}_k)^2 \leq \sum_{i\in\tau_1} |\langle \mathbf{a}_i, \mathbf{x}_k\rangle - b_i|^2 \leq \sigma_{\max}^2 \|\mathbf{x}_k - \mathbf{x}^\star\|^2.$$

Rearranging gives the result. □

Next, we provide an estimate on the coherence of any subset of rows of $\mathbf{A}$ of fixed size. This is necessary to control the adversarial case in which corruptions occur on coherent rows. We replicate ([34], Lemma 4), but without randomness assumptions on $\mathbf{A}$.

Lemma 3.13. *Let $\mathbf{A} \in \mathbb{R}^{m\times n}$ and let $\mathbf{x} \in \mathbb{R}^n$. Then for every set of row indices $\tau \subseteq [m]$, we have*

$$\sum_{i\in\tau} |\langle \mathbf{x}, \mathbf{a}_i\rangle| \leq \sigma_{\max}\sqrt{|\tau|}\,\|\mathbf{x}\|.$$

Proof. As in [34], let $\mathbf{s} \in \mathbb{R}^m$ have entries

$$s_i = \begin{cases} \text{sign}(\langle \mathbf{x}, \mathbf{a}_i \rangle), & \text{if } i \in \tau \\ 0, & \text{otherwise,} \end{cases}$$

for $i \in [m]$. Then we have

$$\sum_{i \in \tau} |\langle \mathbf{x}, \mathbf{a}_i \rangle| = \sum_{i=1}^{m} \langle \mathbf{x}, s_i \mathbf{a}_i \rangle = \langle \mathbf{x}, \sum_{i=1}^{m} s_i \mathbf{a}_i \rangle \leq \left\| \sum_{i=1}^{m} s_i \mathbf{a}_i \right\| \|\mathbf{x}\| = \left\| \mathbf{A}^\top \mathbf{s} \right\| \|\mathbf{x}\| \leq \sigma_{\max} \sqrt{|\tau|} \|\mathbf{x}\|,$$

as desired. □

Lastly, we give a bound on the norm of the sum of a block of rows of $\mathbf{A}$. This will be used in conjunction with Lemma 3.12 to bound the disruptive effects of corrupted rows that pass under the quantile threshold. We note that to the best of our knowledge, Lemma 3.14 is a new result.

Lemma 3.14. *Let $\mathbf{A} \in \mathbb{R}^{m \times n}$. Then for every set of row indices $\tau \subseteq [m]$, we have*

$$\left\| \sum_{i \in \tau} \mathbf{a}_i \right\|^2 \leq \sigma_{\max}^2 |\tau|.$$

Proof. We use a similar trick to Lemma 3.13. Namely, let $\mathbf{s} \in \mathbb{R}^m$ have entries

$$s_i = \begin{cases} 1, & \text{if } i \in \tau \\ 0 & \text{otherwise,} \end{cases}$$

for $i \in [m]$. Note that $\|\mathbf{s}\|^2 = |\tau|$. Then we have that

$$\left\| \sum_{i \in \tau} \mathbf{a}_i \right\|^2 = \left\| \mathbf{A}^T \mathbf{s} \right\|^2 \leq \left\| \mathbf{A}^T \right\|^2 \|\mathbf{s}\|^2 = \sigma_{\max}^2 |\tau|,$$

as claimed. □

3.3.2 General Case

Armed with our lemmas from the previous section, we are now ready to prove Theorem 3.7.

Proof of Theorem 3.7. Denote by τ the set of row indices passing the quantile test at iteration $k+1$, i.e.

$$\tau = \{i \in [m] : |b_i - \mathbf{a}_i^\top \mathbf{x}_k| \leq Q_q(\mathbf{x}_k)\}.$$

Denote by $\tau_1, \tau_2 \subset \tau$ the subsets of indices corresponding to uncorrupted and corrupted rows respectively. Note that we have $|\tau| = qm$, $|\tau_1| \geq (q-\beta)m$, $|\tau_2| \leq \beta m$. As is typical in Kaczmarz-esque convergence proofs, we now attempt to bound $\|\mathbf{x}_{k+1} - \mathbf{x}^\star\|^2$ in terms of $\|\mathbf{x}_k - \mathbf{x}^\star\|^2$. We have:

$$\begin{aligned}
\|\mathbf{x}_{k+1} - \mathbf{x}^\star\|^2 &= \left\| \mathbf{x}_k - \frac{\alpha}{|\tau|} \sum_{i \in \tau} (\mathbf{a}_i^T \mathbf{x}_k - b_i)\mathbf{a}_i - \mathbf{x}^\star \right\|^2 \\
&= \left\| \mathbf{x}_k - \frac{\alpha}{|\tau|} \sum_{i \in \tau_1} (\mathbf{a}_i^T \mathbf{x}_k - b_i)\mathbf{a}_i - \frac{\alpha}{|\tau|} \sum_{i \in \tau_2} (\mathbf{a}_i^T \mathbf{x}_k - b_i)\mathbf{a}_i - \mathbf{x}^\star \right\|^2 \\
&= \left\| (\mathbf{x}_k - \mathbf{x}^\star) - \frac{\alpha}{|\tau|} \sum_{i \in \tau_1} \mathbf{a}_i \mathbf{a}_i^T (\mathbf{x}_k - \mathbf{x}^\star) - \frac{\alpha}{|\tau|} \sum_{i \in \tau_2} (\mathbf{a}_i^T \mathbf{x}_k - b_i)\mathbf{a}_i \right\|^2 \\
&= \|X - Y\|^2 = \|X\|^2 - 2\langle X, Y \rangle + \|Y\|^2,
\end{aligned}$$

where $X := \left(\mathbf{I} - \frac{\alpha}{|\tau|} \mathbf{A}_{\tau_1}^T \mathbf{A}_{\tau_1}\right) \mathbf{e}_k$ with $\mathbf{e}_k = \mathbf{x}_k - \mathbf{x}^\star$, and $Y := \frac{\alpha}{|\tau|} \sum_{i \in \tau_2} (\mathbf{a}_i^T \mathbf{x}_k - b_i)\mathbf{a}_i$. We proceed to analyze these three terms individually.

Term 1: For the first uncorrupted term $\|X\|^2$, we pursue an analysis similar to that of [61], and let $\mathbf{W}_{\tau_1} := \mathbf{A}_{\tau_1}^\top \mathbf{A}_{\tau_1}$. We then have:

$$\begin{aligned}
\|X\|^2 = \left\| \left(\mathbf{I} - \frac{\alpha}{|\tau|} \mathbf{W}_{\tau_1}\right) \mathbf{e}_k \right\|^2 &= \mathbf{e}_k^\top \left(\mathbf{I} - \frac{\alpha}{|\tau|} \mathbf{W}_{\tau_1}\right)^2 \mathbf{e}_k \\
&= \mathbf{e}_k^\top \left(\mathbf{I} - 2\frac{\alpha}{|\tau|} \mathbf{W}_{\tau_1} + \frac{\alpha^2}{|\tau|^2} \mathbf{W}_{\tau_1}^2\right) \mathbf{e}_k. \qquad (3.5)
\end{aligned}$$

Then, we estimate under the positive semi-definite (Loewner) ordering:

$$\sigma^2_{q-\beta,\min} \leq \sigma^2_{\min}(\mathbf{A}_{\tau_1}) \leq \mathbf{W}_{\tau_1} \leq \lambda_{\max}(\mathbf{W}_{\tau_1}) = \sigma^2_{\max}(\mathbf{A}_{\tau_1}) \leq \sigma^2_{\max},$$

recalling that $\sigma^2_{\max} := \sigma^2_{\max}(\mathbf{A})$, and the smallest restricted singular values $\sigma^2_{q-\beta,\min}$ is defined as per (3.1). Furthermore, to ensure a decrease in norm at each iteration, we require $2\alpha/|\tau| - \alpha^2\sigma^2_{\max}/|\tau|^2 \geq 0$. This is a less restrictive bound than the later condition on α for convergence in Equation (3.9), so we proceed. From Equation (3.5) we have

$$\begin{aligned}
\|X\|^2 &\leq \mathbf{e}_k^\top \left(\mathbf{I} - 2\frac{\alpha}{|\tau|}\mathbf{W}_{\tau_1} + \frac{\alpha^2}{|\tau|^2}\sigma^2_{\max}\mathbf{W}_{\tau_1}\right)\mathbf{e}_k \\
&\leq \left(1 - \left(\frac{2\alpha}{|\tau|} - \frac{\alpha^2}{|\tau|^2}\sigma^2_{\max}\right)\sigma^2_{\min}(\mathbf{A}_{\tau_1})\right)\|\mathbf{e}_k\|^2 \\
&\leq \left(1 - \left(\frac{2\alpha}{qm} - \frac{\alpha^2}{q^2m^2}\sigma^2_{\max}\right)\sigma^2_{q-\beta,\min}\right)\|\mathbf{e}_k\|^2 .
\end{aligned}$$

Term 2: For the scalar product between uncorrupted and accepted but corrupted parts $\langle X, Y\rangle$, we make use of Lemma 3.13. We have

$$\begin{aligned}
\langle X, Y\rangle &\leq \frac{2\alpha}{|\tau|}\sum_{i\in\tau_2}\left|\left\langle\left(\mathbf{I} - \frac{\alpha}{|\tau|}\mathbf{W}_{\tau_1}\right)\mathbf{e}_k, (\mathbf{a}_i^\top\mathbf{x}_k - b_i)\mathbf{a}_i\right\rangle\right| \\
&\leq \frac{2\alpha Q_q(\mathbf{x}_k)}{|\tau|}\sqrt{|\tau_2|}\sigma_{\max}\left\|\left(\mathbf{I} - \frac{\alpha}{|\tau|}\mathbf{W}_{\tau_1}\right)\mathbf{e}_k\right\| \\
&\leq \frac{2\alpha Q_q(\mathbf{x}_k)\sqrt{|\tau_2|}}{|\tau|}\sigma_{\max}\left(1 - \frac{\alpha}{|\tau|}\sigma^2_{q-\beta,\min}\right)\|\mathbf{e}_k\| \\
&\leq \frac{2\alpha\sqrt{\beta}\sigma^2_{\max}}{qm\sqrt{1-q-\beta}}\left(1 - \frac{\alpha}{qm}\sigma^2_{q-\beta,\min}\right)\|\mathbf{e}_k\|^2 .
\end{aligned}$$

Here we use that $|\tau| = qm$, $|\tau_2| \leq \beta m$, and estimate $Q_q(\mathbf{x}_k)$ using Lemma 3.12.

Term 3: Lastly, we estimate the maximal total impact of the residual constrained cor-

rupted equations, making use of both Lemma 3.12 and Lemma 3.14:

$$\begin{aligned}
\|Y\|^2 &= \left\| \frac{\alpha}{|\tau|} \sum_{i \in \tau_2} (\mathbf{a}_i^\top \mathbf{x}_k - b_i) \mathbf{a}_i \right\|^2 \\
&= \frac{\alpha^2}{|\tau|^2} \left\| \sum_{i \in \tau_2} (\mathbf{a}_i^\top \mathbf{x}_k - b_i) \mathbf{a}_i \right\|^2 \\
&\leq \frac{\alpha^2 Q_q(\mathbf{x}_k)^2}{|\tau|^2} \left\| \sum_{i \in \tau_2} \mathbf{a}_i \right\|^2 \\
&\leq \frac{\alpha^2 Q_q(\mathbf{x}_k)^2}{|\tau|^2} \sigma_{\max}^2 |\tau_2| \\
&\leq \frac{\alpha^2 \beta \sigma_{\max}^4}{q^2 m^2 (1-q-\beta)} \|\mathbf{e}_k\|^2 .
\end{aligned}$$

Bringing all three estimates together yields

$$\begin{aligned}
\|\mathbf{e}_{k+1}\|^2 \leq \Bigg[1 - \left(\frac{2\alpha}{qm} - \frac{\alpha^2}{q^2 m^2} \sigma_{\max}^2 \right) \sigma_{q-\beta,\min}^2 + \frac{2\alpha\sqrt{\beta}\sigma_{\max}^2}{qm\sqrt{1-q-\beta}} \left(1 - \frac{\alpha}{qm} \sigma_{q-\beta,\min}^2 \right) \\
+ \frac{\alpha^2 \beta \sigma_{\max}^4}{q^2 m^2 (1-q-\beta)} \Bigg] \|\mathbf{e}_k\|^2 = \left(1 - c_1 \alpha + c_2 \alpha^2 \right) \|\mathbf{e}_k\|^2 ,
\end{aligned} \tag{3.6}$$

where

$$c_1 = \frac{2\sigma_{q-\beta,\min}^2}{qm} - \frac{2\sqrt{\beta}\sigma_{\max}^2}{qm\sqrt{1-q-\beta}}, \tag{3.7}$$

$$c_2 = \frac{\sigma_{\max}^2 \sigma_{q-\beta,\min}^2}{q^2 m^2} - \frac{2\sqrt{\beta}\sigma_{\max}^2 \sigma_{q-\beta,\min}^2}{q^2 m^2 \sqrt{1-q-\beta}} + \frac{\beta \sigma_{\max}^4}{q^2 m^2 (1-q-\beta)}. \tag{3.8}$$

In order to achieve convergence we must have $c_1 > 0$. This is equivalent to

$$\frac{\sqrt{\beta}}{\sqrt{1-q-\beta}} < \frac{\sigma_{q-\beta,\min}^2}{\sigma_{\max}^2},$$

which is reminiscent of the relative conditions imposed on q and β in Theorem 3.6, though slightly relaxed. With this restriction, we then have convergence for all α such that

$$1 - c_1 \alpha + c_2 \alpha^2 < 1, \tag{3.9}$$

equivalently, $\alpha \in (0, c_1/c_2)$, with an optimal choice of $\alpha := c_1/2c_2$. With this optimal choice, our per-iteration guarantee becomes

$$\|\mathbf{e}_{k+1}\|^2 \leq \left(1 - \frac{c_1^2}{4c_2}\right) \|\mathbf{e}_k\|^2 .$$

Induction then yields the result. □

Remark 3.15 (Non-optimal choices of α). Note that since the convergence rate has quadratic dependence on α (3.6), taking $\alpha = \xi\alpha_{opt}$ with $\xi \in (0, 2)$ results in the convergence rate that is $(\xi - \xi^2/2)$ times the "optimal" convergence rate. This implies certain stability in the choice of α: an approximation within a small constant factor does not change the dependence of the convergence rate on any characteristics of the matrix $\mathbf{A}$. Further, we focus on estimating the optimal step size $\alpha = \alpha_{opt}$.

We proceed now to prove Corollary 3.8, in effect giving a simplification of the convergence rate derived in Theorem 3.7.

Proof of Corollary 3.8. Given that for some $\epsilon \in (0, 1)$,

$$\frac{\sqrt{\beta}}{\sqrt{1-q-\beta}} = \epsilon \frac{\sigma^2_{q-\beta,\min}}{\sigma^2_{\max}},$$

we may simplify the expression for the rate via its components c_1 and c_2 given by (3.7) and (3.8). Specifically, it simplifies to

$$c_1 = \frac{2(1-\epsilon)\sigma^2_{q-\beta,\min}}{qm}$$

and

$$\begin{aligned} c_2 &= \frac{\sigma^2_{\max}\sigma^2_{q-\beta,\min}}{q^2m^2} - \frac{2\sqrt{\beta}\sigma^2_{\max}\sigma^2_{q-\beta,\min}}{q^2m^2\sqrt{1-q-\beta}} + \frac{\beta\sigma^4_{\max}}{q^2m^2(1-q-\beta)} \\ &= \frac{\sigma^2_{\max}\sigma^2_{q-\beta,\min}}{q^2m^2} - \frac{2\epsilon\sigma^4_{q-\beta,\min}}{q^2m^2} + \frac{\epsilon^2\sigma^4_{q-\beta,\min}}{q^2m^2} \\ &= \frac{\sigma^2_{q-\beta,\min}}{q^2m^2}\left(\sigma^2_{\max} - \epsilon(2-\epsilon)\sigma^2_{q-\beta,\min}\right). \end{aligned}$$

We can thus express the theoretical optimal step size as

$$\alpha = \frac{c_1}{2c_2} = \frac{qm(1-\epsilon)}{\sigma^2_{\max} - \epsilon(2-\epsilon)\sigma^2_{q-\beta,\min}},$$

and our guaranteed convergence rate as

$$1 - \frac{c_1^2}{4c_2} = 1 - \frac{(1-\epsilon)^2\sigma^2_{q-\beta,\min}}{\sigma^2_{\max} - \epsilon(2-\epsilon)\sigma^2_{q-\beta,\min}}.$$

□

Remark 3.16. Our convergence rate in the general case is difficult to compare with the rate for QuantileRK given in Theorem 3.6, as both expressions are complex and quite different. However, comparing leading terms one may show that our rate is $\mathcal{O}(\sigma^2_{q-\beta,\min}/\sigma^2_{\max})$, and the rate found in Theorem 3.6 is $\mathcal{O}(\sigma^2_{q-\beta,\min}/m)$, showing our method yields a speedup by a factor of $m/\sigma^2_{\max} \geq 1$ (recall that the normalization of the rows ensures that $m = \|\mathbf{A}\|_F^2 \geq \sigma^2_{\max}$). In the next section, we are able to make this more precise for the particular case that $\mathbf{A}$ satisfies the random matrix heuristic given in Section 3.1.3.4.

3.3.3 Subgaussian Case

In this section we take the point of view of [34], namely that $\mathbf{A}$ belongs to the class of random matrices described by Section 3.1.3.4. Within this setting, we will show that $\sigma^2_{q-\beta,\min}$ and $\sigma^2_{\max}$ are both on the order of m/n, giving rise to Theorem 3.9: i.e., the theoretical convergence rate in this case is not dependent on m or n.

As mentioned in [34] and discussed in greater detail in [74], a standard example of a matrix satisfying Section 3.1.3.4 is one whose rows have been sampled independently from the uniform distribution on the sphere. Alternatively, one may sample rows from the standard multivariate Gaussian distribution, and then normalize. The benefit of introducing this random matrix model is that the spectra of such matrices are well studied. In particular, it allows for a high probability uniform lower bound on the smallest singular values of uniform-sized submatrices of $\mathbf{A}$: we state ([34], Proposition 1) below.

Proposition 3.17 ([34], Proposition 1)**.** *Let $\delta \in (0,1]$ and let $\mathbf{A} \in \mathbb{R}^{m\times n}$ satisfy Section 3.1.3.4 with constants D and K. Then there exist absolute constants $C_1, C_2 > 0$ such that if $\mathbf{A}$ has large enough aspect ratio, namely,*

$$\frac{m}{n} > C_1 \frac{1}{\delta} \log \frac{DK}{\delta},$$

then the following high probability uniform lower bound holds for the smallest singular values of all its row submatrices that have at least δm rows.

$$\mathbb{P}\left(\inf_{\tau \subseteq [m], |\tau| \geq \delta m} \sigma_{\min}(\mathbf{A}_\tau) \geq \frac{\delta^{3/2}}{24D} \sqrt{\frac{m}{n}} \right) \geq 1 - 3\exp(C_2 \delta m).$$

Equipped with this result, taking $\delta = q - \beta$ gives the following bound on our key quantity of interest $\sigma^2_{q-\beta,\min}(\mathbf{A})$.

Corollary 3.18. *Suppose that $\mathbf{A}$ satisfies Assumptions 1 and 2, and let C_1, C_2 be the absolute constants arising from Proposition 3.17 upon taking $\delta = q - \beta$. If*

$$\frac{m}{n} > C_1 \frac{1}{q-\beta} \log \frac{DK}{q-\beta},$$

then with probability at least $1 - 3\exp(-C_2(q-\beta)m)$,

$$\sigma^2_{q-\beta,\min} \geq \frac{(q-\beta)^3}{(24D)^2} \frac{m}{n}.$$

Furthermore, we have the following standard bound on $\sigma^2_{\max}(\mathbf{A})$ (see, e.g. [79], Theorem 4.6.1.):

Theorem 3.19. *Let $\mathbf{A} \in \mathbb{R}^{m\times n}$ be a random matrix satisfying Section 3.1.3.4 with constants K, D. Then*

$$\sigma^2_{\max} \leq (1 + CK^2)\frac{m}{n}$$

with probability at least $1 - 2\exp(-cm)$, for some absolute constants $C, c > 0$.

Equipped with these, we may conclude Theorem 3.9 directly from Corollary 3.8:

Proof of Theorem 3.9. The restriction on β given in Equation (3.3), and the optimal step size α given in Equation (3.4), follow immediately from plugging the estimates on $\sigma^2_{q-\beta,\min}, \sigma^2_{\max}$ (given in Proposition 3.17 and Theorem 3.19 respectively) into the corresponding restriction and optimal step size formulae found in Corollary 3.8.

For the convergence result, we may similarly apply estimates of $\sigma^2_{\max}$ and $\sigma^2_{q-\beta,\min}$ to the convergence guarantee in Corollary 3.8. Using these, with probability at least $1 - 2\exp(-cm) - 3\exp(-C_2(q-\beta)m) \geq 1 - c_3\exp(-c_q m)$ we have that

$$\begin{aligned}
1 - \frac{(1-\epsilon)^2\sigma^2_{q-\beta,\min}}{\sigma^2_{\max} - \epsilon(2-\epsilon)\sigma^2_{q-\beta,\min}} &\leq 1 - \frac{(1-\epsilon)^2\sigma^2_{q-\beta,\min}}{\sigma^2_{\max}} \\
&\leq 1 - \frac{(1-\epsilon)^2\frac{(q-\beta)^3 m}{(24D)^2 n}}{(1+CK^2)\frac{m}{n}} \\
&= 1 - \frac{(1-\epsilon)^2(q-\beta)^3}{(1+CK^2)(24D)^2} \\
&= 1 - C_q.
\end{aligned}$$

Now, c_q and C_q are absolute constants depending only on q, so we have that for any k,

$$\|\mathbf{x}_k - \mathbf{x}^\star\|^2 \leq (1-C_q)^k \|\mathbf{x}_0 - \mathbf{x}^\star\|^2,$$

as claimed.

