Randomized Kaczmarz with Averaging

Jacob D. Moorman[*,1], Thomas K. Tu[*,2], Denali Molitor[*,3] and Deanna Needell[*,4]

[*]Department of Mathematics, University of California, Los Angeles, Los Angeles, California
[1]jdmoorman@math.ucla.edu, [2]thomastu@ucla.edu, [3]dmolitor@ucla.edu, [4]deanna@math.ucla.edu

***Abstract*— The randomized Kaczmarz (RK) method is an iterative method for approximating the least-squares solution of linear systems of equations. The RK method requires sequential updates, making parallel computation difficult. Here, we study a parallel version of RK where a weighted average of independent updates is used. We analyze the convergence of RK with averaging and demonstrate its performance empirically. We show that as the number of threads increases, the rate of convergence improves and the convergence horizon for inconsistent systems decreases.**

I. Introduction

Randomized Kaczmarz (RK) is a popular iterative method for approximating the least-squares solution of large, overdetermined linear systems [Kac37], [SV09]. At each iteration, a row is chosen with some probability and the current approximation is projected onto the solution space of that row.

In order to take advantage of parallel computation and speed up the convergence of RK, we consider a simple extension of the RK method, where at each iteration multiple independent updates are computed in parallel and a weighted average of the updates is used. We analyze the convergence rate of this RK with averaging, and show that increasing the number of rows used in each update improves both the convergence rate and convergence horizon.

A. Problem Statement

Given $\mathbf{A} \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$, we aim to solve the linear system of equations

$$\mathbf{A}x = b \tag{1}$$

which is overdetermined, typically with $m \gg n$. For simplicity, we assume throughout that $\mathbf{A}$ has full rank so that the solution is unique when it exists. However, this assumption can be relaxed by choosing the solution with least norm when multiple solutions exist [ZF13].

When a solution to Equation (1) exists, we denote the solution $x^\star$ and refer to the problem as *consistent*. Otherwise, the problem is *inconsistent*, and $x^\star$ instead denotes the *least-squares* solution:

$$x^\star = \underset{x\in\mathbb{R}^n}{\operatorname{argmin}} \frac{1}{2}\|b - \mathbf{A}x\|_2^2. \tag{2}$$

The least-squares solution can be equivalently written as $x^\star = \mathbf{A}^\dagger b$, where $\mathbf{A}^\dagger$ is the Moore-Penrose pseudoinverse of $\mathbf{A}$. We denote the least-squares *residual* as $r^\star := b - \mathbf{A}x^\star$, which is zero for consistent systems.

B. Algorithms

Let x^k be the k^{th} iterate, $e^k := x^k - x^\star$ be the error of the k^{th} iterate, and $r^k := b - \mathbf{A}x^k$ be the residual of the k^{th} iterate. We use $\mathbf{A}_i$ to denote the i^{th} row of $\mathbf{A}$ and $\|\cdot\| := \|\cdot\|_2$. The *relaxed RK* update is given by

$$x^{k+1} = x^k - \lambda_{k,i_k} \frac{\mathbf{A}_{i_k}x^k - b_{i_k}}{\|\mathbf{A}_{i_k}\|^2}\mathbf{A}_{i_k}^\top, \tag{3}$$

where i_k is sampled from some fixed distribution $\mathcal{D}$ at each iteration and $\lambda_{k,i}$ are relaxation parameters. Fixing $\lambda_{k,i} = 1$ for all iterations k and indices i leads to the standard RK method in which one projects onto the solution space corresponding to the i_k^{th} row of $\mathbf{A}$ at iteration k [SV09]. Choosing relaxation parameters $\lambda_{k,i} \neq 1$ can be used to accelerate convergence or dampen the effect of noise in the linear system [CZT12], [HN90b], [HN90a].

For consistent systems, RK converges exponentially in expectation to the solution $x^\star$ of $\mathbf{A}x = b$ [SV09]. For inconsistent systems, there exists at least one equation $\mathbf{A}_j x = b_j$ that is not satisfied by $x^\star$. As a result RK cannot converge for inconsistent systems, since it will occasionally project onto the solution space of such an equation. One can, however, guarantee convergence in expectation to within a radius of the least-squares solution, commonly referred to as the *convergence horizon* [Nee10], [ZF13], [NT12].

Instead of carrying out updates with respect to a single row sequentially, we consider a weighted average of independent updates, which can be carried out in parallel to improve the efficiency per iteration. Specifically, we

write the averaged RK update

$$x^{k+1} = x^k - \sum_{i \in \tau_k} \frac{w_i}{|\tau_k|} \frac{\mathbf{A}_i x^k - b_i}{\|\mathbf{A}_i\|^2} \mathbf{A}_i^\top, \tag{4}$$

where τ_k is a random sequence of the row indices taken with replacement and w_i represents the weight corresponding to the i^{th} row. RK with averaging is detailed in Algorithm 1. If τ_k is a sequence of length one and the weights are chosen as $w_i = 1$ for $i = 1, \ldots, m$, we recover the standard RK method.