□

Remark 3.20. Theorem 3.9 shows that in the subgaussian setting, QuantileABK enjoys a speedup by a factor of n over QuantileRK (recall Theorem 3.5). This is, heuristically, due to the fact that averaging the projections onto many rows yields a direction vector that points more directly towards $\mathbf{x}^\star$ than any single projection. Hence, one may take a much larger step size, of order n in the subgaussian setting. The convergence rate then enjoys a corresponding increase of the same order. We note that the optimal step size is likely closely related to the coherence of the matrix, in that larger step sizes may be used for matrices with nearly orthogonal rows (such as those of subgaussian-type). This is because

the coherence, in some sense, determines how much information about the location of $x^\star$ can be obtained from a block of rows. We explore and comment on this phenomenon further in Section 3.4.

3.4 Experimental Results

We divide our experiments into two main sections. We first present results using our method as presented in Algorithm 2, including determining the optimal step size, exploring robustness with respect to the quantile parameter, and comparing performance with QuantileRK. We then perform an analysis of a variation of our method, in which only a subset of rows are taken at each iteration and used for computing the quantile and averaged direction vector. This reduces computational cost (at least when it is not possible to compute the full residual in parallel), but yields potentially slower per-iteration convergence. We explore this trade-off experimentally, and believe that our theoretical results may be extended to this method for sufficiently large sample sizes, but leave such theory to future work. Lastly, we also include a demonstration of how a projective block method may not converge in the sparse corruption setting.

3.4.1 Results without subsampling

We begin with our method as presented in Algorithm 2. We perform experiments on systems lying on two geometrical extremes: "Gaussian" systems, where the entries of each row are sampled i.i.d. $\mathrm{N}(0,1)$ and then each row is normalized; and "coherent" systems, where the entries of each row are sampled i.i.d. $\mathrm{Uniform}(0,1)$ and then each row is normalized. These choices are motivated by the fact that the performance of row projection methods such as ours depends heavily on the geometry of the system, in particular the coherence (that is, the pairwise inner products of rows). Gaussian systems are typically highly incoherent, whilst our coherent construction produces highly coherent systems.

We use $m = 10000$ rows in all experiments. $\mathbf{A}$ is constructed to be either "Gaussian" or

"coherent" as described, and then $\mathbf{x}^\star \in \mathbb{R}^n$ is constructed at random with $\mathrm{N}(0,1)$ entries. We then let $\mathbf{b} = \mathbf{A}\mathbf{x}^\star$. Corruptions are placed uniformly at random and are taken to be of size $\mathrm{Uniform}(-100, 100)$, which is large relative to the magnitude of the entries of $\mathbf{b}$. Other parameters will be specified for each experiment.

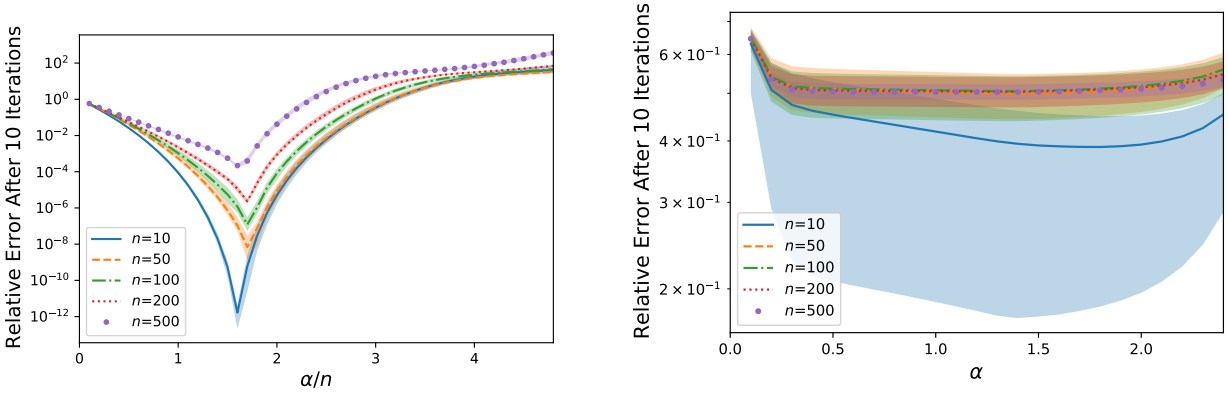
Prior to giving plots showing convergence directly, we first conduct experiments to find the optimal choice of α and q, and then use these optimal choices for convergence plots and comparisons with QuantileRK.

3.4.1.1 Optimal Step Sizes

We begin by determining the optimal choice of step size α for the systems with $\mathbf{A} \in \mathbb{R}^{10000\times n}$, where $n \in \{10, 50, 100, 200, 500\}$. We take $\beta = 0.2$ and $q = 0.7$ and plot the relative error after 10 iterations, $\|\mathbf{x}_k - \mathbf{x}^\star\| \,/\, \|\mathbf{x}_0 - \mathbf{x}^\star\|$, versus α. Since convergence is approximately linear, it suffices to run the method for only a few iterations to determine the optimal parameter. For Gaussian systems the optimal step size appears to scale with the number of columns n, and we present a scaled x-axis to highlight this. The optimal step size is around $1.6n$ to $1.8n$ for each n. For coherent systems, however, the step size does not scale in this fashion, and the optimal step size is approximately 2 for all n. This corresponds to the heuristic that more information about the location of $\mathbf{x}^\star$ is obtained when rows are more incoherent, and thus a larger step size may be taken. Note that a relative error greater than 1 indicates that the method will diverge: for Gaussian systems this happens for step sizes roughly larger than $3n$, and for coherent systems divergence occurs for step sizes roughly larger than 2.5.

3.4.1.2 Optimal choice of q

The relative conditions imposed on q, β in Theorem 3.7 are strict, in the sense that q must in general be much smaller than $1 - \beta$. However, we are able to show in practice that the method is robust even for q very close to $1 - \beta$. This is beneficial as taking q to be larger allows for uncorrupted rows with larger residual entries to be used in the averaged



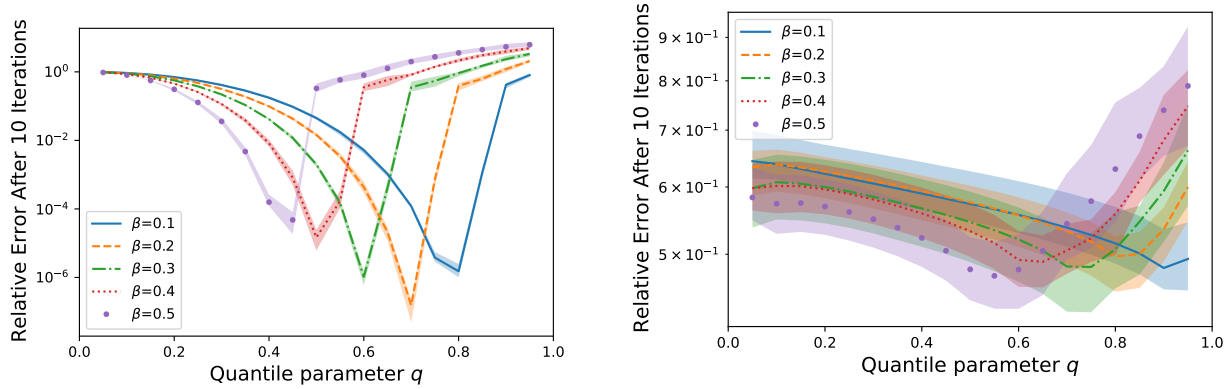
(a) Gaussian systems

(b) Coherent systems

Figure 3.1: Relative error after 10 iterations of QuantileABK applied to $10000 \times n$ normalized Gaussian and coherent systems versus step size, for different numbers of columns n. Note that in (a) the x-axis is scaled by a factor of $1/n$.

projection step, leading to larger movement towards $\mathbf{x}^\star$ and consequently accelerated convergence. In Figure 3.2 we take $\mathbf{A} \in \mathbb{R}^{10000\times 100}$, $\beta \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$, and vary $q \in (0, 1)$. We plot the relative error after 10 iterations of QuantileABK(q) with the optimal step size found experimentally as in the previous subsection.

Our results indicate that in both extremes of system geometry, the method is highly robust to q and q may be taken very close to $1 - \beta$ before convergence begins to slow or fail entirely. In practice, estimating β precisely may be difficult, so one may be more conservative when choosing q.



(a) Gaussian system.

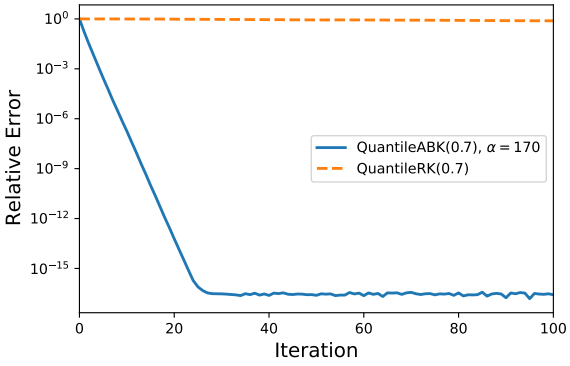
(b) Coherent system.

Figure 3.2: Relative error after 10 iterations of QuantileABK applied to 10000×100 Gaussian and coherent systems versus choice of quantile q, for a range of corruption rates β.

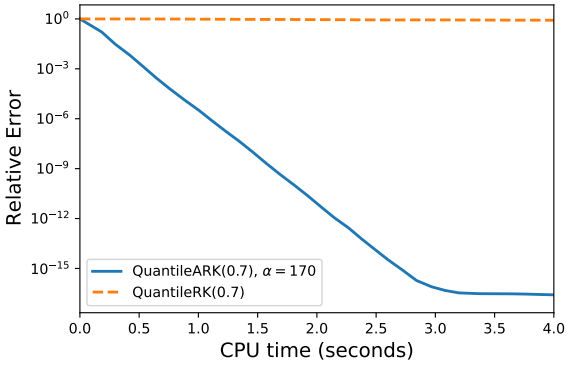
3.4.1.3 Acceleration over QuantileRK

We compare QuantileABK to QuantileRK on 10000×100 systems. We take $\beta = 0.2$, $q = 0.7$ and perform 100 iterations of both methods. In Figure 3.3 we plot the relative error of each method versus iteration, and also versus CPU time. It is clear that QuantileABK outperforms QuantileRK significantly in both the Gaussian and coherent settings, on both a per-iteration and temporal basis. We note that the plateau appearing in the Gaussian plots is due to floating point arithmetic limitations.

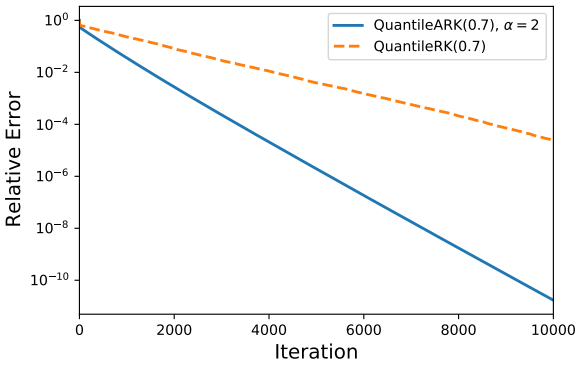
The plots for the coherent system show an initial sharp drop-off in the relative error before a more steady linear convergence. This is a reflection of an initial large movement when $\mathbf{x}_0$ is projected on the first selected hyperplane(s), and then subsequent small movements from further projections as the incident angles between hyperplanes are small. Note that both methods do converge when applied to the consistent system (but slowly, as suggested by Theorem 1.8, since coherency results in small values of $\sigma^2_{q-\beta,\min}$).



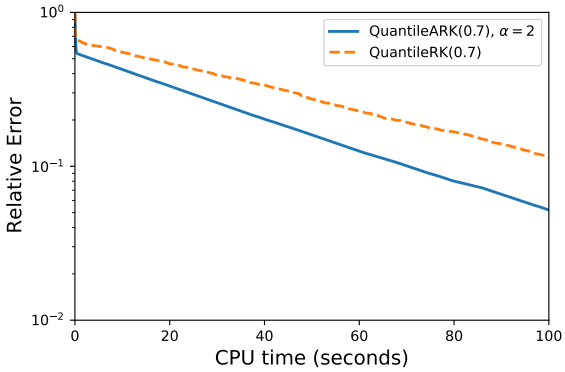
(a) Relative error versus iteration, Gaussian system.



(b) Relative error versus CPU time, Gaussian system.



(c) Relative error versus iteration, coherent system.



(d) Relative error versus CPU time, coherent system.

Figure 3.3: Comparison of QuantileABK to QuantileRK on 10000×100 Gaussian and coherent systems, with quantile $q = 0.7$, corruption rate $\beta = 0.2$.

3.4.2 Results with subsampling

In this section, we perform experiments using a modification of our method, in which at each iteration only a sampled subset of the residual is computed. At each iteration, t rows are sampled uniformly, the subresidual for that block of rows is computed, and its q-quantile is taken. An averaged projection step is then performed using the rows in this block with residual entries below the quantile, with step size α. We call this method SampledQABK and give pseudocode in Algorithm 3. This methodology may be of interest when the full residual cannot be computed in parallel, as in this case subsampling can

substantially reduce the computational cost (accompanied by a trade-off with the per-iteration convergence rate, as we will show). We note that our theoretical results require sampling the full residual, but we believe that this may be relaxed.

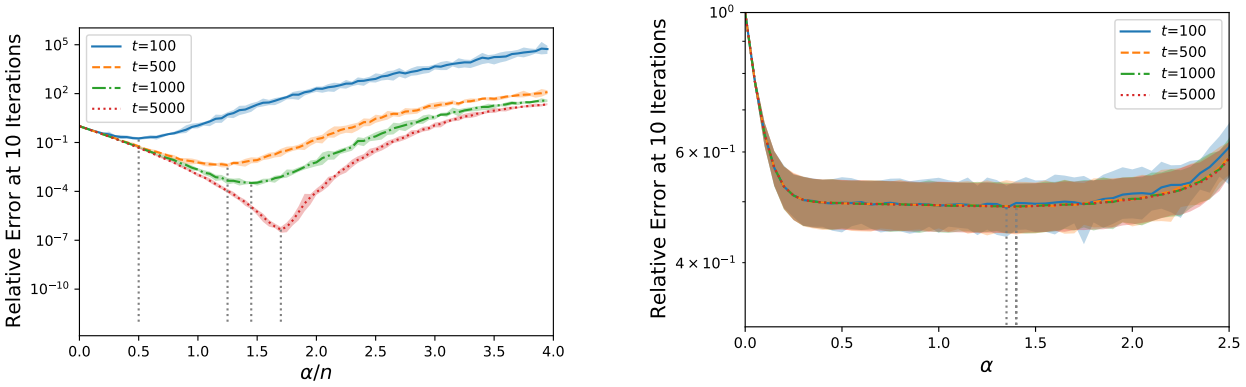
Algorithm 3 Sampled Quantile Averaged Block Kaczmarz

procedure SAMPLEDQABK($\mathbf{A}, \mathbf{b}$, N, q, t, α, $\mathbf{x}_0$)
  for $k = 1, 2, \ldots, N-1$ **do**
    Sample $i_1, \cdots, i_t \in [m]$ uniformly without replacement
    Compute $Q_q(\mathbf{x}_{k-1}) = q^{\text{th}}$ quantile of $\{|\mathbf{a}_i^T \mathbf{x}_{k-1} - b_i| : i \in \{i_1, \cdots, i_t\}\}$
    Set $\tau = \{i \in \{i_1, \cdots, i_t\} : |\mathbf{a}_i^T \mathbf{x}_{k-1} - b_i| < Q_q(\mathbf{x}_{k-1})\}$
    Update $\mathbf{x}_k = \mathbf{x}_{k-1} - \frac{\alpha}{|\tau|} \sum_{i \in \tau} (\mathbf{a}_i^T \mathbf{x}_{k-1} - b_i) \mathbf{a}_i$
  end for
  return $\mathbf{x}_N$
end procedure

Our experimental setup is the same as in the previous section: Gaussian and coherent systems are constructed in the same manner, corruptions are taken $\text{Uniform}(-100, 100)$ and placed uniformly at random, and other parameters will be specified for each experiment.

3.4.2.1 Optimal Step Sizes

We again begin by finding the optimal step size experimentally. We fix 10000×100 Gaussian and coherent systems and take $\beta = 0.2$, $q = 0.7$. We then take sample sizes $t \in \{100, 500, 1000, 5000\}$ and run SampledQABK for 10 iterations for a range of step sizes α, and present our results in Figure 3.4. As mentioned previously, the method converges linearly and so it is sufficient to run it for only a few iterations for comparison purposes. We note that the method was unstable or did not converge for $t < 100$, which is a consequence of the method being unable to accurately distinguish corrupted and uncorrupted rows given such a small sample.



(a) Gaussian system.

(b) Coherent system.

Figure 3.4: Relative error after 10 iterations of SampledQABK versus step size α on 10000×100 Gaussian and coherent systems, for a range of sample sizes t, quantile parameter $q = 0.7$, corruption rate $\beta = 0.2$. Note the scaling of the x-axis in (a).

We observe that for the Gaussian system, the optimal step size is again on the order of n, and that the method becomes more sensitive to the choice of step size as the sample size t increases (that is, the 'valleys' at the optima become sharper). This may be explained by the heuristic that the amount of information that may be obtained from a block of rows (and in turn, the step size that may be taken) is limited by the rank of the matrix, which in the Gaussian case is n almost surely. Hence, significant increases in sample size do not yield corresponding increases in the optimal step size.

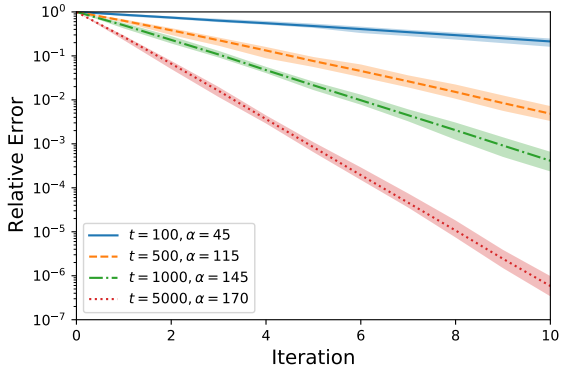
For the coherent system, however, we see that the behavior is almost exactly the same across sample sizes. Again, the optimal choice of α is roughly constant, and there is little further information to leverage from taking larger sample sizes.

3.4.2.2 Effect of Sample Size on Convergence

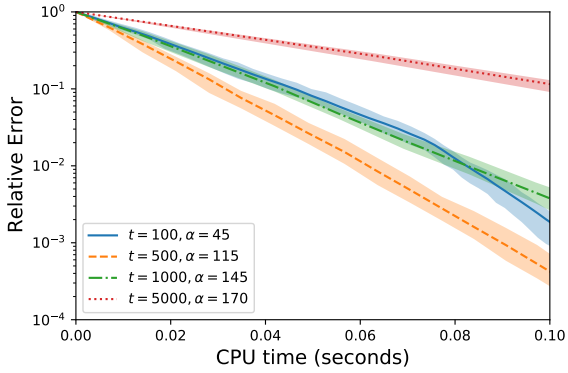
We proceed to compare the convergence of SampledQABK with a variety of sample sizes. In Figure 3.5 we include plots of the relative error versus iteration and versus CPU time, for both Gaussian and coherent systems. We compare sample sizes $t \in \{100, 500, 1000, 5000\}$, and use the optimal step sizes found in the previous section (for simplicity, we take $\alpha = 1.4$ for all sample sizes for the coherent system). We present our results in Figure 3.5.

For the Gaussian system, we see that taking a larger sample size greatly improves the per-iteration convergence. However, when computation time is taken into account, there is a clear trade-off: for example, $t = 500$ converges much faster in terms of CPU time than $t = 5000$.

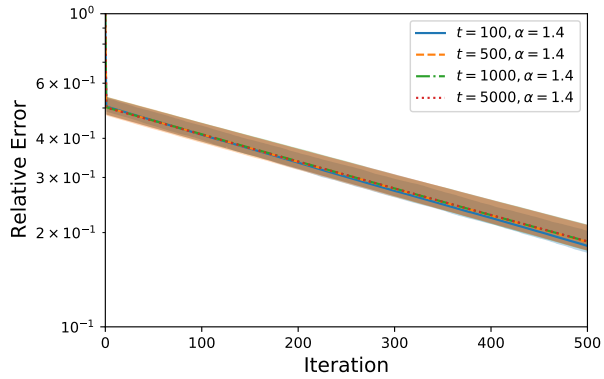
For the coherent system, we see that on a per-iteration basis there is essentially no difference between different sample sizes. The first iteration provides an initial jump, and then convergence proceeds much more slowly than the Gaussian case. Indeed, there is no trade-off between per-iteration convergence and computational cost in this case: subsampling greatly improves convergence over CPU time. We see that taking $t = 100$ is significantly faster than any other sample size. In this case, t should be taken as small as possible while still achieving convergence, and as noted previously we observed that convergence fails for $t < 100$.



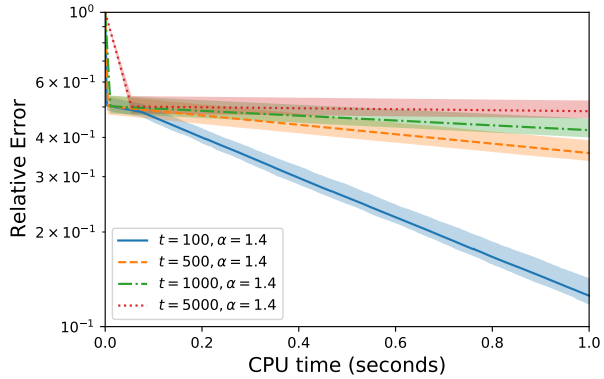
(a) Relative error versus iteration, Gaussian system.



(b) Relative error versus CPU time, Gaussian system.



(c) Relative error versus iteration, coherent system.



(d) Relative error versus CPU time, coherent system.

Figure 3.5: Convergence of SampledQABK for a range of sample sizes t.

3.4.3 Projective vs Averaged

In this section, we give an experiment to support earlier discussion in Section 3.1.3.2 regarding averaged versus projective block variants. We argued that in projective block methods, where iterates are projected onto the intersection of the hyperplanes corresponding to an entire block of rows, the presence of even a single corrupted row in each block can prevent convergence. To illustrate this we first give a natural quantile-based block Kaczmarz variant, QuantilePBK, in Algorithm 4. Similar to QuantileABK, at each iteration, a quantile of the residual is taken, and then the previous iterate is projected onto the intersection of the hyperplanes of every row with residual entry beneath the quantile.

Algorithm 4 Quantile Projective Block Kaczmarz

procedure QUANTILEPBK($\mathbf{A}, \mathbf{b}, N, q, \mathbf{x}_0$)
  for $k = 1, 2, \ldots, N-1$ **do**
    Compute $Q_q(\mathbf{x}_{k-1}) = \mathrm{q}^{\text{th}}$ quantile of $\{|\mathbf{a}_i^T \mathbf{x}_{k-1} - b_i| : i \in [m]\}$
    Set $\tau = \{i \in [m] : |\mathbf{a}_i^T \mathbf{x}_{k-1} - b_i| < Q_q(\mathbf{x}_{k-1})\}$
    Update $\mathbf{x}_k = \mathbf{x}_{k-1} + \mathbf{A}_\tau^\dagger(\mathbf{b}_\tau - \mathbf{A}_\tau \mathbf{x}_{k-1})$
  end for
  return $\mathbf{x}_N$
end procedure

We construct an example to demonstrate how QuantilePBK may fail as follows. We construct a matrix $\mathbf{A} \in \mathbb{R}^{1250 \times 100}$, where 1000 rows are sampled by taking i.i.d. $N(0,1)$ entries and then normalizing, and where 250 rows are identical copies of one further Gaussian row. A solution vector $\mathbf{x}^\star$ is then constructed with i.i.d $N(0,1)$ entries. Then, the 250 identical rows have their entries in b corrupted in order to all equal 500, given 250 identical full rows in the system. Denoting one such row by $\mathbf{a}^\top \mathbf{x} = 500$, the initial iterate $\mathbf{x}_0$ is then taken to be $\mathbf{x}_0 = (500 - \mathbf{a}^\top \mathbf{1})\mathbf{a}^\top$, i.e., the projection of the vector of all ones $\mathbf{1}$ onto the hyperplane $\{\mathbf{a}^\top \mathbf{x} = 500\}$. Under these choices, the iterates will always lie in this (corrupted) hyperplane, hence these rows will always have residual entry zero. This ensures they always pass the quantile test, and ensures that QuantilePBK cannot converge. We note that even taking projections on smaller sub-blocks of the accepted index set τ won't improve robustness, as long as each block contains one of the corrupted rows. So, in the worst case, even 250 blocks of 5 equations each can be such that the iterates never leave the corrupted hyperplane.

We perform QuantilePBK and QuantileABK on this system with $q = 0.7$, initial iterate as described above, and for QuantileABK a step size of $\alpha = 10$. In Figure 3.6, we plot the relative error versus iteration, and indeed observe as expected that QuantilePBK fails to converge, whilst QuantileABK continues to enjoy linear convergence even in this adversarial setting.

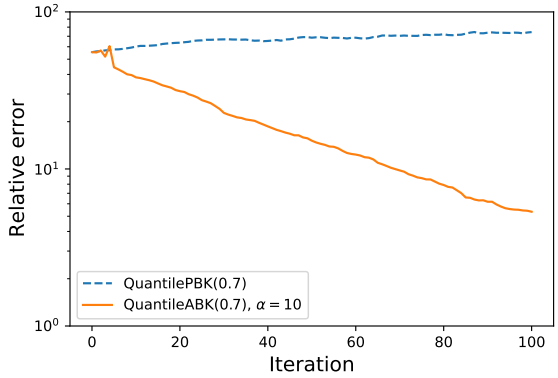


Figure 3.6: Relative Error versus iteration, QuantilePBK versus QuantileABK.

3.5 Conclusion

In this work, we propose a novel method, QuantileABK, for solving large-scale systems of equations that suffer from arbitrarily large, but sparse, corruptions in the measurement vector. This sparse corruption model arises in a wide range of applications, including sensor networks, computerized tomography, and many problems in distributed computing, and finding methods that are able to detect and avoid these corruptions has been a popular recent problem.

Our method combines an averaged blocking technique, that has experienced recent popularity in related literature, with the use of a quantile of the residual at each iteration. This provides a large acceleration over the preeminent existing method for this setting, QuantileRK, by leveraging far more information from the computed residual at each iteration.

We prove that our method enjoys linear convergence under certain conditions on the quantile parameter q, and the fraction of corruption rates β, for all matrices such that the uniform smallest singular value over all row-submatrices with at least $(q - \beta)m$ rows is positive, i.e. $\sigma^2_{q-\beta,\min} > 0$. Notably, our results place no restriction on the size or (potentially adversarial) placement of corruptions. We show theoretically and experimentally that our method converges faster than QuantileRK. In particular by specializing to the case of a

matrix of subgaussian-type, we are able to quantify this speed-up more precisely, and show that our method converges faster than QuantileRK by a factor of n (the number of columns of the system). Whilst this speed up is per-iteration, both methods require computing the full residual at each iteration, so the per-iteration computational cost is of the same order.

Experimentally, we show that our method significantly outperforms QuantileRK, by iteration and by CPU time. We provide experiments on both geometric extremes (that is, matrices with nearly parallel rows and matrices with nearly orthogonal rows), and demonstrate the scaling behavior of the optimal step size in these cases, as well as direct performance comparisons to QuantileRK, in which the increase in convergence rate is clear. We also introduce a variant of our method that uses only a subsample of rows at each iteration, and provide step size and convergence results for a range of sample sizes.

As future work, we propose that there is still further information to be gained from the residual. In particular, we believe that historical residual information may be used to estimate the likelihood of a row being corrupted. That is, if one row's residual entries are continually greater than the quantile threshold, then that row is more likely to be corrupted than others. This could potentially then be used to reduce the number of corrupted rows that are deemed acceptable for projection at each iteration.

CHAPTER 4

Randomized Block Gossip Algorithms for Average Consensus

This chapter is a version of [33] and is joint work with Chen Yap and Jamie Haddock. Chen Yap contributed the codebase, and Chen Yap and I contributed the experiments. Jamie Haddock and I contributed the theory and wrote the paper together. We discuss the average consensus problem, a well-studied model problem in distributed computing, and offer a new perspective: we show that certain popular protocols for the problem, namely gossip protocols, are equivalent to applying a block variant of the Kaczmarz method to a certain system of linear equations. This equivalence brings forth a new theoretical foundation for the convergence of gossip protocols, enabling practitioners to make better-informed choices about their particular implementation. We also generalize much of the theory behind block Kaczmarz methods and broaden the class of systems of linear equations for which convergence guarantees for such methods exist.

4.1 Introduction

Consider a network in which every node has an value that is initially unknown to the other nodes, and the goal is for all nodes to learn the average of these values. This problem is a classical and fundamental problem in distributed computing and multi-agent systems known as *average consensus* [44]; it has additional real-world applications in clock synchronization [24], PageRank, localization without GPS [85], opinion formation, distributed data fusion in sensor networks [82], blockchain technology, and load balancing [12]. See Figure 4.1 for a visualization of an average consensus problem.

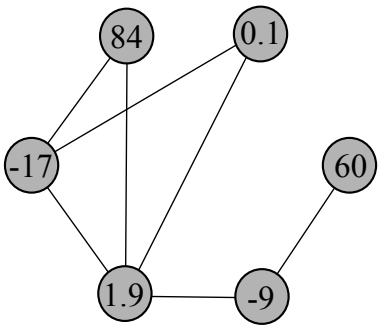


Figure 4.1: Average consensus problem with initial (unknown) values listed. The consensus value for this problem is $\bar{c} = 20$.

First approaches for this problem may be to allow all nodes to pass their initial unknown value to a single *hub* node, by which we mean a node that is edge connected to each other node, which would then perform the averaging and pass this value back to the others, or for every node to share its stored knowledge of all other nodes' inital unknown values with its neighbours until all nodes have learned all stored values (a process known as *flooding* [83]). However, such methods are problematic. The first requires communication that may be infeasible as it may not respect the topology of the underlying network; in particular, there may be no hub node (see Figure 4.1). The second may require many instances of communication between nodes and thus struggle to scale to modern large-scale networks. An attractive class of methods that go some way towards rectifying these issues are *gossip* protocols, where at each time-step some subset of nodes are 'activated' to share their stored information with each other across network edges (note that the hub and flooding methods may still be considered special cases of this, but these protocols typically involve activating *proper* node subsets to avoid the aforementioned large-scale communication issues). As is common in large-scale settings, such protocols are *randomized* in the sense that activated nodes are often chosen at random, with distribution depending on the particular gossip protocol. Over many iterations, the values held at each node converge to the average over the whole network (with mild assumptions including connectivity of the underlying network) and the problem is solved in a distributed manner. Such methods have been a topic of popular study this side of the millenium, with the seminal 2006 paper of Boyd et al. [4] sparking a flurry of research on the topic [1, 15, 14, 36, 52, 62, 69, 87].

In [55], Loizou and Richtárik united the study of randomized gossip protocols with randomized numerical linear algebra. They showed that a wide class of gossip protocols can be interpreted, under certain assumptions, as randomized iterative linear system solvers applied to a linear system derived from the network at hand. This remarkable connection prompted a variety of convergence results, new accelerated and weighted gossip variants, and dual edge-based gossip protocols, opening up the possibility for further links to be developed. In this paper, we hone in on the class of *block* gossip algorithms discussed in their work: we show that several popular gossip protocols are members of this class, give strong convergence results for said methods, and generalize the theory on both the linear algebra and gossip protocol sides of the problem.

4.1.1 Contributions

Our main contributions in this work are threefold:

- We generalize previous results on randomized block iterative methods for solving linear systems (see [61, 64, 66]) to include the case where the system is less than full rank (which is vital for the average consensus case, as we will see), to include a wider class of block sampling protocols, and to sharpen the resulting convergence rate guarantee.

- We derive new convergence rates for popular gossip protocols such as *path* [7], *clique* [53], and *edge-independent set* [4] gossiping, by showing they can be interpreted as block gossip methods. We furthermore generalize to include the case where multiple node sets can be activated simultaneously, and analyze the dependence of a protocol's performance on the spanning trees of the activated subgraphs.

- We give new analyses of gossip for *inconsistent* consensus models. We link edge communication errors in a network to inconsistent, or noisy, linear systems, and connect the performance of iterative solvers on said systems to the performance of gossip protocols on said networks.

Furthermore, we provide a wide range of experiments to demonstrate the comparative performance of the discussed algorithms, both previously existing and new, on a variety of network structures.

We remark that we do not go into detail regarding the specific communication protocols that networks may have. For example, certain methods (e.g., path gossiping [7]) require multiple instances of information sharing at each iteration, which may not be implementable in all networks, and others (e.g., edge independent set gossiping [4]) require communication across each edge only once at each time step. We point the reader to [4, 68] for more details on network communication protocols and their use in gossip algorithms.

4.1.2 Organization

The rest of the paper is organized as follows. In Section 4.1.3, we introduce notation that will be used throughout the work. In Section 4.1.4, we describe in detail the methods considered in this paper and present our main theoretical results. In Section 4.1.5, we provide detailed background on the average consensus problem, and recent work on block iterative methods for solving linear systems. In Section 4.2, we prove our generalization of the block iterative method theory, and go into detail on the connection between this and block gossip algorithms for average consensus. In Section 4.3, we connect particularly popular block gossip algorithms with our theory and produce explicit convergence rates for them. In Section 4.4, we explore the case of average consensus models with edge miscommunication and provide convergence results for the gossip method on this type of faulty model. Lastly, in Section 4.5, we provide numerous experiments and compare our theoretical results with the empirical behavior of the considered gossip methods.

4.1.3 Notation and Definitions

Throughout, we let $[m]$ denote the set of integers from one to m; $[m] := \{1, 2, \cdots, m\}$. We additionally let $\mathbf{0}$ denote the vector of all zeros and let $\mathbf{1}$ denote the vector of all ones (the dimensions of these vectors will be given or obvious from context).

In what follows, we let $\mathbf{A}_\tau$ denote the subset of matrix $\mathbf{A}$ with rows indexed by the set τ. We let $\lambda_{\min}(\mathbf{A}), \lambda_{\min+}(\mathbf{A})$, and $\lambda_{\max}(\mathbf{A})$ denote the minimum, minimum non-zero, and maximum eigenvalues of the matrix $\mathbf{A}$, respectively. Additionally, we denote by $\mathbf{A}^\dagger$ the Moore-Penrose pseudoinverse of the matrix $\mathbf{A}$.

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ denote an undirected network where $\mathcal{V}$ denotes the n nodes of the network and $\mathcal{E}$ denotes the m edges. Let $\mathbf{Q} \in \mathbb{R}^{m \times n}$ denote the incidence matrix of the network. Each row of $\mathbf{Q}$ corresponds to an edge of the network and each column corresponds to a node. If row l of $\mathbf{Q}$, denoted $\mathbf{q}_l^T$, corresponds to the edge $e_{ij} \in \mathcal{E}$ connecting node i and j, then all but the ith and jth entries of $\mathbf{q}_l$ are zero with these entries containing a one and negative one (the order of the positive and negative entries does not matter). Note that $\mathbf{Q}^\top \mathbf{Q} = \mathbf{L} \in \mathbb{R}^{n \times n}$ is the *Laplacian* matrix of the graph $\mathcal{G}$. The second smallest eigenvalue of $\mathbf{L}$ is called the *algebraic connectivity* of the graph $\mathcal{G}$; we denote this value $\lambda_2(\mathcal{G}) := \lambda_2(\mathbf{L})$. Finally, we note that if a graph $\mathcal{G}$ is connected, then the algebraic connectivity $\lambda_2(\mathcal{G})$ is positive, and thus $\lambda_2(\mathcal{G}) = \lambda_{\min+}(\mathbf{L})$.

We recall the notion of several special subgraph structures. We remind the reader of the definition of an *independent edge set*, that is a subset of edges of the graph in which no two edges are incident to the same node. Additionally, we remind the reader that a *clique subgraph*, or *complete subgraph*, is a subset of edges of the graph which together form a subgraph in which every pair of nodes is connected by an edge; that is, the edge-induced subgraph is a clique. Finally, we remind the reader of the definition of a *path subgraph*, a subset of edges which together form a path graph; that is, the edge-induced subgraph is a path.

We additionally recall the notion of a *row paving* of a matrix $\mathbf{A}$, which controls the conditioning of a set of submatrices that partition the matrix $\mathbf{A}$. We include here the definition provided by Needell and Tropp in [64], and note that earlier work on the construction of pavings for block projection methods is due to Popa [71]. In the wider operator theory context, pavings have long been a topic of interest; see [80] for a review.

Definition. A (d, α, β) *row paving* of a matrix $\mathbf{A}$ is a partition $T = \{\tau_1, \tau_2, \cdots, \tau_d\}$ of the row

indices that satisfies

$$\alpha \leq \lambda_{\min}(\mathbf{A}_\tau \mathbf{A}_\tau^\top) \text{ and } \lambda_{\max}(\mathbf{A}_\tau \mathbf{A}_\tau^\top) \leq \beta \text{ for each } \tau \in T.$$

We will additionally be interested in subsets of the row indices that do not necessarily partition the rows. We define a *row covering* to be the natural relaxation of a row paving in which the requirement that the subsets partition the row indices be relaxed and where the parameter α provides a lower-bound for the minimum *non-zero* eigenvalue rather than the minimum eigenvalue. Each of these are important generalizations for our application of interest, gossip algorithms for average consensus.

Definition. A (d, α, β, r, R) *row covering* of a matrix $\mathbf{A}$ is a collection of subsets $T = \{\tau_1, \tau_2, \cdots, \tau_d\}$ of the row indices, $\tau_i \subset [m]$ for all $i = 1, \cdots, d$, that covers the row indices, for each $i \in [m]$ we have $i \in \tau_l$ for some $l = 1, \cdots, d$, and that satisfies

$$\alpha \leq \lambda_{\min +}(\mathbf{A}_\tau \mathbf{A}_\tau^\top) \text{ and } \lambda_{\max}(\mathbf{A}_\tau \mathbf{A}_\tau^\top) \leq \beta \text{ for each } \tau \in T,$$

where r and R are the minimum and maximum, respectively, number of blocks in which a single row appears, i.e., $r = \min_{i \in [m]} |\{\tau_l \in T : i \in \tau_l\}|$ and $R = \max_{i \in [m]} |\{\tau_l \in T : i \in \tau_l\}|$.

First, note that the minimum and maximum number of repeated occurrences of a row in the blocks, r and R, satisfy

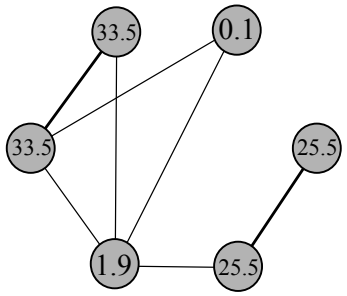
$$1 \leq r \leq R \leq d.$$

Further, note that a (d, α, β) row paving is by definition a $(d, \alpha, \beta, 1, 1)$ row covering. Next, we remind the readers that *every* sufficiently tall row-normalized matrix $\mathbf{A}$ admits a row paving with $\alpha > 0$ and $\beta < 2$ [64]; the authors illustrate that additional mild assumptions on the matrix allows one to produce such a row paving via random partitioning. Additionally, we note that if the matrix in question is an incidence matrix $\mathbf{Q}$, where rows correspond to edges in a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, then the row covering $T = \{\tau_1, \cdots, \tau_d\}$ corresponds to a collection of sets of edges such that every edge appears in at least one set; that is $\tau_i \subset \mathcal{E}$ for

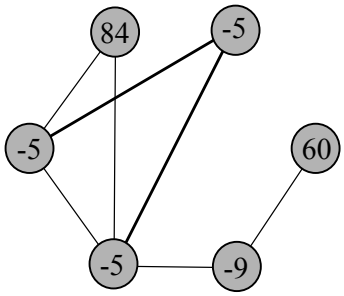
all $i \in [d]$ and for all $e \in \mathcal{E}$, we have $e \in \tau_l$ for some $l \in [d]$.

4.1.4 Main Results

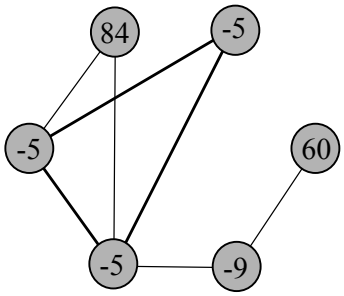
The average consensus problem is defined over an undirected network $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. We let $\mathbf{c} = (c^{(1)}, c^{(2)}, ..., c^{(n)})^T$ denote the vector of inital unknown values (i.e., $c^{(i)}$ is the initial unknown value of the ith node) initally held by the nodes of the network. The average consensus problem is then to ensure, after some communication protocol is applied, that each node stores the averaged value $\bar{c} := \text{mean}(\mathbf{c})$; that is the final vector of updated node values is $\mathbf{c}^* = \bar{c}\mathbf{1}$ where $\mathbf{1} \in \mathbb{R}^n$ is the vector of all ones. One can see that $\mathbf{c}^*$ is the orthogonal projection of $\mathbf{c}$ onto the kernel of $\mathbf{Q}$, $\{\mathbf{x} \mid \mathbf{Q}\mathbf{x} = \mathbf{0}\}$. Here $\mathbf{0} \in \mathbb{R}^{|\mathcal{E}|}$ denotes the all zeros vector. The average consensus problem may be formulated in this way using either the incidence matrix or the Laplacian matrix, $L = D - A$ where D is the diagonal matrix of node degrees and A is the adjacency matrix, or more generally as an *average consensus system* as defined in [56]. In this work, we will focus upon the *block gossip* methods for this problem: at each iteration, a group (block) of edges is activated, and the nodes belonging to each connected component of the subgraph induced by this block average their values. Pseudocode for this method is provided in Algorithm 5. See Figure 4.2 for a visualization of one step of Algorithm 5 on the average consensus problem of Figure 4.1 with independent edge set, path, and clique block sampling.



(a) One iteration of block gossip with a block corresponding to an independent edge set of size 2.



(b) One iteration of block gossip with a block corresponding to a path of length 2.



(c) One iteration of block gossip with a block corresponding to a clique of size 3.

Figure 4.2: The average consensus problem of Figure 4.1 after an iteration of the block gossip method (Algorithm 5) with various types of block structures. The edges defining the sampled block are represented by bold lines.