Algorithm 1 Randomized Kaczmarz with Averaging

1: **Input** $\mathbf{A} \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$, $x^0 \in \mathbb{R}^n$, weights $w \in \mathbb{R}^m$, maximum iterations K, distribution $\mathcal{D}$, # of threads $|\tau_k|$
2: **for** $k = 0, \ldots, K-1$ **do**
3: $\quad \tau_k \leftarrow |\tau_k|$ indices sampled from $\mathcal{D}$
4: $\quad$ Compute $\delta \leftarrow \sum_{i \in \tau_k} \frac{w_i}{|\tau_k|} \frac{\mathbf{A}_i x^k - b_i}{\|\mathbf{A}_i\|^2} \mathbf{A}_i^\top$ in parallel
5: $\quad$ Update $x^{k+1} \leftarrow x^k - \delta$
6: **Output** x^K

C. Contributions

We prove that using averaged parallel updates as described in Algorithm 1 with uniform weights improves the guaranteed convergence rate as compared to RK (Corollary 3) and reduces the convergence horizon for inconsistent systems. We provide a general convergence result for Algorithm 1 in Theorem 2.

Our experiments illustrate the improvement in the convergence rate and convergence horizon as the number of threads used per iteration increases. The experiments explore the effects of the relaxation parameter α, the weights w_i and distribution $\mathcal{D}$. We recover the standard convergence for RK when $|\tau_k| = 1$ and appropriate weights and probabilities are chosen [SV09], [Nee10], [ZF13]. For uniform weights, we relate Algorithm 1 to a more general parallel sketch-and-project method discussed in [RT17].

D. Related Work

RK is a well-studied method with many variants [Kac37], [SV09], [Nee10], [ZF13]. We do not provide an exhaustive review of the related literature, but instead only remark on a few closely related parallel extensions of RK.

The CARP algorithm [GG05] distributes rows of $\mathbf{A}$ into blocks. The Kaczmarz method is then applied to the rows contained within each block and a component-averaging operator combines the approximations from each block. While the CARP method is shown to converge for consistent systems and to converge cyclically for inconsistent systems, no exponential convergence rate is given.

AsyRK [LWS14] is an asynchronous parallel RK method that results from applying Hogwild! [NRRW11] to the least-squares objective. In AsyRK, each thread chooses a row $\mathbf{A}_i$ at random and updates a random coordinate within the support of that row $\mathbf{A}_i$ with a weighted RK update. AsyRK is shown to have exponential convergence, given conditions on the step size. Their analysis requires that $\mathbf{A}$ is sparse, while we do not make this restriction here.

RK falls under a more general class of methods often called sketch-and-project methods [GR15]. For a linear system $\mathbf{A}x = b$, sketch-and-project methods iteratively project onto the solution space of a sketched subsystem $\mathbf{S}^\top \mathbf{A}x - \mathbf{S}^\top b$. In particular, RK is a sketch-and-project method with $\mathbf{S}^\top = \mathbf{I}_i$, where $\mathbf{I}_i$ is the i^{th} row of the identity matrix. Other popular iterative methods such as coordinate descent can also be framed as sketch-and-project methods. In [RT17], the authors discuss a more general version of Algorithm 1 for sketch-and-project methods with averaging. Their analysis and discussion, however, focus on consistent systems and require uniform weights. We instead focus on the more general case in which the system may be inconsistent and allow for more general weights w_i.

II. Convergence Results

For inconsistent systems, RK satisfies the error bound

$$\mathbb{E}\left[\|e^{k+1}\|^2\right] \leq \left(1 - \frac{\sigma_{\min}^2(\mathbf{A})}{\|\mathbf{A}\|_F^2}\right) \mathbb{E}\left[\|e^k\|^2\right] + \frac{\|r^\star\|^2}{\|\mathbf{A}\|_F^2}, \tag{5}$$

where $\sigma_{\min}(\mathbf{A})$ is the smallest singular value of $\mathbf{A}$, $\|\mathbf{A}\|_F^2 = \sum_{i,j} \mathbf{A}_{ij}^2$, and $r^\star$ is the least-squares residual [SV09], [Nee10], [ZF13]. Iterating this error bound yields

$$\mathbb{E}\left[\|e^k\|^2\right] \leq \left(1 - \frac{\sigma_{\min}^2(\mathbf{A})}{\|\mathbf{A}\|_F^2}\right)^k \|e^0\|^2 + \frac{\|r^\star\|^2}{\sigma_{\min}^2(\mathbf{A})}.$$

For consistent systems, $r^\star = 0$ and this bound guarantees exponential convergence in expectation at a rate $1 - \frac{\sigma_{\min}^2(\mathbf{A})}{\|\mathbf{A}\|_F^2}$. For inconsistent systems, this bound only guarantees exponential convergence in expectation to within a convergence horizon $\|r^\star\|^2 / \sigma_{\min}^2(\mathbf{A})$.