Algorithm 5 Block Gossip Method

procedure BG($\mathcal{G}, \mathbf{c}_0 = \mathbf{c}, T = \{\tau_1, \cdots, \tau_d\}$)
  $k = 0$
  repeat
    $k \leftarrow k + 1$
    Choose edge subset τ uniformly at random from elements of T.
    Form $\mathcal{G}_\tau = (\mathcal{V}_\tau, \mathcal{E}_\tau)$, the edge-induced subgraph of $\mathcal{G}$ defined by edges in τ.
    Nodes outside of $\mathcal{V}_\tau$ do not update values, $(\mathbf{c}_k)_{\mathcal{V}_\tau^C} \leftarrow (\mathbf{c}_{k-1})_{\mathcal{V}_\tau^C}$.
    for all connected components $\mathcal{G}' = (\mathcal{V}', \mathcal{E}')$ of $\mathcal{G}_\tau$ **do**
      Nodes in $\mathcal{V}'$ solve average consensus on $\mathcal{G}'$, $(\mathbf{c}_k)_{\mathcal{V}'} \leftarrow \left[\frac{1}{|\mathcal{V}'|}\sum_{i\in\mathcal{V}'}(\mathbf{c}_k)_i\right]\mathbf{1}_{|\mathcal{V}'|}$.
    end for
  until stopping criterion reached
  return $\mathbf{c}_k$
end procedure

We next present our main result which illustrates that the block gossip method converges at least linearly in expectation to consensus if the underlying graph is connected. Moreover, we specialize this result to three special cases: the case when the blocks of edges sampled

in each iteration are independent edge sets, when they are clique subgraphs, and when they are path subgraphs, and give refined convergence rates for these cases.

Corollary 4.1. *Suppose graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is connected, $\mathbf{Q} \in \mathbb{R}^{|\mathcal{E}|\times|\mathcal{V}|}$ is the incidence matrix for $\mathcal{G}$, and $T = \{\tau_1, \cdots, \tau_d\}$ is a (d, α, β, r, R) row covering for $\mathbf{Q}$ with $M = \max_{i\in[d]} |\tau_i|$. Then the block gossip method with blocks determined by T converges at least linearly in expectation with the guarantee*

$$\mathbb{E}\left\|\mathbf{c}_k - \mathbf{c}^*\right\|^2 \leq \left(1 - \frac{r\lambda_2(\mathcal{G})}{\beta d}\right)^k \left\|\mathbf{c} - \mathbf{c}^*\right\|^2,$$

where $\lambda_2(\mathcal{G})$ is the algebraic connectivity of graph $\mathcal{G}$.

(1) *If T consists of independent edge sets, the rate constant can be bounded by*

$$\left(1 - \frac{r\lambda_2(\mathcal{G})}{2d}\right).$$

(2) *If T consists of path subgraphs, the rate constant can be bounded by*

$$\left(1 - \frac{r\lambda_2(\mathcal{G})}{(2 - 2\cos\frac{M\pi}{M+1})d}\right) \leq \left(1 - \frac{r\lambda_2(\mathcal{G})}{4d}\right).$$

(3) *If T consists of clique subgraphs, the rate constant can be bounded by*

$$\left(1 - \frac{r\lambda_2(\mathcal{G})}{(2 - 2\cos\frac{M\pi}{M+1})d}\right) \leq \left(1 - \frac{r\lambda_2(\mathcal{G})}{4d}\right).$$

(4) *If T consists of arbitrary connected subgraphs, the rate constant can be bounded by*

$$\left(1 - \frac{r\lambda_2(\mathcal{G})}{Md}\right).$$

Remark 4.2. In the gossip algorithm literature, convergence results are often presented in terms of the ϵ-averaging time of the algorithm, defined in [4] as

$$T_{\text{ave}}(\epsilon) := \sup_{\mathbf{c}} \inf_{k=0,1,2,\cdots} \left\{\mathbb{P}\left[\frac{\|\mathbf{c}_k - \mathbf{c}^*\|}{\|\mathbf{c}\|} \geq \epsilon\right] \leq \epsilon\right\}.$$

Our work goes beyond this by providing explicit convergence rates, but a bound on $T_{\text{ave}}(\epsilon)$ can be easily derived: by noting that $\|\mathbf{c} - \mathbf{c}^*\| \leq \|\mathbf{c}\|$ and applying Markov's inequality, one may apply Corollary 4.1 to obtain

$$T_{\text{ave}}(\epsilon) \leq \frac{3 \log \epsilon}{\log\left(1 - \frac{r\lambda_2(\mathcal{G})}{\beta d}\right)}.$$

However, it is challenging to directly compare this result to others in the literature as these are often only presented as asymptotic bounds without specific dependence on block parameters like r, d, and β. On the other hand, the presence of r, d, and β make clear the dependence of this bound on the block selection strategy.

Corollary 4.1 follows from a new bound on the convergence rate of the *block Kaczmarz method* on a potentially rank-deficient least-squares problem. The block Kaczmarz method samples blocks of rows of the matrix $\mathbf{A}$ in each iteration and performs an update which projects the previous iterate onto the solution space of the subset of sampled equations; note that the standard single-row Kaczmarz updates are a special case of block Kaczmarz with block size one. The details of this method are provided in Algorithm 6. The block gossip method with blocks T produces the same iterates as the block Kaczmarz method performed with $\mathbf{A} = \mathbf{Q}, \mathbf{b} = \mathbf{0}$, and $\mathbf{x}_0 = \mathbf{c}$ with blocks T.

Algorithm 6 Block Kaczmarz Method

procedure BK($\mathbf{A}, \mathbf{b}, \mathbf{x}_0, T = \{\tau_1, \cdots, \tau_d\}$)
    $k = 0$
    repeat
        $k \leftarrow k + 1$
        Choose row block τ uniformly at random from T.
        $\mathbf{x}_k \leftarrow \mathbf{x}_{k-1} + \mathbf{A}_\tau^\dagger(\mathbf{b}_\tau - \mathbf{A}_\tau \mathbf{x}_{k-1})$
    until stopping criterion reached
    return $\mathbf{x}_k$
end procedure

Our main result regarding the block Kaczmarz method generalizes the main result of [64] in several ways:

- Generalizes to the case when the least-squares problem is rank-deficient.

- Relaxes the requirement that the row blocks be sampled from a matrix paving.
- Demonstrates that the convergence horizon depends upon the minimum *nonzero* singular value of the blocks $\mathbf{A}_\tau$ rather the absolute minimum singular value (often 0).

These generalizations are important in our main application to the average consensus problem and block gossip methods, but are likely of independent interest in other applications.

Theorem 4.3. *Consider the least-squares problem*

$$\min \|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2$$

where $\mathbf{A} \in \mathbb{R}^{m\times n}$ is not necessarily full-rank and $\mathbf{b} \in \mathbb{R}^m$. Let $\mathbf{e} = A\mathbf{x}_\mathbf{e} - \mathbf{b}$ for some $\mathbf{x}_\mathbf{e}$ and let $T = \{\tau_1, \cdots, \tau_d\}$ be a (d, α, β, r, R) covering (not necessarily a paving) of the rows of $\mathbf{A}$. Let $\mathbf{x}_j$ denote the jth iterate produced by Block RK on the system defined by $\mathbf{A}$ and $\mathbf{b}$ with initial iterate $\mathbf{x}_0$, and let $\mathbf{x}^* := \left(\mathbf{I} - \mathbf{A}^\dagger\mathbf{A}\right)\mathbf{x}_0 + \mathbf{A}^\dagger(\mathbf{b}+\mathbf{e}) = \left(\mathbf{I} - \mathbf{A}^\dagger\mathbf{A}\right)\mathbf{x}_0 + \mathbf{A}^\dagger\mathbf{A}\mathbf{x}_\mathbf{e}$. Then we have

$$\mathbb{E}\left(\|\mathbf{x}_j - \mathbf{x}^*\|^2\right) \leq \rho(\mathbf{A}, T)^j \|\mathbf{x}_0 - \mathbf{x}^*\|^2 + \frac{\beta R}{\alpha r \sigma^2_{\min+}(\mathbf{A})} \|\mathbf{e}\|^2, \tag{4.1}$$

where $\rho(\mathbf{A}, T) = 1 - \frac{r\sigma^2_{\min+}(\mathbf{A})}{\beta d}$ and $\sigma_{\min+}(\mathbf{A})$ is the smallest nonzero singular value of $\mathbf{A}$.

Remark 4.4. The convergence horizon term, $\frac{\beta R}{\alpha r \sigma^2_{\min+}(\mathbf{A})} \|\mathbf{e}\|^2$, is minimized in the case that $\mathbf{e} = \mathbf{e}^*$ where $\mathbf{e}^*$ is the minimum norm residual, i.e.,

$$\mathbf{e}^* = \text{argmin } \|\mathbf{e}\|^2 \text{ s.t. } \mathbf{e} = \mathbf{A}\mathbf{x} - \mathbf{b} \text{ for some } \mathbf{x} \in \mathbb{R}^n.$$

In this case, the iterates converge to $\mathbf{x}^* := \left(\mathbf{I} - \mathbf{A}^\dagger\mathbf{A}\right)\mathbf{x}_0 + \mathbf{A}^\dagger(\mathbf{b}+\mathbf{e}^*)$.

Remark 4.5. In other Kaczmarz literature, including Needell's original result for Randomized Kaczmarz applied to inconsistent linear systems [63, Theorem 2.1], the problem is formulated by adding some vector of measurement noise $\mathbf{r}$ to a consistent system $\mathbf{A}\mathbf{x} = \mathbf{b}$, leading to a convergence horizon term proportional to $\|\mathbf{r}\|$. We instead work with $\mathbf{A}\mathbf{x} = \mathbf{b}$

being inconsistent and give a horizon proportional to a residual, but we note that the two formulations are equivalent.

Remark 4.6. We include a generalization of Theorem 4.3 to the case when the right hand side, $\mathbf{b}$, is varying in each iteration according to mean-zero randomly distributed additive noise in Proposition 4.12.

4.1.5 Related Work

In this subsection we offer some related reading in the fields of average consensus, gossip protocols, and block iterative methods, and draw connections between them and our work.

Average consensus and gossip protocols. As mentioned, average consensus has been a fundamental topic in distributed computing since the inception of the field. We refer the reader to the classical work of DeGroot [13] for an inception of the consensus problem, and to the work of Tsitsiklis, Bertsekas, and Athans [78] for a first look into stochastic protocols for distributed computing. As networks have grown larger in size and have appeared in more applications, the need for more efficient average consensus solvers motivated the development of *gossip algorithms*: protocols that, in general, select some subset of nodes and allow them to 'gossip', i.e., share and average their stored values. Boyd et al.'s 2006 paper [4] provided a fundamental exposition of said protocols, particularly on the connection between their convergence rate and the underlying network topology, and has motivated research in the topic ever since [15].

There have been a variety of works analyzing different node-selection protocols for gossiping. In [7], at each epoch a path of nodes in the network is formed and the values along said path are averaged. In [53], the network is decomposed beforehand into cliques (connected subgraphs), and at each epoch one such clique is activated. Lastly in [4], at each epoch a selection of pairs are chosen and each pair computes its own average (we call this *edge independent set* gossiping throughout). These analyses are, however, somewhat disjoint, so we believe the unified convergence analysis presented in this paper (which covers all of the aforementioned methods) to be novel.

Block randomized Kaczmarz. The Kaczmarz method [43] is an iterative linear system solver whose popularity boomed after its randomized variant was proven to have exponential convergence by Strohmer and Vershynin [75]. After this work proved convergence for full-rank, consistent systems, further work was done to generalize to the case of inconsistent [63] and rank-deficient [86] systems.

A well-studied family of variants are *block Kaczmarz methods*, in which iterates are projected onto subspaces corresponding to blocks of rows rather than single equations. Early references include [19, 21, 71], but we focus on the block randomized Kaczmarz method introduced by Needell and Tropp [64]. The authors proved that under certain restrictions on the choice of blocks, the method achieves exponential convergence, and converges up to a threshold if the system is inconsistent. In our work we significantly relax these conditions and achieve a similar convergence guarantee.

The connection between gossip algorithms and Kaczmarz methods for linear systems was analyzed in depth by Loizou and Richtárik in their 2019 paper [55]; others considering this connection include [30, 87]. This connection was exploited in [55] to build a framework giving new convergence guarantees for gossip protocols, accelerations via momentum, and other interesting discussions such as dual gossip algorithms. In our work we hone in on their general exposition of block gossip algorithms, and give more explicit links and convergence guarantees for previously mentioned existing gossip protocols.

4.2 Convergence of Block Randomized Kaczmarz

In this section, we prove our main result, Theorem 4.3, which illustrates that the block randomized Kaczmarz method converges at least linearly in expectation on least-squares problems even in the case that the matrix $\mathbf{A}$ is rank-deficint and the blocks are not sampled from a matrix paving. We will then illustrate how this result specializes to prove Corollary 4.1.

Proof of Theorem 4.3. First, note that by definition, $\mathbf{e} = \mathbf{A}\mathbf{x}_\mathbf{e} - \mathbf{b}$ and so

$$\begin{aligned}\mathbf{A}\mathbf{x}^* - \mathbf{b} &= (\mathbf{A}\mathbf{x}_0 - \mathbf{A}\mathbf{A}^\dagger\mathbf{A}\mathbf{x}_0) + \mathbf{A}\mathbf{A}^\dagger\mathbf{b} + \mathbf{A}\mathbf{A}^\dagger\mathbf{e} - \mathbf{b} \\ &= \mathbf{A}\mathbf{A}^\dagger\mathbf{b} + \mathbf{A}\mathbf{A}^\dagger(\mathbf{A}\mathbf{x}_\mathbf{e} - \mathbf{b}) - \mathbf{b} \\ &= \mathbf{A}\mathbf{x}_\mathbf{e} - \mathbf{b} = \mathbf{e},\end{aligned} \tag{4.2}$$

where the second and third equation used the fact that $\mathbf{A}\mathbf{A}^\dagger\mathbf{A} = \mathbf{A}$.

Recall that our updates take the form $\mathbf{x}_{j+1} = \mathbf{x}_j + \mathbf{A}_\tau^\dagger(\mathbf{b}_\tau - \mathbf{A}_\tau\mathbf{x}_j)$, where τ is chosen uniformly at random from our set of blocks. We then have

$$\begin{aligned}\|\mathbf{x}_{j+1} - \mathbf{x}^*\|^2 &= \left\|\mathbf{x}_j + \mathbf{A}_\tau^\dagger(\mathbf{b}_\tau - \mathbf{A}_\tau\mathbf{x}_j) - \mathbf{x}^*\right\|^2 \\ &= \left\|(\mathbf{I} - \mathbf{A}_\tau^\dagger\mathbf{A}_\tau)\mathbf{x}_j + \mathbf{A}_\tau^\dagger(\mathbf{A}_\tau\mathbf{x}^* - \mathbf{e}_\tau) - \mathbf{x}^*\right\|^2 \\ &= \left\|(\mathbf{I} - \mathbf{A}_\tau^\dagger\mathbf{A}_\tau)(\mathbf{x}_j - \mathbf{x}^*) - \mathbf{A}_\tau^\dagger\mathbf{e}_\tau\right\|^2 \\ &= \left\|(\mathbf{I} - \mathbf{A}_\tau^\dagger\mathbf{A}_\tau)(\mathbf{x}_j - \mathbf{x}^*)\right\|^2 + \left\|\mathbf{A}_\tau^\dagger\mathbf{e}_\tau\right\|^2,\end{aligned} \tag{4.3}$$

where we used that $\mathbf{A}_\tau^\dagger(\mathbf{b}_\tau + \mathbf{e}_\tau) = \mathbf{A}_\tau^\dagger\mathbf{A}_\tau\mathbf{x}^*$ and that $\mathrm{Im}(\mathbf{A}_\tau^\dagger)$ and $\mathrm{Im}(\mathbf{I} - \mathbf{A}_\tau^\dagger\mathbf{A}_\tau)$ are orthogonal. Then since $\mathbf{I} - \mathbf{A}_\tau^\dagger\mathbf{A}_\tau$ is an orthogonal projector, we have

$$\left\|(\mathbf{I} - \mathbf{A}_\tau^\dagger\mathbf{A}_\tau)(\mathbf{x}_j - \mathbf{x}^*)\right\|^2 = \|\mathbf{x}_j - \mathbf{x}^*\|^2 - \left\|\mathbf{A}_\tau^\dagger\mathbf{A}_\tau(\mathbf{x}_j - \mathbf{x}^*)\right\|^2. \tag{4.4}$$

Note that $\left\|\mathbf{A}_\tau^\dagger\mathbf{e}_\tau\right\|^2 \leq \sigma_{\max}^2(\mathbf{A}_\tau^\dagger)\|\mathbf{e}_\tau\|^2 = \frac{1}{\sigma_{\min+}^2(\mathbf{A}_\tau)}\|\mathbf{e}_\tau\|^2 \leq \frac{1}{\alpha}\|\mathbf{e}_\tau\|^2$. Using this fact and taking expectations, we obtain

$$\mathbb{E}_j\left(\|\mathbf{x}_{j+1} - \mathbf{x}^*\|^2\right) \tag{4.5}$$

$$\begin{aligned}&\leq \|\mathbf{x}_j - \mathbf{x}^*\|^2 - \mathbb{E}_j\left(\left\|\mathbf{A}_\tau^\dagger\mathbf{A}(\mathbf{x}_j - \mathbf{x}^*)\right\|^2\right) + \frac{1}{\alpha}\mathbb{E}_j\left(\|\mathbf{e}_\tau\|^2\right) \\ &\leq \|\mathbf{x}_j - \mathbf{x}^*\|^2 - \frac{1}{\beta}\mathbb{E}_j\left(\|\mathbf{A}_\tau(\mathbf{x}_j - \mathbf{x}^*)\|^2\right) + \frac{1}{\alpha}\mathbb{E}_j\left(\|\mathbf{e}_\tau\|^2\right) \qquad (4.6) \\ &\leq \|\mathbf{x}_j - \mathbf{x}^*\|^2 - \frac{r}{\beta d}\|\mathbf{A}(\mathbf{x}_j - \mathbf{x}^*)\|^2 + \frac{R}{\alpha d}\|\mathbf{e}\|^2,\end{aligned}$$

where the last inequality follows from the fact that

$$\mathbb{E}_j(\|\mathbf{v}_\tau\|^2) = \frac{1}{d}\sum_{l=1}^{d}\|\mathbf{v}_{\tau_l}\|^2 = \frac{1}{d}\sum_{l=1}^{d}\sum_{i=1}^{m}\mathbf{1}(i\in\tau_l)v_i^2$$
$$= \frac{1}{d}\sum_{i=1}^{m}\left[\sum_{l=1}^{d}\mathbf{1}(i\in\tau_l)\right]v_i^2$$

so we have $\frac{r}{d}\|\mathbf{v}\|^2 \le \mathbb{E}_j(\|\mathbf{v}_\tau\|^2) \le \frac{R}{d}\|\mathbf{v}\|^2$.

We now claim that for all j, $\mathbf{x}_j - \mathbf{x}^* \in \mathrm{Im}(\mathbf{A}^T)$, i.e., the row space of $\mathbf{A}$. We do so by induction; firstly for $j = 0$ we have

$$\mathbf{x}_0 - \mathbf{x}^* = \mathbf{x}_0 - (\mathbf{I} - \mathbf{A}^\dagger\mathbf{A})\mathbf{x}_0 - \mathbf{A}^\dagger(\mathbf{b}+\mathbf{e}) = \mathbf{A}^\dagger(\mathbf{A}\mathbf{x}_0 - (\mathbf{b}+\mathbf{e})),$$

and since $\mathrm{Im}(\mathbf{A}^\dagger) = \mathrm{Im}(\mathbf{A}^T)$, we are done.

Now assume $\mathbf{x}_l - \mathbf{x}^* \in \mathrm{Im}(\mathbf{A}^T)$. Then we have, for some τ,

$$\mathbf{x}_{l+1} - \mathbf{x}^* = \mathbf{x}_l + \mathbf{A}_\tau^\dagger(\mathbf{b}_\tau - \mathbf{A}_\tau\mathbf{x}_l) - \mathbf{x}^*.$$

By assumption $\mathbf{x}_l - \mathbf{x}^* \in \mathrm{Im}(\mathbf{A}^T)$, and furthermore since $\mathbf{A}_\tau$ is a row submatrix of $\mathbf{A}$, we have

$$\mathrm{Im}(\mathbf{A}_\tau^\dagger) = \mathrm{Im}(\mathbf{A}_\tau^T) \subseteq \mathrm{Im}(\mathbf{A}^T).$$

Thus $\mathbf{x}_{l+1} - \mathbf{x}^* \in \mathrm{Im}(A^T)$, and we are done by induction.

Now, returning to (4.5), since $\mathbf{x}_j - \mathbf{x}^* \in \mathrm{Im}(\mathbf{A}^T) = \mathrm{Ker}(\mathbf{A})^\perp$, we have

$$\|\mathbf{A}(\mathbf{x}_j - \mathbf{x}^*)\|^2 \ge \sigma_{\min+}^2(\mathbf{A})\|\mathbf{x}_j - \mathbf{x}^*\|.$$

This yields

$$\mathbb{E}_j\left(\|\mathbf{x}_{j+1} - \mathbf{x}^*\|^2\right) \le \rho(\mathbf{A},T)\|\mathbf{x}_j - \mathbf{x}^*\|^2 + \frac{R}{\alpha d}\|\mathbf{e}\|^2,$$

where $\rho(\mathbf{A}, T) = 1 - \frac{r\sigma^2_{\min+}(\mathbf{A})}{\beta d}$ and by induction we obtain

$$\begin{aligned}
&\mathbb{E}\left(\|\mathbf{x}_{j+1} - \mathbf{x}^*\|^2\right) && (4.7)\\
&\leq \rho(\mathbf{A}, T)^{j+1} \|\mathbf{x}_0 - \mathbf{x}^*\|^2 + \left[\sum_{i=0}^{j} \rho(\mathbf{A}, T)^i\right] \frac{R}{\alpha d} \|\mathbf{e}\|^2 \\
&\leq \rho(\mathbf{A}, T)^{j+1} \|\mathbf{x}_0 - \mathbf{x}^*\|^2 + \left[\sum_{i=0}^{\infty} \rho(\mathbf{A}, T)^i\right] \frac{R}{\alpha d} \|\mathbf{e}\|^2 && (4.8)\\
&= \rho(\mathbf{A}, T)^{j+1} \|\mathbf{x}_0 - \mathbf{x}^*\|^2 + \frac{\beta R}{\alpha r \sigma^2_{\min+}(\mathbf{A})} \|\mathbf{e}\|^2 .
\end{aligned}$$

□

Remark 4.7. We note that in most applications, including our application of average consensus, $\sigma_{\min+}(\mathbf{A})$ is fixed, and so it is natural to seek to maximize $\frac{r}{\beta d}$. For the average consensus problem, this equates to careful selection of the block set T.