We derive a convergence result for Algorithm 1 which is similar to Equation (5) and leads to a better convergence rate and a smaller convergence horizon for inconsistent systems when using uniform weights. In

analyzing the convergence, it will be useful to consider the update to the error at each iteration. Subtracting $x^\star$ from both sides of the update rule in Equation (3) and using the fact that $\mathbf{A}_i e^k - r_i^\star = \mathbf{A}_i x^k - b_i$, we derive the update

$$e^{k+1} = e^k - \sum_{i \in \tau_k} \frac{w_i}{|\tau_k|} \frac{\mathbf{A}_i e^k - r_i^\star}{\|\mathbf{A}_i\|^2} \mathbf{A}_i^\top. \tag{6}$$

To simplify notation, we define the following matrices.

Definition 1. *Define the weighted sampling matrix*

$$\mathbf{M}_k := \sum_{i \in \tau_k} \frac{w_i}{|\tau_k|} \frac{\mathbf{I}_i^\top \mathbf{I}_i}{\|\mathbf{A}_i\|^2},$$

where τ_k is a sequence of indices sampled independently from $\mathcal{D}$ with replacement and $\mathbf{I}$ is the identity matrix.

Using Definition 1, the error update in Equation (6) can be rewritten as

$$e^{k+1} = (\mathbf{I} - \mathbf{A}^\top \mathbf{M}_k \mathbf{A}) e^k + \mathbf{A}^\top \mathbf{M}_k r^\star. \tag{7}$$

Definition 2. *Let $\mathbf{Diag}(d_1, d_2, \ldots, d_m)$ denote the diagonal matrix with $d_1, d_2, \ldots d_m$ on the diagonal. Define the normalization matrix*

$$\mathbf{D} := \mathbf{Diag}(\|\mathbf{A}_1\|, \|\mathbf{A}_2\|, \ldots, \|\mathbf{A}_m\|)$$

so that the matrix $\mathbf{D}^{-1}\mathbf{A}$ has rows with unit norm, the probability matrix

$$\mathbf{P} := \mathbf{Diag}(p_1, p_2, \ldots, p_m),$$

where $p_j = \mathbb{P}(i = j)$ with $i \sim \mathcal{D}$, and the weight matrix

$$\mathbf{W} := \mathbf{Diag}(w_1, w_2, \ldots, w_m).$$

The convergence analysis additionally relies on the expectations given in Lemma 1, whose proof can be found in Appendix I.

Lemma 1. *Let $\mathbf{M}_k, \mathbf{P}, \mathbf{W}$, and $\mathbf{D}$ be defined as in Definitions 1 and 2. Then*

$$\mathbb{E}[\mathbf{M}_k] = \mathbf{P}\mathbf{W}\mathbf{D}^{-2}$$

and

$$\mathbb{E}\left[\mathbf{M}_k^\top \mathbf{A}\mathbf{A}^\top \mathbf{M}_k\right] = \frac{1}{|\tau_k|}\mathbf{P}\mathbf{W}^2\mathbf{D}^{-2} + \left(1 - \frac{1}{|\tau_k|}\right)\mathbf{P}\mathbf{W}\mathbf{D}^{-2}\mathbf{A}\mathbf{A}^\top\mathbf{P}\mathbf{W}\mathbf{D}^{-2}.$$

A. Coupling of Weights and Probabilities

Note that

$$\lim_{|\tau_k| \to \infty} \mathbf{M}_k = \mathbb{E}[\mathbf{M}_k].$$

Therefore, as $|\tau_k| \to \infty$, the RK with averaging update approaches the deterministic update

$$x^{k+1} = (\mathbf{I} - \mathbf{A}^\top \mathbb{E}[\mathbf{M}_k]\mathbf{A}) x^k + \mathbf{A}^\top \mathbb{E}[\mathbf{M}_k] b.$$

Since we want the method to converge to the least-squares solution, we should require that it have $x^\star$ as a fixed point. However, any fixed point x must solve

$$\mathbf{A}^\top \mathbb{E}[\mathbf{M}_k]\mathbf{A}x = \mathbf{A}^\top \mathbb{E}[\mathbf{M}_k] b, \tag{8}$$

which corresponds to minimizing $\frac{1}{2}\|b - \mathbf{A}x\|^2_{\mathbb{E}[\mathbf{M}_k]}$. This coincides with the least-squares solution defined in Equation (2) only if Assumption 1 holds.