We next include a proof of Corollary 4.1 which follows from Theorem 4.3 due to the fact that block gossip on the network $\mathcal{G}$ with initial unknown node values $\mathbf{c}$ coincides with block randomized Kaczmarz on the AC problem.

Proof of Corollary 4.1. Consider running block RK with $\mathbf{A} = \mathbf{Q}, \mathbf{b} = \mathbf{0}$, and $\mathbf{x}_0 = \mathbf{c}$ with samples $\tau \in T$ determined by the run of block gossip on $\mathcal{G}$ with $\mathbf{c}_0 = \mathbf{c}$. Note that $\ker(\mathbf{Q})$ is nonempty since $\operatorname{rank}(\mathbf{Q}) \leq n - 1$, so $\mathbf{e} = \mathbf{0}$. It follows from [55, Theorem 5] that block gossip iterate $\mathbf{c}_k$ and block RK iterate $\mathbf{x}_k$ coincide for all k.

If we take $\mathbf{x}_0 = \mathbf{c}$, then by Theorem 4.3 we know that block RK will converge to $(\mathbf{I} - \mathbf{Q}^\dagger\mathbf{Q})\mathbf{c}$ (since $\mathbf{b} = \mathbf{e} = \mathbf{0}$ in this application). This is exactly the orthogonal projection of $\mathbf{c}$ onto $\ker(\mathbf{Q}) = \operatorname{span}\{\mathbf{1}\}$, where $\mathbf{1}$ is the length-$|\mathcal{E}|$ vector of all ones. Then, since $\left\{\frac{1}{\sqrt{|\mathcal{E}|}}\mathbf{1}\right\}$ is an

orthonormal basis for $\ker(\mathbf{Q})$, we can compute this projection as

$$(\mathbf{I} - \mathbf{Q}^\dagger \mathbf{Q})\mathbf{c} = \left\langle \mathbf{c}, \frac{1}{\sqrt{|\mathcal{E}|}}\mathbf{1} \right\rangle \frac{1}{\sqrt{|\mathcal{E}|}}\mathbf{1} \tag{4.9}$$

$$= \frac{1}{|\mathcal{E}|}\left(\sum_{i=1}^{|\mathcal{E}|} c_i\right)\mathbf{1} \tag{4.10}$$

$$= \mathbf{c}^*. \tag{4.11}$$

Finally, note that $\sigma^2_{\min+}(\mathbf{Q}) = \lambda_{\min+}(\mathbf{L}) = \lambda_2(\mathcal{G})$ where $\mathbf{L}$ is the Laplacian matrix of $\mathcal{G}$. The specific results enumerated in Corollary 4.1 follow from the singular value upper bounds presented in Section 4.3. □

4.3 Block Gossip Sampling

In this section, we consider particular cases when the blocks used in the block gossip method correspond to special subgraph structures, namely independent edge sets, clique subgraphs, path subgraphs, and arbitrary connected subgraphs.

We begin with a lemma that will be used to strengthen our convergence results for these cases.

Fact 4.8. Let τ be a subset of edges of $\mathcal{G}$, and let $\tau' \subseteq \tau$ be the edge set of a spanning tree of $\mathcal{G}_\tau$. Then the block gossip updates produced by choosing blocks τ and τ' are identical.

This follows immediately from the fact that $\mathcal{G}_\tau$ and $\mathcal{G}_{\tau'}$ have the same vertex set, and block gossip simply averages the stored values of said vertex set at each iteration. The value of Fact 4.8 comes from the fact that our theoretical convergence rate bound depends (in part) on the maximum singular values of our blocks: removing rows from a matrix (i.e., using a subset of edges in our block) decreases said singular values, tightening the bound.

Throughout this section, we make use of the fact that for any collection of row indices τ, the spectra of $\mathbf{Q}_\tau \mathbf{Q}_\tau^\top$ and $\mathbf{L}_\tau$ are the same, up to zeros. In particular, we can identify our covering constants α and β by analyzing the spectrum of $\mathbf{L}_\tau$, which is well understood for

many of the graph structures we will consider.

4.3.1 Independent edge set blocks

In the case that $T = \{\tau_1, \tau_2, \cdots, \tau_d\}$ is a row covering of $\mathbf{Q}$ where each $\mathcal{G}_{\tau_i}$ is an independent edge set, we have that

$$\mathbf{Q}_{\tau_i}\mathbf{Q}_{\tau_i}^\top = 2\mathbf{I}$$

for each $i \in [d]$. Thus, T is a $(d, 2, 2, r, R)$ row covering of $\mathbf{Q}$ and we have that

$$\left(1 - \frac{r\lambda_2(\mathcal{G})}{\beta d}\right) = \left(1 - \frac{r\lambda_2(\mathcal{G})}{2d}\right).$$

Note that each independent each set is its own spanning tree, so there is no improvement to be made here via Fact 4.8.

4.3.2 Path blocks

In the case that $T = \{\tau_1, \tau_2, \cdots, \tau_d\}$ is a row covering of $\mathbf{Q}$ where each $\mathcal{G}_{\tau_i}$ is a path, we make use of the following fact about the eigenvalues of the Laplacian of a path subgraph (see e.g., [27]).

Fact 4.9. Let P_n be a path subgraph of G of length n. Then the eigenvalues of L_{P_n} are

$$2 - 2\cos\frac{\pi k}{n} \quad \text{for } k = 0, 1, \cdots, n-1. \tag{4.12}$$

We then have for each $i \in [d]$ that

$$\lambda_{\max}(\mathbf{Q}_{\tau_i}\mathbf{Q}_{\tau_i}^\top) = \lambda_{\max}(\mathbf{Q}_{\tau_i}^\top\mathbf{Q}_{\tau_i}) = 2 - 2\cos\frac{|\tau_i|\pi}{|\tau_i| + 1} \tag{4.13}$$

$$\leq 2 - 2\cos\frac{M\pi}{M+1}, \tag{4.14}$$

so T is a $(d, \alpha, 2 - 2\cos\frac{M\pi}{M+1}, r, R)$ row covering. Thus we have that

$$\left(1 - \frac{r\lambda_2(\mathcal{G})}{\beta d}\right) = \left(1 - \frac{r\lambda_2(\mathcal{G})}{(2 - 2\cos\frac{M\pi}{M+1})d}\right) \leq \left(1 - \frac{r\lambda_2(\mathcal{G})}{4d}\right).$$

Again, each path is its own spanning tree, so this rate cannot be improved via Fact 4.8.

4.3.3 Clique blocks

In the case that $T = \{\tau_1, \tau_2, \cdots, \tau_d\}$ is a row covering of $\mathbf{Q}$ where each $\mathcal{G}_{\tau_i}$ is a complete subgraph of $\mathcal{G}$, we make use of Fact 4.8; in particular, for each $\tau \in T$ there exists $\tau' \subset \tau$ such that $\mathcal{G}_{\tau'}$ is a spanning *path* of G_τ. See Figures 4.2b and 4.2c for an example of how a spanning path block update coincides with the update produced by a clique block update.

In this way, we see that the bound on the convergence rate constant for complete subgraphs must be no larger than that of path subgraphs and we recover the constant

$$\left(1 - \frac{r\lambda_2(\mathcal{G})}{(2 - 2\cos\frac{M\pi}{M+1})d}\right) \leq \left(1 - \frac{r\lambda_2(\mathcal{G})}{4d}\right).$$

4.3.4 Arbitrary connected subgraph blocks

In the case that $T = \{\tau_1, \tau_2, \cdots, \tau_d\}$ is a row covering of $\mathbf{Q}$ where each $\mathcal{G}_{\tau_i}$ is an arbitrary connected subgraph of $\mathcal{G}$, we again make use of the fact that we may replace every block $\tau \in T$ with a block $\tau' \subset \tau$ such that $\mathcal{G}_{\tau'}$ is a spanning tree of $\mathcal{G}_\tau$ to form $T' = \{\tau'_1, \tau'_2, \cdots, \tau'_d\}$. We note that the block gossip method with blocks sampled from T will produce the same set of iterates as those sampled identically from T'.

Now, we use that fact that the eigenvalues of the Laplacian of a tree, $L_{\mathcal{T}}$, are bounded above by $|\mathcal{E}(\mathcal{T})|$ (and that in fact this bound is tight for the star graph), see [27]. Thus, we have

$$\lambda_{\max}(\mathbf{Q}_{\tau'_i}\mathbf{Q}_{\tau'_i}^\top) = \lambda_{\max}(\mathbf{Q}_{\tau'_i}^\top\mathbf{Q}_{\tau'_i}) \leq |\tau_i| \leq M.$$

We use this bound to recover the constant upper bound

$$\left(1 - \frac{r\lambda_2(\mathcal{G})}{Md}\right).$$

4.3.5 Multiple subgraph blocks

If the network allows for multiple disjoint components to be activated at a single instance, we may form blocks consisting of multiple disjoint subgraphs of $\mathcal{G}$. In this case, to compute β we may use the following Lemma.

Lemma 4.10. *Let $\mathcal{G}_\tau$ be a subgraph of $\mathcal{G}$ consisting of disjoint connected subgraphs $\mathcal{G}_{\tau_1}, \cdots, \mathcal{G}_{\tau_k}$. Then*

$$\lambda_{\max}(\mathbf{Q}_\tau \mathbf{Q}_\tau^\top) = \max_i \lambda_{\max}(\mathbf{Q}_{\tau_i} \mathbf{Q}_{\tau_i}^\top). \tag{4.15}$$

Proof. This follows immediately from the fact that since said subgraphs are edge-disjoint, we have that the Laplacian of their union is the direct sum of their individual Laplacians:

$$\mathbf{L}_\tau = \mathbf{L}_{\tau_1} \oplus \cdots \oplus \mathbf{L}_{\tau_k}, \tag{4.16}$$

and so the spectrum of $\mathbf{L}_\tau$ is exactly the union of the spectra of $\mathbf{L}_{\tau_1}, \cdots, \mathbf{L}_{\tau_k}$. □

Suppose then that we take $T = \{\tau_1, \tau_2, \cdots, \tau_d\}$, where each τ_i is the union of a collection of rows corresponding to disjoint connected subgraphs, say $\tau_i = \bigcup_j \tau_i^j$. Then by Lemma 4.10 we have

$$\lambda_{\max}(\mathbf{Q}_{\tau_i} \mathbf{Q}_{\tau_i}^\top) = \max_j \lambda_{\max}(\mathbf{Q}_{\tau_i^j} \mathbf{Q}_{\tau_i^j}^\top).$$

One may then apply the relevant previous results of Section 4.3 to compute an upper bound on this quantity, yielding β.

4.4 Inconsistent Consensus Models

In this section, we consider two models of inconsistent average consensus where communication across edges is noisy and provide analyses of the natural block gossip method in these cases.

4.4.1 Constant edge communication error

Consider the average consensus problem in the presence of constant edge communication error; that is, some blocks of nodes do not update to local consensus during iterations of the block gossip method, but instead update according to an attempt to satisfy constant edge miscommunication values. We denote $\mathbf{m} \in \mathbb{R}^{|\mathcal{E}|}$ as the edge miscommunication values and consider the block gossip updates under this edge miscommunication to be

$$\mathbf{c}_k = \mathbf{c}_{k-1} + \mathbf{Q}_\tau^\dagger(\mathbf{m}_\tau - \mathbf{Q}_\tau \mathbf{c}_{k-1}). \tag{4.17}$$

We can apply Theorem 4.3 to prove the following corollary. This result yields a guarantee of convergence to a convergence horizon that depends upon the edge miscommunication vector $\mathbf{m}$.

Corollary 4.11. *Suppose graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is connected, $\mathbf{Q} \in \mathbb{R}^{|\mathcal{E}|\times|\mathcal{V}|}$ is the incidence matrix for $\mathcal{G}$, and $T = \{\tau_1, \cdots, \tau_d\}$ is a (d, α, β, r, R) row covering for $\mathbf{Q}$. Then the block gossip method under edge miscommunication $\mathbf{m}$ as defined in (4.17) with blocks determined by T converges at least linearly in expectation to a horizon determined by $\mathbf{m}$ with the guarantee*

$$\mathbb{E}\,\|\mathbf{c}_k - \mathbf{c}^*\|^2 \leq \left(1 - \frac{r\lambda_2(\mathcal{G})}{\beta d}\right)^k \|\mathbf{c} - \mathbf{c}^*\|^2 + \frac{\beta R}{\alpha r \lambda_2(\mathcal{G})}\|\mathbf{m}\|^2,$$

where $\lambda_2(\mathcal{G})$ is the algebraic connectivity of graph $\mathcal{G}$.

Proof. This result follows from Theorem 4.3 where $\mathbf{b} = \mathbf{m}$, $\mathbf{A} = \mathbf{Q}$, and $\mathbf{e} = -\mathbf{m}$. Note that $\mathbf{x}^* = (\mathbf{I} - \mathbf{Q}^\dagger\mathbf{Q})\mathbf{c} + \mathbf{Q}^\dagger(\mathbf{m} - \mathbf{m}) = (\mathbf{I} - \mathbf{Q}^\dagger\mathbf{Q})\mathbf{c} = \mathbf{c}^*$. □

4.4.2 Randomly varying edge communication error

We now consider the average consensus problem in the presence of randomly varying edge communication error. During the block gossip method, blocks do not update to local consensus but instead update to attempt to satisfy the iteration dependent edge miscommunication values. We denote $\mathbf{m}_k \in \mathbb{R}^{|\mathcal{E}|}$ as the edge miscommunication values during the kth iteration and consider the block gossip updates under edge miscommunication to be

$$\mathbf{c}_k = \mathbf{c}_{k-1} + \mathbf{Q}_\tau^\dagger((\mathbf{m}_k)_\tau - \mathbf{Q}_\tau \mathbf{c}_{k-1}). \tag{4.18}$$

We prove a generalization of Theorem 4.3 and use it to prove a guarantee of convergence to a convergence horizon that depends upon the distribution of the edge miscommunication values.

Proposition 4.12. *Let $\mathbf{b} \in r(\mathbf{A})$ with $\mathbf{A}$ not necessarily full rank. Consider running the block Kaczmarz method with matrix $\mathbf{A}$ and vector $\mathbf{b}_k = \mathbf{b} + \mathbf{e}_k$ in the kth iteration; that is*

$$\mathbf{x}_k = \mathbf{x}_{k-1} + \mathbf{A}_\tau^\dagger((\mathbf{b}_k)_\tau - \mathbf{A}_\tau \mathbf{x}_{k-1}).$$

Assume that $\{\mathbf{e}_k\}$ is sampled i.i.d. according to distribution $\mathcal{D}$, $\mathbf{e}_k \sim \mathcal{D}$, with $\mathbb{E}_\mathcal{D}[\mathbf{e}_k] = \mathbf{0}$ and $cov(\mathbf{e}_k) = \mathbb{E}_\mathcal{D}[\mathbf{e}_k \mathbf{e}_k^\top] = \mathbf{\Sigma}$. Let $T = \{\tau_1, \tau_2, \cdots, \tau_d\}$ be a (d, α, β, r, R) row covering of A. Let $\mathbf{x}^* = (\mathbf{I} - \mathbf{A}^\dagger \mathbf{A})\mathbf{x}_0 + \mathbf{A}^\dagger \mathbf{b}$. Then we have

$$\mathbb{E}\|\mathbf{x}_{j+1} - \mathbf{x}^*\|^2 \leq \rho(\mathbf{A}, T)^{j+1} \|\mathbf{x}_0 - \mathbf{x}^*\|^2 + \frac{\beta R}{\alpha r \sigma_{\min +}^2(\mathbf{A})} tr(\mathbf{\Sigma}),$$

where $\rho(\mathbf{A}, T) = 1 - \frac{r\sigma_{\min +}^2(\mathbf{A})}{\beta d}$.

Proof. This proof proceeds in a manner highly similar to that of Theorem 4.3. First, note that in a calculation similar to (4.2), we have that $\mathbf{A}\mathbf{x}^* - \mathbf{b}_k = \mathbf{e}_k$. Then, recalling that the updates take the form $\mathbf{x}_{j+1} = \mathbf{x}_j + \mathbf{A}_\tau^\dagger((\mathbf{b}_k)_\tau - \mathbf{A}_\tau \mathbf{x}_j)$ where τ is chosen uniformly from T, we compute

$$\|\mathbf{x}_{j+1} - \mathbf{x}^*\|^2 = \|(\mathbf{I} - \mathbf{A}_\tau^\dagger \mathbf{A}_\tau)(\mathbf{x}_j - \mathbf{x}^*)\|^2 + \|\mathbf{A}_\tau^\dagger (\mathbf{e}_j)_\tau\|^2$$

in a manner similar to that of (4.3). Using (4.4) and taking expectation with respect to the sampled block in iteration j, τ_j, conditioned on all previously sampled blocks, we have

$$\begin{aligned}
&\mathbb{E}_{\tau_j}\|\mathbf{x}_{j+1}-\mathbf{x}^*\|^2 \\
&\quad = \|\mathbf{x}_j-\mathbf{x}^*\|^2 - \mathbb{E}_j\|\mathbf{A}_\tau^\dagger\mathbf{A}_\tau(\mathbf{x}_j-\mathbf{x}^*)\|^2 + \frac{1}{\alpha}\mathbb{E}_j\|(\mathbf{e}_j)_\tau\|^2 \\
&\quad \leq \left(1-\frac{r\sigma^2_{\min+}(\mathbf{A})}{\beta d}\right)\|\mathbf{x}_j-\mathbf{x}^*\|^2 + \frac{R}{\alpha d}\|\mathbf{e}_j\|^2.
\end{aligned}$$

Now, taking expectation with respect to the sampled error $\mathbf{e}_j \sim \mathcal{D}$ conditioned upon all previously sampled errors, we arrive at

$$\begin{aligned}
&\mathbb{E}_{\mathbf{e}_j}\left[\mathbb{E}_{\tau_j}\|\mathbf{x}_{j+1}-\mathbf{x}^*\|^2\right] \\
&\quad \leq \left(1-\frac{r\sigma^2_{\min+}(\mathbf{A})}{\beta d}\right)\|\mathbf{x}_j-\mathbf{x}^*\|^2 + \frac{R}{\alpha d}\mathbb{E}_{\mathbf{e}_j}\|\mathbf{e}_j\|^2 \\
&\quad = \left(1-\frac{r\sigma^2_{\min+}(\mathbf{A})}{\beta d}\right)\|\mathbf{x}_j-\mathbf{x}^*\|^2 + \frac{R}{\alpha d}\mathbf{tr}(\mathbf{\Sigma}).
\end{aligned}$$

Iterating this expectation and proceeding inductively as in (4.7), we arrive at the desired result. □

We may now use this result to prove the following guarantee for convergence of block gossip methods in the presence of randomly varying edge miscommunication error.

Corollary 4.13. *Suppose graph $\mathcal{G}=(\mathcal{V},\mathcal{E})$ is connected, $\mathbf{Q}\in\mathbb{R}^{|\mathcal{E}|\times|\mathcal{V}|}$ is the incidence matrix for $\mathcal{G}$, and $T=\{\tau_1,\cdots,\tau_d\}$ is a (d,α,β,r,R) row covering for $\mathbf{Q}$. Assume that $\{\mathbf{m}_k\}$ is sampled i.i.d. according to distribution $\mathcal{D}$, $\mathbf{m}_k\sim\mathcal{D}$, with $\mathbb{E}_{\mathcal{D}}[\mathbf{m}_k]=0$ and $cov(\mathbf{m}_k)=\mathbb{E}_{\mathcal{D}}[\mathbf{m}_k\mathbf{m}_k^\top]=\mathbf{\Sigma}$. Then the block gossip method under randomly varying edge miscommunication $\mathbf{m}_k$ as defined in (4.18) with blocks determined by T converges at least linearly in expectation to a horizon determined by $tr(\mathbf{\Sigma})$ with the guarantee*

$$\mathbb{E}\,\|\mathbf{c}_k-\mathbf{c}^*\|^2 \leq \left(1-\frac{r\lambda_2(\mathcal{G})}{\beta d}\right)^k\|\mathbf{c}-\mathbf{c}^*\|^2 + \frac{\beta R}{\alpha r\lambda_2(\mathcal{G})}tr(\mathbf{\Sigma}),$$

where $\lambda_2(\mathcal{G})$ is the algebraic connectivity of graph $\mathcal{G}$.

Proof. This result follows from Proposition 4.12 where $\mathbf{b} = \mathbf{0}$, $\mathbf{e}_k = \mathbf{m}_k$, and $\mathbf{A} = \mathbf{Q}$. Note that $\mathbf{x}^* = (\mathbf{I} - \mathbf{Q}^\dagger \mathbf{Q})\mathbf{c} = \mathbf{c}^*$. □

Remark 4.14. In the case that the blocks consist of single rows, then updates (4.17) and (4.18) correspond to nodes updating to satisfy a misspecified (nonhomogenous) equation in the AC system, which could model link communication failure. However, for larger blocks, the interpretation of these updates break down and it is less clear that they model a natural gossip process, as the individual node value updates are produced by the collection of edge miscommunications. We note, however, that such nonhomogenous systems of equations arise in practice elsewhere, e.g., in rank aggregration from pairwise comparisons via Massey's method [58] or Hodgerank [42].

4.5 Experiments

In this section we present empirical results from applying block gossip with various choices of block structure to AC problems on multiple graph structures, including Erdös-Rényi graphs of varying connectivity, square lattice graphs, and the Les Misérables network [46]. All experiments were conducted in Python 3.8 with the NetworkX [35] package used to generate and work with graph structures. We also present results from applying said protocols to inconsistent consensus models as detailed in Section 4.4.

4.5.1 Preliminaries

Recall that an Erdös-Rényi graph ER(n,p) on n vertices is formed by randomly including edges between each pair of nodes independently with probability p. We choose to experiment on such graphs as they are popular models for real-life networks and highlight the effects of varying connectivity (by varying p) on the convergence of the considered gossip protocols. We also run experiments on $n \times n$ square lattice graphs, another widely-studied network structure; see e.g., [7]. In each experiment, initial unknown values are drawn randomly for each node from a Uniform$(0, 1)$ distribution.

For all graphs we perform experiments with four block sampling protocols: independent edge sets (IES), cliques, paths, and randomly selected blocks of fixed size. These protocols and the graph structures underlying them are detailed in Subsection 4.1.4 and Section 4.3.

To produce an IES cover we use a greedy algorithm that repeatedly finds the largest independent edge set, then removes it from the graph until there are no remaining edges. Similarly, a clique edge cover of the graph is generated by a greedy algorithm that repeatedly finds the largest clique, then removes it from the graph until there are no remaining edges. For path gossip, paths are formed by selecting a node uniformly at random, adding a randomly selected neighbour to it, and continuing to sequentially add neighbours until we have a path of the desired length l. Randomly selected blocks are sampled by selecting edges uniformly at random to form a block of a specified size. The blocks generated by these algorithms are then passed into our block gossip algorithm, which randomly samples a block from the list of blocks at each iteration. We note that these block enumeration processes are not part of the block gossip method, they are simply a necessary step for our simulations.

We produce two types of plot: *collapse plots*, which are a visualization of individual node values by iteration (as in [6]), and *error plots*, which show the error at each iteration, $\|\mathbf{c}_k - \mathbf{c}^*\|$, and for some examples also display the predicted upper bound on convergence given by Corollary 4.1.

4.5.2 Erdös-Rényi Graphs

We apply each of our block sampling protocols to Erdös-Rényi graphs ER(n, p) with $p = 0.2, 0.4, 0.6, 0.8, 1$ and $n = 200$.

In Figures 4.3 to 4.6 we compare the performance of each protocol (with path length $l = 10$ for path gossip) across a range of p. It appears that, in the ER case, connectivity does not have a substantial effect on the relative performance of our protocols, with IES gossip consistently being the strongest by some margin. This aligns with the fact that independent edge sets will have the greatest *node* overlap, particularly when compared to cliques – in

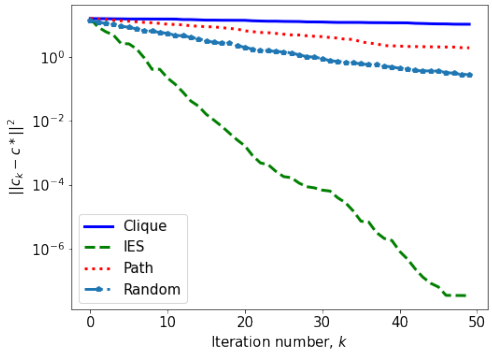


Figure 4.3: Errors for all protocols applied to ER$(200, 0.2)$.

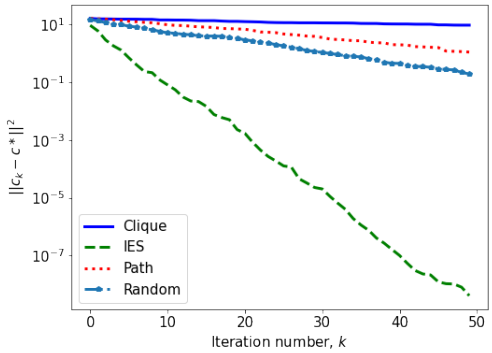


Figure 4.4: Errors for all protocols applied to ER$(200, 0.4)$.

the sense that a single node is likely to be in many more independent edge sets than cliques – and so a larger amount of information is transferred per iteration.

In Figures 4.7 to 4.10, we present a closer look at best (IES) and worst (clique) performing protocols from the previous experiments. The dramatic difference in convergence rate can be partially explained by the collapse plots: we see that during clique gossip certain nodes will hold their value for many iterations before updating, leading to dramatically slower convergence than IES gossip seen in the error plots. This is again connected to the greater amount of node overlap that persists in IES blocks, compared to cliques. We see from the error plots that both protocols respect the upper bound on convergence given by Corollary 4.1, and that said bound predicts that IES gossip should outperform clique gossip.

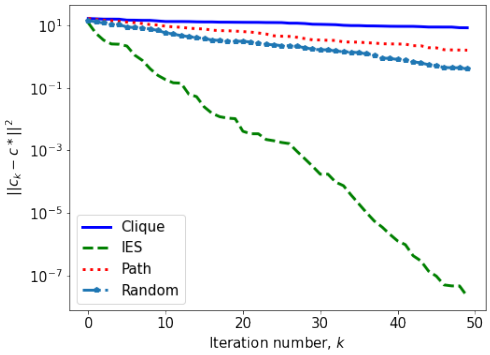


Figure 4.5: Errors for all protocols applied to ER$(200, 0.6)$.

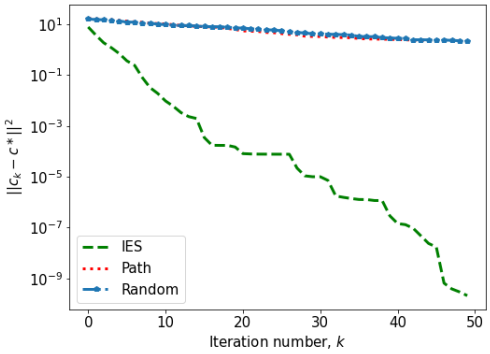


Figure 4.6: Errors for all protocols applied to ER$(200, 1)$.

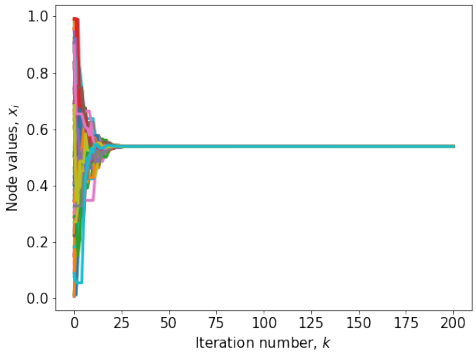


Figure 4.7: Collapse plot for ER$(100, 0.6)$ under IES gossip.

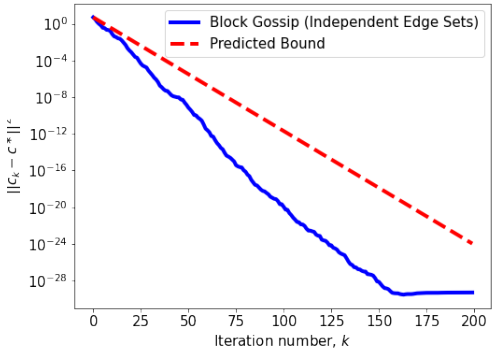


Figure 4.8: Error plot for ER$(100, 0.6)$ under IES gossip.

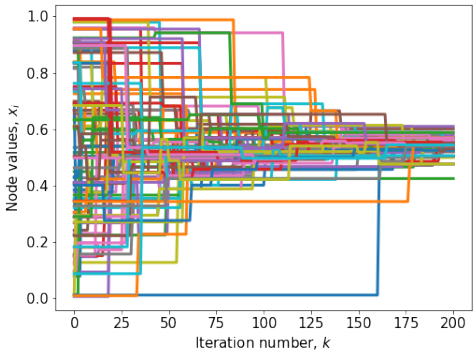


Figure 4.9: Collapse plot for ER$(100, 0.6)$ under clique gossip.

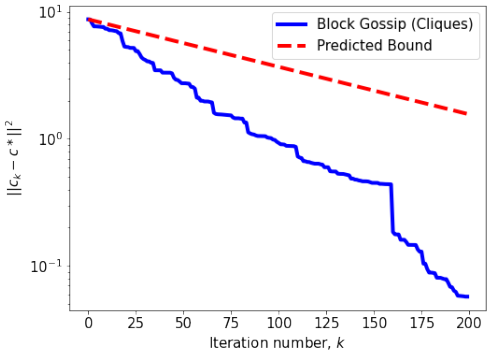


Figure 4.10: Error plot for ER$(100, 0.6)$ under clique gossip.

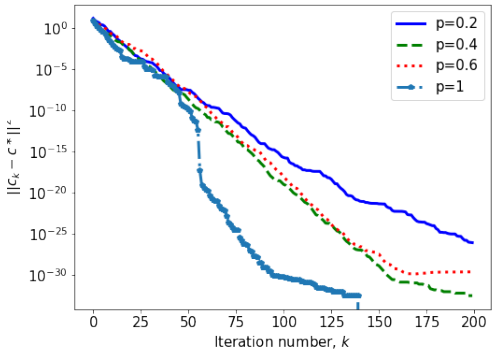


Figure 4.11: IES gossip errors on ER$(200, p)$ for various p.

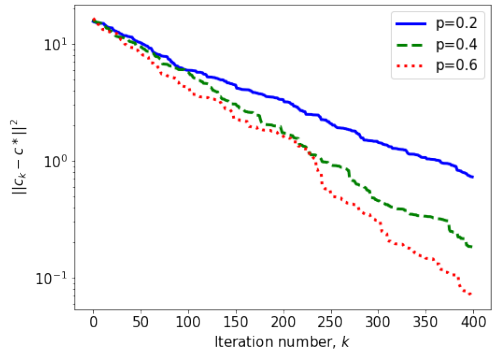


Figure 4.12: Clique gossip errors on ER$(200, p)$ for various p.

We analyze the effect of increasing p (and thus the connectivity of the graph) in Figures 4.11 to 4.14. It can be seen that IES gossip is both fast and robust to variations in connectivity compared to other protocols. This can be attributed heuristically to the fact that independent edge sets are formed by selecting edges which separate nodes well, rather than selecting edges which join nodes well (as in forming cliques). The performance of clique gossip improves significantly with p, corresponding to the fact that ER(n, p) is likely to have a greater number of larger cliques as p, and thus its average degree, increases. Note that we exclude clique gossip for $p = 1$, as the entire graph will be selected as a clique and consensus will be reached in a single iteration.

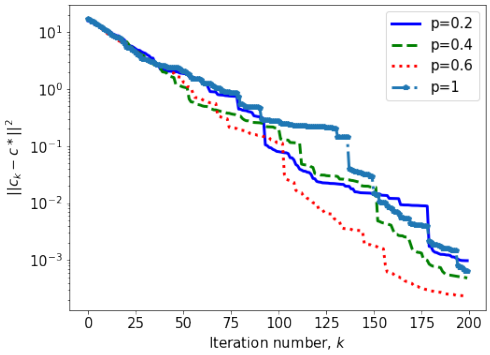


Figure 4.13: Path gossip errors on ER$(200, p)$ for various p.

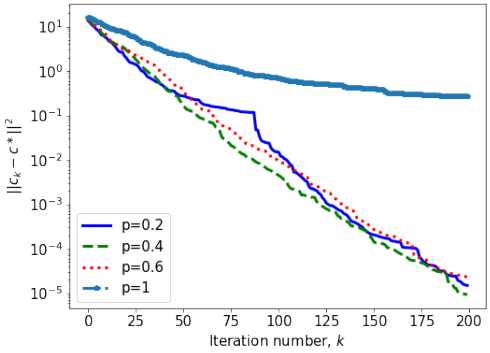


Figure 4.14: Random gossip errors on ER$(200, p)$ for various p.

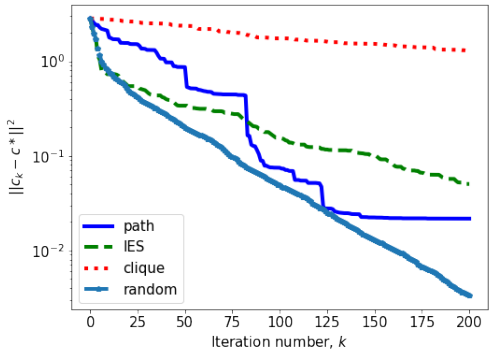


Figure 4.15: Comparison of sampling protocols applied to a 10×10 square lattice.

4.5.3 Square Lattice

In Figure 4.15 we show the error plots from each of our sampling protocols applied to the 10×10 square lattice. The graph structure restricts the size of cliques to only single edges, leading to poor performance versus other protocols. Moreover, path gossip struggles as nodes on opposite sides of the grid are a large graph distance apart. In this scenario, with our 'special' structures being limited, it is sensible to choose blocks at random to attempt to maximise the dispersement of information, and this strategy indeed yields the best performance.

4.5.4 Real World Network

We apply each of our block sampling protocols to the Les Misérables network [46]. This network represents co-occurences of characters in the novel "Les Misérables" by Victor Hugo. Nodes represent the 77 characters, and an edge between two nodes indicates that these two characters appear in the same chapter of the novel. There are 254 edges and the graph is connected. We plot the gossip method errors with independent edge sets, paths, cliques, and arbitrary sets and the upper bound on convergence given by Corollary 4.1 in Figures 4.16, 4.17, 4.18, and 4.19, respectively. Note that the best performing protocol (gossip with independent edge sets) also has the lowest bound on convergence; however,

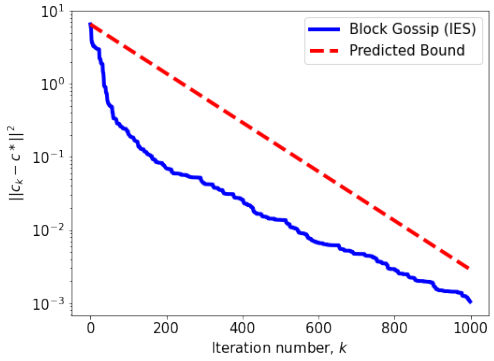


Figure 4.16: Error plot for Les Misérables graph under independent edge set gossip.

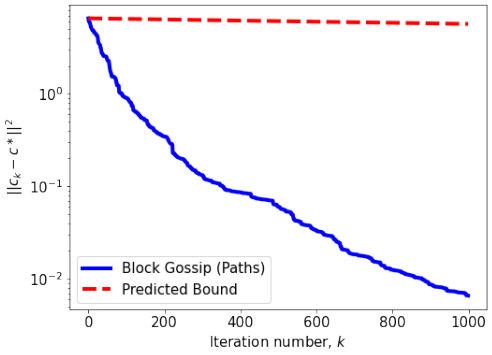


Figure 4.17: Error plot for Les Misérables graph under path gossip.

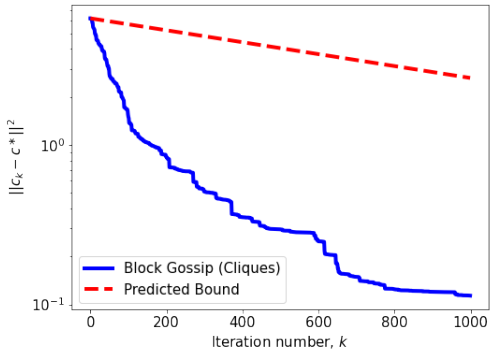


Figure 4.18: Error plot for Les Misérables graph under clique gossip.

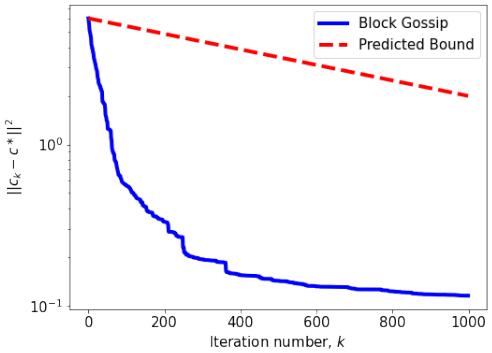


Figure 4.19: Error plot for Les Misérables graph under block gossip with arbitrary subsets.

the next best performing protocol on this network (path gossip) has the worst bound on convergence.

4.5.5 Inconsistent Consensus Models

We experimented with the two types of inconsistent average consensus systems described in Section 4.4, namely systems with a constant edge communication error (CECE) and systems with a randomly varying edge communication error (VECE). In CECE, all iterations are affected by the same edge miscommunication values $\mathbf{m} \in \mathbb{R}^{|\mathcal{E}|}$ which is generated (in advance) with entries sampled independently from $\mathcal{N}(0, 0.01)$, the mean-zero Gaussian distribution with variance 0.01. In VECE, the edge miscommunication for the kth iteration,

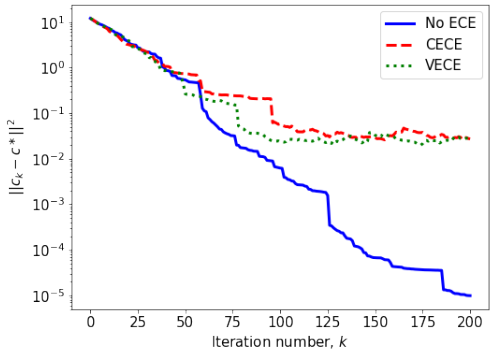


Figure 4.20: Comparison of effect of differ-ent communication errors on path gossip.

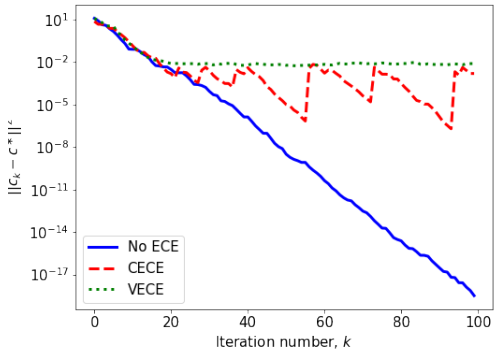


Figure 4.21: Comparison of effect of differ-ent communication errors on IES gossip.

$\mathbf{m}_k \in \mathbb{R}^{|\mathcal{E}|}$, has entries sampled independently from $\mathcal{N}(0, 0.01)$.

In Figures 4.20 and 4.21, we display the errors for consistent block gossip updates ($\mathbf{m} = \mathbf{0}$), CECE updates (4.17) with constant edge communication values $\mathbf{m}$ sampled as described above, and VECE updates (4.18) with varying edge communication values $\mathbf{m}_k$ sampled as described above, on an ER(n, p) graph with $n = 150$ and $p = 0.6$ with path block sampling and IES block sampling, respectively. Note that IES block sampling gossip converges far more quickly than path block sampling, and so we are able to see more clearly the effect of varying error in this plot. Note that VECE has a fairly smooth convergence horizon, whereas the convergence horizon for CECE varies widely as the updates sample the (fixed) values of edge miscommunication of differing magnitude.

In Figures 4.22 and 4.23, we display the errors for consistent block gossip updates ($\mathbf{m} = \mathbf{0}$), CECE updates (4.17) with constant edge communication values $\mathbf{m}$ sampled as described above, and VECE updates (4.18) with varying edge communication values $\mathbf{m}_k$ sampled as described above, on an ER(n, p) graph with $n = 150$ and $p = 0.6$ with clique block sampling and random block sampling, respectively. Note that clique block gossip and random block gossip converge so slowly that it takes far longer to see clearly the convergence horizon for the CECE and VECE updates.

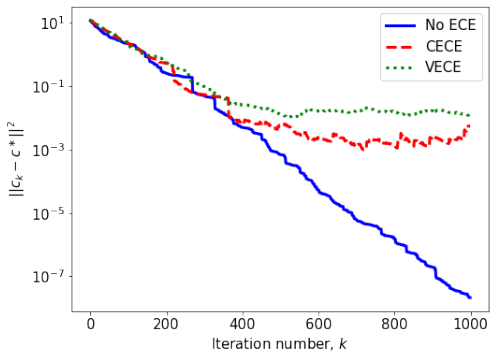


Figure 4.22: Comparison of effect of different communication errors on clique gossip.

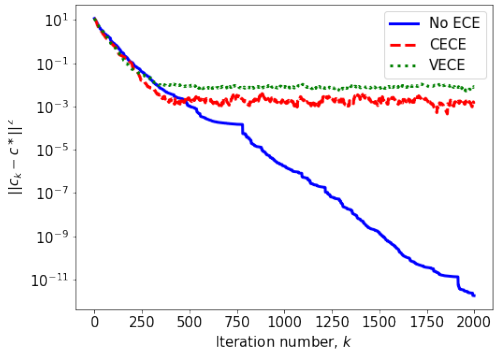


Figure 4.23: Comparison of effect of different communication errors on random gossip.

4.6 Conclusions

In this work, we prove a new convergence result for the block gossip method for the average consensus problem and specialize this theoretical result to *path* [7], *clique* [53], and *edge-independent set* [4] gossip protocols. We prove this result by exploiting the fact that these methods are generalized by the block randomized Kaczmarz method for solving linear systems.

We prove a generalized convergence result for the block randomized Kaczmarz method which generalizes the main result of [64] to the case of rank-deficient systems and relaxes requirements on the set of blocks to be sampled. While these generalization are highly important for the average consensus problem, we expect that they will be of interest in other applications as well.

We additionally prove convergence results for the block gossip method on inconsistent consensus models, and perform a broad set of experiments to compare our theoretical results to the empirical behavior of the block gossip methods on various network structures.

Future directions include further exploration of inconsistent average consensus models, including random link failure and adversarial nodes, bounded confidence models, which generalize the average consensus model, and block Kaczmarz variants for randomly varying noise.

CHAPTER 5

Online Signal Recovery via Heavy Ball Kaczmarz

This chapter is a version of [41] and is joint work with Yotam Yaniv and Prof. Deanna Needell. I proposed the project and Deanna Needell supervised. Yotam Yaniv and I contributed the convergence analysis and experiments. We present an application of the Kaczmarz method to an online signal recovery setting, where instead of sampling rows from a fixed system of linear equation, one samples linear measurements from some general source distribution. When such settings arise, as in medical imaging, these measurements are frequently highly coherent (that is, have corresponding hyperplanes with small incident angles). We introduce a variant of the Kaczmarz method with an additional heavy-ball momentum term, and show via theoretical results and experiments that it handles this particular setting well.

5.1 Introduction

5.1.1 The Kaczmarz Method

Recovering a signal $x^* \in \mathbb{R}^n$ from a collection of linear measurements is an important problem in computerized tomography [60], sensor networks [72], compressive sensing [20, 23], machine learning subroutines [3], and beyond. When the collection of linear measurements is finite, say of size m, and accessible at any time, the problem is equivalent to solving a system of linear equations $Ax = b$ with $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$, which has been well-studied. A popular method for solving this classical problem is the Kaczmarz method [43]: beginning with an initial iterate x_0, at each iteration a row of the system is sampled and the previous iterate is projected onto the hyperplane defined by the solution space

given by that row. More precisely, if the row $a_i^\top x = b_i$ is sampled at iteration k, the update has the form

$$x_k = x_{k-1} - \frac{\langle a_i, x_{k-1} \rangle - b_i}{\|a_i\|^2} a_i.$$

The original method proposed cycling through rows in order, such that $i = k \bmod m$. In [37] it was observed empirically that randomized row selection accelerates convergence, and in the landmark work [75] it was proven that selecting rows at random with probability proportional to their Euclidean norm yields linear convergence in expectation.

In this work, we consider an online model in which at each discrete time $t = 1, 2, \dots$ a linear measurement $(\varphi_t, y_t) \in \mathbb{R}^n \times \mathbb{R}$ is received. We assume that each measurement is noiseless, i.e. $\langle \varphi_t, x^* \rangle = y_t$ for all t, and that measurements are *streamed* through memory and are not stored. Note that the linear system setting described above is a special case of this model, but we now allow for measurements to be sampled from a more general source. The Kaczmarz method is well-suited to this setting as it requires access to only a single measurement at each iteration. See, for example, [10], where measurement data is viewed as being sampled i.i.d. from some distribution $\mathcal{D}$ on $\mathbb{R}^n$. We assume the noiseless, i.i.d. setting throughout this paper. A Kaczmarz update in this setting has the following form, when initialized with some arbitrary x_0: at discrete times $t = 1, 2, \dots$, a measurement $(\varphi_t, y_t) \in \mathbb{R}^n \times \mathbb{R}$ is received, where $\varphi_t \sim \mathcal{D}$, and a Kaczmarz iteration is computed

$$x_t = x_{t-1} - \frac{\langle \varphi_t, x_{t-1} \rangle - y_t}{\|\varphi_t\|^2} \varphi_t.$$

In [10] it was shown that under certain conditions on $\mathcal{D}$, the method enjoys linear convergence in expectation. Further related works have placed online Kaczmarz in the context of learning theory [51], and have analyzed sparse online variants [48, 57]. Random vector models have also appeared in analyses of Kaczmarz methods for phase retrieval [77] and for sparsely corrupted data [34].

5.1.2 Heavy Ball Momentum

Heavy ball momentum is a popular addition to gradient descent methods, in which an additional step is taken in the direction of the previous iteration's movement. Proposed initially in [70], it has proven very popular in machine learning [76, 47, 26, 81], with a guarantee of linear convergence for stochastic gradient methods with heavy ball momentum proven in [54] (improving on earlier sublinear guarantees in [84, 25]). A gradient descent method itself [67], the Kaczmarz method may be modified with heavy ball momentum to give updates of the following form:

$$x_{t+1} = x_t - \frac{\langle \varphi_t, x_t \rangle - y_t}{\|\varphi_t\|^2} \varphi_t + \beta(x_t - x_{t-1}),$$

where $\beta \geq 0$ is a momentum parameter. In [54] it was shown that when applied to a linear system (i.e., when each φ_t is sampled from the rows of a matrix A), the Kaczmarz method with heavy ball momentum converges linearly in expectation. Experimental results indicate accelerated convergence compared to the standard Kaczmarz method on a range of datasets, while the momentum term does not affect the order of the computational cost.

In this work, we propose an *online* variant of the Kaczmarz method with heavy ball momentum. We prove that our method converges linearly in expectation for a wide range of distributions $\mathcal{D}$, and offer particular examples. This theory is supported by numerical experiments on both synthetic and real-world data, which in particular demonstrate the benefits of adding momentum when measurements are highly coherent.

5.2 Proposed Method & Empirical Results

We propose an online variant of the Kaczmarz method, modified to include a heavy ball momentum term $\beta \in (0, 1)$, which we call OHBK(β) (see Algorithm 7). We note that our method is a generalization of the momentum Kaczmarz method for systems of linear equations introduced in [54]. The method requires only a single measurement to be held in storage at a time, while leveraging information about previous measurements through

the momentum term.

Algorithm 7 Online Heavy Ball Kaczmarz

procedure OHBK(β) (Input: initial iterate x_0, measurements $\{(\varphi_t, y_t)\}_{t=1}^{\infty}$, momentum parameter β)
  Set $x_1 = x_0$
  for $t = 1, 2, \ldots$ **do**
    Update $x_{t+1} = x_t - \frac{\langle \varphi_t, x_t \rangle - y_t}{\|\varphi_t\|^2} \varphi_t + \beta(x_t - x_{t-1})$
  end for
end procedure

We test our method on synthetic and real-world data. For each data source, we compare our method OHBK(β) for a variety of β to an online Kaczmarz method without momentum, which we denote by OK (equivalently, OHBK(0)).

We first experiment on synthetic data. We sample $x^* \in \mathbb{R}^{50}$ with standard Gaussian entries, and take $\{\varphi_t\}_{t=1}^{\infty}$ to be vectors of length 50 with $U[0, 1]$ entries. We note that this process produces particularly coherent data, that is, the vectors $\{\varphi_t\}_{t=1}^{\infty}$ have small pairwise inner products. Each y_t is then computed as $y_t = \langle \varphi_t, x^* \rangle$ to ensure measurements are noiseless. In Figure 5.2 we perform a parameter search over 100 trials for β and plot the median error after 100 iterations versus β with shading for the 25th through 75th percentiles. Introducing some amount of momentum provides an acceleration, however, taking β to be too large places too much weight on previous information and is less effective. In Figure 5.1 we show convergence down to machine epsilon of OHBK(β) versus online randomized Kaczmarz (i.e. OHBK(0)) for a selection of β (averaging over 10 trials), and the acceleration provided by momentum is clear.

In Figure 5.3, we investigate the effect of momentum on highly coherent systems further. We perform 4000 iterations of OHBK(β) on $U[\varepsilon, 1]$ signals of length 50, for $\varepsilon \in [0, 1]$, for a range of momentum parameters β (again averaged over 10 trials). We see that momentum provides a significant speedup in convergence even for highly coherent systems (i.e. for large ϵ). However, as $\epsilon \to 1$, recovering the signal becomes intractable.

We compare the effect of the signal length n on the optimal momentum parameter β in Figure 5.4. We perform parameter searches for signals of length $n \in \{50, 100, 500, 1000\}$ and

mark the optimal values of β. The optimal choice of β does not appear to vary significantly with n.

In Figure 5.5 we use a system generated from the Wisconsin Diagnostic Breast Cancer (WDBC) dataset, where each measurement is computed from a digitized image of a fine needle aspirate of a breast mass and describes characteristics of the cell nuclei present [17]. We stream through each measurement of the 699-row, 10-feature dataset once to replicate the online model, and again see that the addition of momentum provides a noteworthy acceleration to convergence.

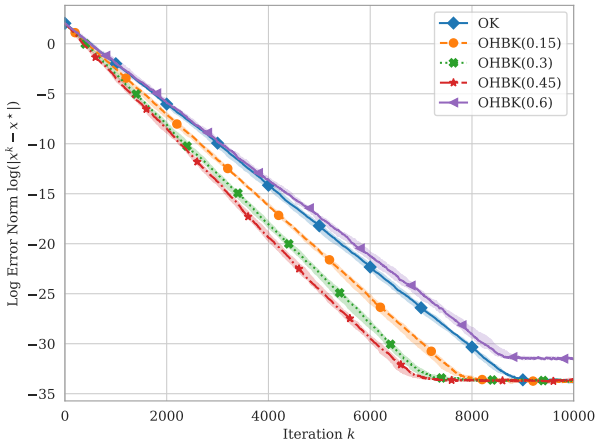


Figure 5.1: Error versus iteration for OHBK(β) applied to $U[0,1]$ signals of length 50.

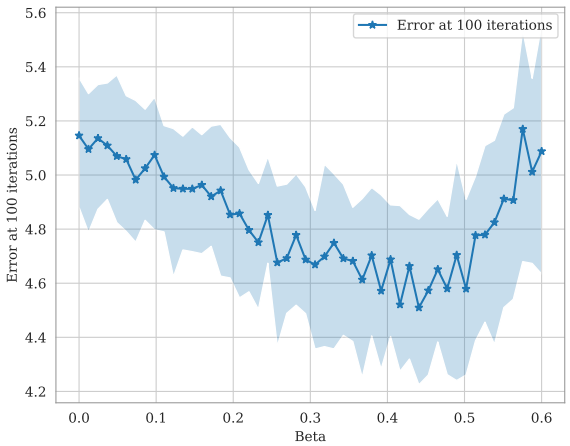


Figure 5.2: $\|x_{100} - x^*\|$ versus β for a range of $\beta \in [0, 0.6]$, for $U[0,1]$ signals of length 50.

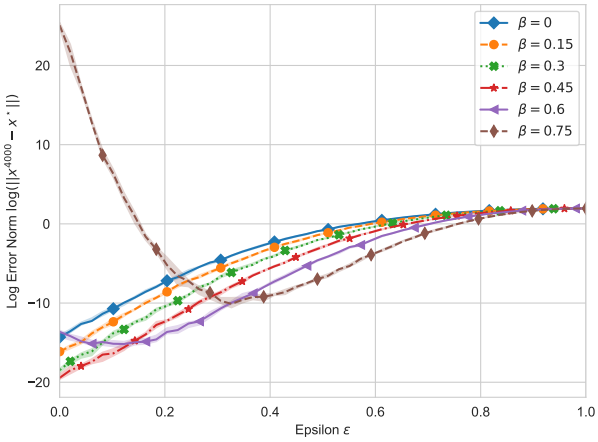


Figure 5.3: $\log \|x_{4000} - x^*\|$ versus ε for OHBK(β) applied to $U[\varepsilon, 1]$ signals of length 50.

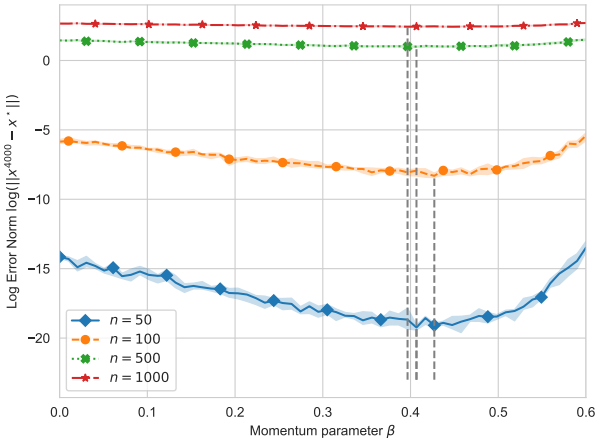


Figure 5.4: $\log \|x_{4000} - x^*\|$ versus β for OHBK(β) applied to $U[0, 1]$ signals of length n. The gray verticals show the value of β yielding the minimum error.

5.3 Theoretical Results

Throughout our theory, we assume that $\{\varphi_t\}_{t=1}^{\infty}$ is a sequence of independent samples from some distribution $\mathcal{D}$. We provide a general linear convergence (in expectation) result with a rate depending on the matrix $W := \mathbb{E}_{\mathcal{D}}\left[\frac{\varphi\varphi^\top}{\|\varphi\|^2}\right]$, in particular on its smallest and largest singular values $\sigma_{\min}(W)$ and $\sigma_{\max}(W)$.

Theorem 5.1 (Convergence in Expectation of OHBRK)**.** *Suppose that measurement vectors* $\{\varphi_t\}_{t=1}^{\infty}$ *are sampled independently from* $\mathcal{D}$*, and* $W = \mathbb{E}_{\mathcal{D}}\left[\frac{\varphi\varphi^\top}{\|\varphi\|^2}\right]$*. Then if* β *is small enough such*

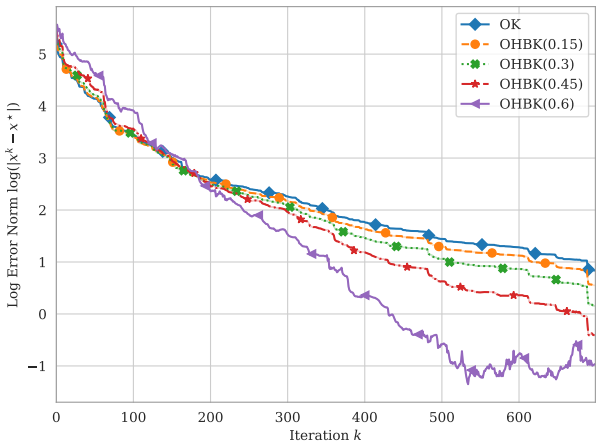


Figure 5.5: Error versus iteration for OHBK(β) applied to the WDBC dataset.

that

$$4\beta + 4\beta^2 - (1+\beta)\sigma_{\min}(W) + \beta\sigma_{\max}(W) < 0,$$

the iterates produced by OHBK(β) satisfy the following guarantee: for some $\delta > 0$, $q \in (0,1)$, we have

$$\mathbb{E}[\|x_t - x^*\|^2] \leq q^t(1+\delta)\|x_0 - x^*\|^2.$$

More interpretable conditions on β may be obtained for particular classes of distribution $\mathcal{D}$. In particular, if $\varphi/\|\varphi\|$ is distributed uniformly on the unit sphere (which is the case if $\mathcal{D}$ itself is the uniform distribution on the unit sphere, or if $\mathcal{D}$ is the standard n-dimensional Gaussian), then $W = \frac{1}{n}I$ and we require

$$\beta + \beta^2 < \frac{1}{4n}$$

to guarantee linear convergence in expectation.

5.4 Proof of Main Result

In this section we prove Theorem 5.1 by following the steps of ([54], Theorem 1), making modifications for the online case and simplifications to some of the constants for our special case. First we present a lemma from [54] which we will use in our convergence proof.

Lemma 5.2 ([54], Lemma 9)**.** *Let $\{F_t\}_{t\geq 0}$ be a sequence of non-negative real numbers with $F_0 = F_1$ that satisfies the relation $F_{t+1} \leq a_1 F_t + a_2 F_{t-1}$ for all $t \geq 1$, with $a_2 > 0$ and $a_1 + a_2 < 1$. Then the following inequality hold for all $t \geq 1$*

$$F_{t+1} \leq q^t(1+\delta)F_0,$$

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

A proof of this lemma can be found in [54].

We begin our convergence analysis by writing the squared $L2$ error at the $(t+1)^{\text{th}}$ iteration and substituting the OHBK(β) update into it,

$$\|x_{t+1} - x^*\|^2 = \left\| x_t - \frac{\langle \varphi_t, x_t\rangle - y_t}{\|\varphi_t\|^2}\varphi_t + \beta(x_t - x_{t-1}) - x^* \right\|^2.$$

Next, we group our equation into three terms:

$$\|x_{t+1} - x^*\|^2 = \left\| x_t - x^* - \frac{\langle \varphi_t, x_t\rangle - y_t}{\|\varphi_t\|^2}\varphi_t \right\|^2 + \beta^2 \|x_t - x_{t-1}\|^2 \tag{5.1}$$

$$+ 2\beta\langle x_t - x^* - \frac{\langle \varphi_t, x_t\rangle - y_t}{\|\varphi_t\|^2}\varphi_t, x_t - x_{t-1}\rangle. \tag{5.2}$$

We bound the first term of Equation (5.2) by following a standard Kaczmarz convergence argument and the fact that $y_t = \langle \varphi_t, x^*\rangle$. We have that

$$\begin{aligned}
\left\| x_t - x^* - \frac{\langle \varphi_t, x_t\rangle - y_t}{\|\varphi_t\|^2}\varphi_t \right\|^2 &= \|x_t - x^*\|^2 + \left\| \frac{\langle \varphi_t, x_t\rangle - y_t}{\|\varphi_t\|^2}\varphi_t \right\|^2 - 2\left\langle \frac{\langle \varphi_t, x_t\rangle - y_t}{\|\varphi_t\|^2}\varphi_t, x_t - x^* \right\rangle \\
&= \|x_t - x^*\|^2 + \frac{(\langle \varphi_t, x_t\rangle - y_t)^2}{\|\varphi_t\|^2} - 2\frac{(\langle \varphi_t, x_t\rangle - y_t)^2}{\|\varphi_t\|^2} \\
&= \|x_t - x^*\|^2 - \frac{(\langle \varphi_t, x_t\rangle - y_t)^2}{\|\varphi_t\|^2}.
\end{aligned}$$

We bound the second term of Equation (5.2) by first adding and subtracting x^*

$$\beta^2 \|x_t - x_{t-1}\|^2 = \beta^2 \|(x_t - x^*) + (x^* - x_{t-1})\|^2.$$

Then by applying the fact that $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$ we have that

$$\beta^2 \|(x_t - x^*) + (x^* - x_{t-1})\|^2 \leq 2\beta^2 \|x_t - x^*\|^2 + 2\beta^2 \|x_{t-1} - x^*\|.$$

Thus we have that

$$\beta^2 \|x_t - x_{t-1}\|^2 \leq 2\beta^2 \|x_t - x^*\|^2 + 2\beta^2 \|x_{t-1} - x^*\|.$$

Finally we bound the third term of Equation (5.2) as

$$\begin{aligned}
&2\beta \left\langle x_t - x^* - \frac{\langle \varphi_t, x_t\rangle - y_t}{\|\varphi_t\|^2}\varphi_t, x_t - x_{t-1} \right\rangle \\
&= 2\beta\langle x_t - x^*, x_t - x_{t-1}\rangle + 2\beta \left\langle \frac{\langle \varphi_t, x_t\rangle - y_t}{\|\varphi_t\|^2}\varphi_t, x_{t-1} - x_t \right\rangle \\
&= 2\beta \|x_t - x^*\|^2 + 2\beta\langle x_t - x^*, x^* - x_{t-1}\rangle + 2\beta \left\langle \frac{\langle \varphi_t, x_t\rangle - y_t}{\|\varphi_t\|^2}\varphi_t, x_{t-1} - x_t \right\rangle \\
&= \beta \|x_t - x^*\|^2 + \beta \|x_t - x_{t-1}\|^2 - \beta \|x_{t-1} - x^*\|^2 + 2\beta \left\langle \frac{\langle \varphi_t, x_t\rangle - y_t}{\|\varphi_t\|^2}\varphi_t, x_{t-1} - x_t \right\rangle \\
&\leq \beta \|x_t - x^*\|^2 + \beta \|x_t - x_{t-1}\|^2 - \beta \|x_{t-1} - x^*\|^2 - \beta\langle \frac{\langle \varphi_t, x_t\rangle - y_t}{\|\varphi_t\|^2}\varphi_t, x_t - x^*\rangle \\
&\quad + \beta\langle \frac{\langle \varphi_t, x_{t-1}\rangle - y_t}{\|\varphi_t\|^2}\varphi_t, x_{t-1} - x^*\rangle.
\end{aligned}$$

Combining the three bounds, we have

$$\begin{aligned}
\|x_{t+1} - x^*\|^2 \leq \|x_t - x^*\|^2 &- \frac{(\langle \varphi_t, x_t\rangle - y_t)^2}{\|\varphi_t\|^2} + 2\beta^2 \|x_t - x^*\|^2 + 2\beta^2 \|x_{t-1} - x^*\|^2 + \beta \|x_t - x^*\|^2 \\
&+ \beta \|x_t - x_{t-1}\|^2 - \beta \|x_{t-1} - x^*\|^2 - \beta \left\langle \frac{\langle \varphi_t, x_t\rangle - y_t}{\|\varphi_t\|^2}\varphi_t, x_t - x^* \right\rangle \\
&+ \beta \left\langle \frac{\langle \varphi_t, x_{t-1}\rangle - y_t}{\|\varphi_t\|^2}\varphi_t, x_{t-1} - x^* \right\rangle.
\end{aligned}$$

Simplifying and grouping like terms we have

$$\|x_{t+1} - x^*\|^2 \leq (1 + 2\beta^2 + \beta)\|x_t - x^*\|^2 + (2\beta^2 - \beta)\|x_{t-1} - x^*\|^2 - \frac{(\langle \varphi_t, x_t \rangle - y_t)^2}{\|\varphi_t\|^2}$$
$$+ \beta\|x_t - x_{t-1}\|^2 - \beta\langle \frac{\langle \varphi_t, x_t \rangle - y_t}{\|\varphi_t\|^2}\varphi_t, x_t - x^* \rangle + \beta\langle \frac{\langle \varphi_t, x_{t-1} \rangle - y_t}{\|\varphi_t\|^2}\varphi_t, x_{t-1} - x^* \rangle.$$

Applying the simplification for the second term of Equation (5.2) and simplifying the inner products, we have

$$\|x_{t+1} - x^*\|^2 \leq (1 + 2\beta^2 + 3\beta)\|x_t - x^*\|^2 + (2\beta^2 + \beta)\|x_{t-1} - x^*\|^2 - (\beta + 1)\frac{\langle \varphi_t, x_t - x^* \rangle^2}{\|\varphi_t\|^2}$$
$$+ \beta\frac{\langle \varphi_t, x_{t-1} - x^* \rangle^2}{\|\varphi_t\|^2}.$$

Taking an expectation over our signal of our simplified equation, we have

$$\mathbb{E}[\|x_{t+1} - x^*\|^2] \leq (1 + 2\beta^2 + 3\beta)\|x_t - x^*\|^2 + (2\beta^2 + \beta)\|x_{t-1} - x^*\|^2$$
$$- (\beta + 1)\mathbb{E}[\frac{(\langle \varphi_t, x_t - x^* \rangle)^2}{\|\varphi_t\|^2}] + \beta\mathbb{E}[\frac{(\langle \varphi_t, x_{t-1} - x^* \rangle)^2}{\|\varphi_t\|^2}]$$
$$= (1 + 2\beta^2 + 3\beta)\|x_t - x^*\|^2 + (2\beta^2 + \beta)\|x_{t-1} - x^*\|^2$$
$$- (1 + \beta)(x_t - x^*)^T\mathbb{E}\left[\frac{\varphi_t\varphi_t^T}{\|\varphi_t\|^2}\right](x_t - x^*) + \beta(x_{t-1} - x^*)^T\mathbb{E}\left[\frac{\varphi_t\varphi_t^T}{\|\varphi_t\|^2}\right](x_{t-1} - x^*).$$

Let $W := \mathbb{E}\left[\frac{\varphi_t\varphi_t^T}{\|\varphi_t\|^2}\right]$. We can then bound the above in terms of the largest and smallest singular values of W:

$$\begin{aligned}\mathbb{E}[\|x_{t+1}-x^*\|^2] &\leq (1+2\beta^2+3\beta)\|x_t-x^*\|^2+(2\beta^2+\beta)\|x_{t-1}-x^*\|^2 \\ &\quad -(1+\beta)\sigma_{\min}(W)\|x_t-x^*\|^2+\beta\sigma_{\max}(W)\|x_{t-1}-x^*\|^2 \\ &= (1+2\beta^2+3\beta-(1+\beta)\sigma_{\min}(W))\|x_t-x^*\|^2 \\ &\quad +(2\beta^2+\beta+\beta\sigma_{\max}(W))\|x_{t-1}-x^*\|^2.\end{aligned}$$

Finally, we apply Lemma 5.2, wherein the two coefficients are given by $a_1 = 1+2\beta^2+3\beta-(1+\beta)\sigma_{\min}(W)$ and $a_2 = 2\beta^2+\beta+\beta\sigma_{\max}(W)$. Since we assumed that $a_1+a_2 = 1+4\beta^2+4\beta+(1+\beta)\sigma_{\min}(W)+\beta\sigma_{\max}(W) < 1$ and since $\beta > 0$ then $a_2 = 2\beta^2+\beta+\beta\sigma_{\max}(W) > 0$ thus the assumptions for Lemma 5.2 hold, so we have that

$$\mathbb{E}[\|x_t-x^*\|^2] \leq q^t(1+\delta)\|x_0-x^*\|^2$$

where $q = \frac{a_1+\sqrt{a_1^2+4a_1}}{2}$, $\delta = q-a_1$ and $a_1+a_2 \leq q < 1$. Since $q \in (0,1)$ we have shown that the norm squared error of the iterates produced by OHBK(β) converges linearly in expectation.

5.5 Conclusion and Future Directions

In this work we discuss using a Kaczmarz method variant with momentum to solve an online signal recovery problem. We leverage a heavy ball momentum term, a classical acceleration method, to improve the convergence rate. We prove a theoretical convergence rate for OHBK(β), and verify this convergence empirically on both synthetic and real-world data. We demonstrate empirically that for coherent measurements, the addition of momentum indeed accelerates convergence, and provided some initial exploration into the dependence of the convergence rate on the signal length n and momentum strength β.

It is notable that in our convergence analysis, we did not recover a theoretically optimal value for β. Doing so, and comparing this value to empirically best values, would be an

interesting future direction. Furthermore, we would like to obtain theoretical parameter relationships: for example, how the optimal momentum strength depends on the signal length and coherency of the measurements. It may in fact be optimal to adaptively adjust the momentum parameter across iterations based on the current iterate and properties of incoming measurements. Additionally, we would like to leverage other accelerated gradient methods such as ADAM [45]. Finally, we would like to consider solving the online signal recovery problem in the case where each measurement is no longer exact, but instead contains some amount of noise [63]. This could be achieved, for example, using relaxation.

CHAPTER 6

Conclusion

As the volume of data being collected, transmitted, and stored continues to increase, it is inevitable that some of it will be affected by noise or corruption. Subsequently feeding this data into downstream tasks can lead to errors propogating and causing potentially harmful decisions. In this thesis, we studied variants of the Kaczmarz method for solving systems of linear equations that are designed specifically to detect, avoid, and minimize the impact of such errors.

We analyzed the convergence of a quantile-based variant of the Kaczmarz method for corrupted systems of linear equations, QuantileRK, with the addition of a noise term in the measurement data. Our theoretical analysis showed that the residual quantile is sufficiently robust to noise to still guarantee convergence to the solution.

Next, we introduced a block variant of QuantileRK that can better exploit the residual information that is necessarily computed at each iteration of QuantileRK. We showed through both theoretical and empirical analyses that our method enjoys significantly faster convergence than QuantileRK, with very little additional computation. Since data in practice is frequently affected by such corruptions, we discussed implementation considerations for practitioners, including parameter selection and the ability to parallelize the method.

We showed that a popular class of iterative methods for solving the average consensus problem on networks is equivalent to applying a block Kaczmarz method to a particular linear system. In doing so, we generalized existing convergence theory for block Kaczmarz methods to a wider class of linear systems and a wider variety of block structures. We were then able to contribute new convergence theory for gossip protocols on networks, in particular guaranteeing convergence at a certain rate. We additionally discussed the

possibility of future connections between other models in network theory, and stochastic iterative methods for solving systems of linear equations.

Finally, we considered a variant of the Kaczmarz method with an additional heavy ball momentum term. We gave general convergence theory for the method, and showed empirically that our method offers an improvement over the standard Kaczmarz method in the particular case of successive measurements being close to parallel.

Solving systems of linear equations is a widespread component of many more complex models in numerous areas, and improvements in this component lead directly to improvements in downstream tasks. Developing mathematical theory to back up these methods is key to providing practitioners with trustworthy tools, and expanding the range of settings in which these tools are applicable is necessary to keep up with the expanding range of settings that modern practioners are faced with.

BIBLIOGRAPHY

[1] T. C. Aysal, M. E. Yildiz, A. D. Sarwate, and A. Scaglione. Broadcast gossip algorithms for consensus. *IEEE Transactions on Signal Processing*, 57(7):2748–2761, 2009.

[2] Z.-Z. Bai and W.-T. Wu. On greedy randomized Kaczmarz method for solving large sparse linear systems. *SIAM Journal on Scientific Computing*, 40(1):592–606, 2018.

[3] L. Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186, 2010.

[4] S. Boyd, A. Ghosh, B. Prabhakar, and D. Shah. Randomized gossip algorithms. *IEEE Transactions on Information Theory*, 52(6):2508–2530, 2006.

[5] J. Briskman and D. Needell. Block Kaczmarz method with inequalities. *Journal of Mathematical Imaging and Vision*, 52(3):385–396, 2015.

[6] H. Z. Brooks and M. A. Porter. A model for the influence of media on the ideology of content in online social networks. *Physical Review Research*, 2, 2020.

[7] F. Bénézit, A. G. Dimakis, P. Thiran, and M. Vetterli. Order-optimal consensus through randomized path averaging. *IEEE Transactions on Information Theory*, 56(10):5150–5167, 2010.

[8] E. Candes and T. Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215, 2005.

[9] Y. Censor, P. Eggermont, and D. Gordon. Strong underrelaxation in Kaczmarz's method for inconsistent systems. *Numerische Mathematik*, 41:83–92, 1983.

[10] X. Chen and A. M. Powell. Almost sure convergence of the Kaczmarz algorithm with random measurements. *Journal of Fourier Analysis and Applications*, 18:1195–1214, 2012.

[11] L. Cheng, B. Jarman, D. Needell, and E. Rebrova. On block accelerations of quantile randomized Kaczmarz for corrupted systems of linear equations. *Inverse Problems*, 39 (2):024002, 2022.

[12] G. Cybenko. Dynamic load balancing for distributed memory multiprocessors. *Journal of Parallel and Distributed Computing*, 7(2):279–301, 1989.

[13] M. H. DeGroot. Reaching a consensus. *Journal of the American Statistical Association*, 69(345):118–121, 1974.

[14] A. D. G. Dimakis, A. D. Sarwate, and M. J. Wainwright. Geographic gossip: Efficient averaging for sensor networks. *IEEE Transactions on Signal Processing*, 56(3):1205–1216, 2008.

[15] 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.

[16] K. Du, W.-T. Si, and X.-H. Sun. Randomized extended average block Kaczmarz for solving least squares. *SIAM Journal on Scientific Computing*, 42(6):3541–3559, 2020.

[17] D. Dua and C. Graff. UCI machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

[18] N. Durgin, R. Grotheer, C. Huang, S. Li, A. Ma, D. Needell, and J. Qin. Randomized Kaczmarz for support recovery of jointly sparse corrupted multiple measurement vectors. *Research in Data Science, Proc. WiSDM (ICERM)*, 2018.

[19] P. P. Eggermont, G. T. Herman, and A. Lent. Iterative algorithms for large partitioned linear systems, with applications to image reconstruction. *Linear Algebra and its Applications*, 40:37–67, 1981.

[20] Y. Eldar and G. Kutyniok. *Compressed Sensing: Theory and Applications*. Cambridge University Press, 2012.

[21] T. Elfving. Block-iterative methods for consistent and inconsistent linear equations. *Numerische Mathematik*, 35:1–12, 1980.

[22] H. G. Feichtinger, C. Cenker, M. Mayer, H. Steier, and T. Strohmer. New variants of the POCS method using affine subspaces of finite codimension with applications to irregular sampling. In *Visual Communications and Image Processing '92*, volume 1818, pages 299 – 310, 1992.

[23] S. Foucart and H. Rauhut. *A Mathematical Introduction to Compressive Sensing*. Birkhäuser New York, 2013.

[24] N. M. Freris and A. Zouzias. Fast distributed smoothing of relative measurements. In *51st IEEE Conference on Decision and Control (CDC)*, pages 1411–1416, 2012.

[25] S. Gadat, F. Panloup, and S. Saadane. Stochastic heavy ball. *Electronic Journal of Statistics*, 12(1):461–529, 2018.

[26] I. Gitman, H. Lang, P. Zhang, and L. Xiao. Understanding the role of momentum in stochastic gradient methods. *Advances in Neural Information Processing Systems*, 32, 2019.

[27] C. Godsil and G. Royle. *Algebraic graph theory*. Springer New York, NY, 2001.

[28] R. M. Gower and P. Richtárik. Randomized iterative methods for linear systems. *SIAM Journal on Matrix Analysis and Applications*, 36(4):1660–1690, 2015.

[29] R. M. Gower, D. Molitor, J. Moorman, and D. Needell. On adaptive sketch-and-project for solving linear systems. *SIAM Journal on Matrix Analysis and Applications*, 42(2): 954–989, 2021.

[30] J. Haddock and A. Ma. Greed works: An improved analysis of sampling Kaczmarz-Motzkin. *SIAM Journal on Mathematics of Data Science*, 3(1):342–368, 2021.

[31] J. Haddock and D. Needell. On Motzkin's method for inconsistent linear systems. *BIT Numerical Mathematics*, 59:387–401, 2019.

[32] J. Haddock and D. Needell. Randomized projection methods for linear systems with arbitrarily large sparse corruptions. *SIAM Journal on Scientific Computing*, 41(5): S19–S36, 2019.

[33] J. Haddock, B. Jarman, and C. Yap. Paving the way for consensus: Convergence of block gossip algorithms. *IEEE Transactions on Information Theory*, 68(11):7515–7527, 2022.

[34] J. Haddock, D. Needell, E. Rebrova, and W. Swartworth. Quantile-based iterative methods for corrupted systems of linear equations. *SIAM Journal on Matrix Analysis and Applications*, 43(2):605–637, 2022.

[35] A. Hagberg, P. Swart, and D. Schult. Exploring network structure, dynamics, and function using networkx. In *Proceedings of the 7th Python in Science conference (SciPy 2008)*, pages 11–15, 2008.

[36] F. Hanzely, J. Konečnỳ, N. Loizou, P. Richtárik, and D. Grishchenko. Privacy preserving randomized gossip algorithms. *arXiv preprint arXiv:1706.07636*, 2017.

[37] G. Herman and L. Meyer. Algebraic reconstruction techniques can be made computationally efficient (positron emission tomography application). *IEEE Transactions on Medical Imaging*, 12(3):600–609, 1993.

[38] A. Hickok, B. Jarman, M. Johnson, J. Luo, and M. A. Porter. Persistent homology for resource coverage: A case study of access to polling sites. *arXiv preprint, arXiv:2206.04834*, 2022.

[39] G. Hounsfield. Computerized transverse axial scanning (tomograph): Part I. description of system. *British Journal of Radiology*, 46(552):1016–1022, 1973.

[40] B. Jarman and D. Needell. QuantileRK: Solving large-scale linear systems with corrupted, noisy data. In *55th Asilomar Conference on Signals, Systems, and Computers*, pages 1312–1316, 2021.

[41] B. Jarman, Y. Yaniv, and D. Needell. Online signal recovery via heavy-ball Kaczmarz. In *56th Asilomar Conference on Signals, Systems, and Computers*, pages 276–280, 2022.

[42] X. Jiang, L.-H. Lim, Y. Yao, and Y. Ye. Statistical ranking and combinatorial Hodge theory. *Mathematical Programming*, 127:203–244, 2011.

[43] S. Kaczmarz. Angenäherte auflösung von systemen linearer gleichungen. *Bulletin International de l'Académie Polonaise des Sciences et des Lettres*, 35:355–357, 1937.

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

[45] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

[46] D. E. Knuth. *The Stanford GraphBase: a platform for combinatorial computing*. ACM, 1993.

[47] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

[48] Y. Lei and D.-X. Zhou. Learning theory of randomized sparse Kaczmarz method. *SIAM Journal on Imaging Sciences*, 11(1):547–574, 2018.

[49] J. Leskovec, A. Rajaraman, and J. D. Ullman. *Mining of Massive Datasets*. Cambridge University Press, 2014.

[50] P. Li, C. Tseng, Y. Zheng, J. Chew, L. Huang, B. Jarman, and D. Needell. Guided semi-supervised non-negative matrix factorization. *Algorithms*, 15:136, 2022.

[51] J. Lin and D.-X. Zhou. Learning theory of randomized Kaczmarz algorithm. *Journal of Machine Learning Research*, 16(103):3341–3365, 2015.

[52] J. Liu, B. D. Anderson, M. Cao, and A. S. Morse. Analysis of accelerated gossip algorithms. *Automatica*, 49(4):873–883, 2013.

[53] Y. Liu, B. Li, B. D. O. Anderson, and G. Shi. Clique gossiping. *IEEE/ACM Transactions on Networking*, 27(6):2418–2431, 2019.

[54] N. Loizou and P. Richtárik. Momentum and stochastic momentum for stochastic gradient, Newton, proximal point and subspace descent methods. *Computational Optimization and Applications*, 77(3):653–710, 2020.

[55] N. Loizou and P. Richtárik. Revisiting randomized gossip algorithms: General framework, convergence rates and novel block and accelerated protocols. *IEEE Transactions on Information Theory*, 67(12):8300–8324, 2021.

[56] 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, 2016.

[57] D. A. Lorenz, S. Wenger, F. Schöpfer, and M. Magnor. A sparse Kaczmarz solver and a linearized Bregman method for online compressed sensing. In *2014 IEEE International Conference on Image Processing (ICIP)*, pages 1347–1351, 2014.

[58] K. Massey. Statistical models applied to the rating of sports teams. *Bluefield College*, 1997.

[59] J. Moorman, T. Tu, D. Molitor, and D. Needell. Randomized Kaczmarz with averaging. *BIT Numerical Mathematics*, 61:337–359, 2021.

[60] F. Natterer. *The Mathematics of Computerized Tomography*. SIAM, 2001.

[61] I. Necoara. Faster randomized block Kaczmarz algorithms. *SIAM Journal on Matrix Analysis and Applications*, 40(4):1425–1452, 2019.

[62] 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.

[63] D. Needell. Randomized Kaczmarz solver for noisy linear systems. *BIT Numerical Mathematics*, 50(2):395–403, 2010.

[64] D. Needell and J. Tropp. Paved with good intentions: Analysis of a randomized block Kaczmarz method. *Linear Algebra and its Applications*, 441:199–221, 2014.

[65] D. Needell and R. Ward. Two-subspace projection method for coherent overdetermined systems. *Journal of Fourier Analysis and Applications*, 19(2):256–269, 2013.

[66] D. Needell, R. Zhao, and A. Zouzias. Randomized block Kaczmarz method with projection for solving least squares. *Linear Algebra and its Applications*, 484:322–343, 2015.

[67] D. Needell, N. Srebro, and R. Ward. Stochastic gradient descent, weighted sampling, and the randomized Kaczmarz algorithm. *Mathematical Programming*, 155:549–573, 2016.

[68] R. Olfati-Saber, J. A. Fax, and R. M. Murray. Consensus and cooperation in networked multi-agent systems. *Proceedings of the IEEE*, 95(1):215–233, 2007.

[69] A. Olshevsky and J. N. Tsitsiklis. Convergence speed in distributed consensus and averaging. *SIAM Journal on Control and Optimization*, 48(1):33–55, 2009.

[70] B. T. Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17, 1964.

[71] C. Popa. Block-projections algorithms with blocks containing mutually orthogonal rows and columns. *BIT Numerical Mathematics*, 441:323–338, 1999.

[72] A. Savvides, C.-C. Han, and M. B. Strivastava. Dynamic fine-grained localization in ad-hoc networks of sensors. In *Proceedings of the 7th Annual International Conference on Mobile Computing and Networking*, pages 166–179, 2001.

[73] S. Steinerberger. A weighted randomized Kaczmarz method for solving linear systems. *Mathematics of Computation*, 90:2815–2826, 2021.

[74] S. Steinerberger. Quantile-based random Kaczmarz for corrupted linear systems of equations. *Information and Inference: A Journal of the IMA*, 2022.

[75] T. Strohmer and R. Vershynin. A randomized Kaczmarz algorithm with exponential convergence. *Journal of Fourier Analysis and Applications*, 15(2):262–278, 2009.

[76] I. Sutskever, J. Martens, G. Dahl, and G. Hinton. On the importance of initialization and momentum in deep learning. In *Proceedings of the 30th International Conference on Machine Learning*, pages 1139–1147, 2013.

[77] Y. S. Tan and R. Vershynin. Phase retrieval via randomized Kaczmarz: theoretical guarantees. *Information and Inference: A Journal of the IMA*, 8(1):97–123, 2018.

[78] J. Tsitsiklis, D. Bertsekas, and M. Athans. Distributed asynchronous deterministic and stochastic gradient optimization algorithms. *IEEE Transactions on Automatic Control*, 31(9):803–812, 1986.

[79] R. Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018.

[80] E. Weber. Algebraic aspects of the paving and feichtinger conjectures. In *Topics in Operator Theory*, pages 569–578. Birkhäuser Basel, 2010.

[81] H. Xia, V. Suliafu, H. Ji, T. Nguyen, A. Bertozzi, S. Osher, and B. Wang. Heavy ball neural ordinary differential equations. *Advances in Neural Information Processing Systems*, 34:18646–18659, 2021.

[82] L. Xiao, S. Boyd, and S. Lall. A scheme for robust distributed sensor fusion based on average consensus. In *IPSN 2005. Fourth International Symposium on Information Processing in Sensor Networks*, 2005.

[83] L. Xioa, S. Boyd, and S.-J. Kim. Distributed average consensus with least-mean-square deviation. *Journal of Parallel and Distributed Computing*, 67(1):33–46, 2007.

[84] T. Yang, Q. Lin, and Z. Li. Unified convergence analysis of stochastic momentum methods for convex and non-convex optimization. *arXiv preprint arXiv:1604.03257*, 2016.

[85] J. Zhang, J. Cui, Z. Wang, Y. Ding, and Y. Xia. Distributed joint cooperative self-localization and target tracking algorithm for mobile networks. *Sensors*, 19(18), 2019.

[86] A. Zouzias and N. Freris. Randomized extended Kaczmarz for solving least squares. *SIAM Journal on Matrix Analysis and Applications*, 34(2):773–793, 2013.

[87] A. Zouzias and N. M. Freris. Randomized gossip algorithms for solving Laplacian systems. In *2015 European Control Conference (ECC)*, pages 1920–1925, 2015.