Assumption 1. *The probability matrix $\mathbf{P}$ and weight matrix $\mathbf{W}$ are chosen to satisfy*

$$\mathbb{E}[\mathbf{M}_k] = \mathbf{P}\mathbf{W}\mathbf{D}^{-2} \propto \mathbf{I}.$$

B. General Result

We now state a general convergence result for RK with averaging in Theorem 2. The proof is given in Appendix II. Theorem 2 in its general form is difficult to interpret, so we defer a detailed analysis to Section II-C in which the assumption of uniform weights simplifies the bound significantly.

Theorem 2. *Suppose $\mathbf{P}$ and $\mathbf{W}$ of Definition 2 are chosen such that $\mathbf{P}\mathbf{W}\mathbf{D}^{-2} = \frac{\alpha}{\|\mathbf{A}\|_F^2}\mathbf{I}$ for relaxation parameter $\alpha > 0$. Then the error at each iteration of Algorithm 1 satisfies*

$$\mathbb{E}\left[\|e^{k+1}\|^2\right] \leq \sigma_{\max}\left(\left(\mathbf{I} - \alpha\frac{\mathbf{A}^\top\mathbf{A}}{\|\mathbf{A}\|_F^2}\right)^2 - \frac{\alpha^2}{|\tau_k|}\left(\frac{\mathbf{A}^\top\mathbf{A}}{\|\mathbf{A}\|_F^2}\right)^2\right)\|e^k\|^2 + \frac{\alpha}{|\tau_k|}\frac{\|r^k\|^2_{\mathbf{W}}}{\|\mathbf{A}\|_F^2},$$

where $\|\cdot\|^2_{\mathbf{W}} = \langle \cdot, \mathbf{W}\cdot\rangle$ and $\|\mathbf{A}\|_F^2 = \sum_{i,j}\mathbf{A}_{ij}^2$.

Here, and for the remainder of the paper, we take the expectation $\mathbb{E}\left[\|e^{k+1}\|^2\right]$ conditioned on e_k.

As we shall see in Sections II-C and III, the relaxation parameter α and number of threads $|\tau_k|$ are closely tied to both the convergence horizon and convergence rate. The convergence horizon is proportional to $\frac{\alpha^2}{|\tau_k|}$, so smaller α and larger $|\tau_k|$ lead to a smaller convergence horizon. Increasing the value of α improves the convergence rate of the algorithm up to a critical point beyond which further increasing α leads to slower convergence

rates. Increasing the number of threads $|\tau_k|$ improves the convergence rate, asymptotically approaching an optimal rate as $|\tau_k| \to \infty$.

C. Uniform Weights

Suppose $\mathbf{W} = \alpha\mathbf{I}$, or equivalently that the weights are uniform. In this case, the update for each iteration becomes

$$x^{k+1} = x^k - \frac{\alpha}{|\tau_k|} \sum_{i \in \tau_k} \frac{\mathbf{A}_i x^k - b_i}{\|\mathbf{A}_i\|^2} \mathbf{A}_i^\top,$$

where $i \in \tau_k$ are independent samples from $\mathcal{D}$ with $p_i = \frac{\|\mathbf{A}_i\|^2}{\|\mathbf{A}\|_F^2}$. Under these conditions, the expected error bound of Theorem 2 can be simplified to remove the dependence on r^k. This simplification leads to the more interpretable error bound given in Corollary 3. In particular, increasing $|\tau_k|$ leads to both a faster convergence rate and smaller convergence horizon. If the relaxation parameter α is chosen to be one and a single row is selected at each iteration, we arrive at the RK method of [SV09]. Using uniform weights other than one results in the relaxed RK method [HN90a], [HN90b].

Corollary 3. *Suppose* $p_i = \frac{\|\mathbf{A}_i\|^2}{\|\mathbf{A}\|_F^2}$ *and* $\mathbf{W} = \alpha\mathbf{I}$. *Then the expected error at each iteration of Algorithm 1 satisfies*

$$\mathbb{E}\left[\|e^{k+1}\|^2\right] \leq \sigma_{\max}\left(\left(\mathbf{I} - \alpha\frac{\mathbf{A}^\top\mathbf{A}}{\|\mathbf{A}\|_F^2}\right)^2 + \frac{\alpha^2}{|\tau_k|}\left(\mathbf{I} - \frac{\mathbf{A}^\top\mathbf{A}}{\|\mathbf{A}\|_F^2}\right)\frac{\mathbf{A}^\top\mathbf{A}}{\|\mathbf{A}\|_F^2}\right)\|e^k\|^2 + \frac{\alpha^2\|r^\star\|^2}{|\tau_k|\|\mathbf{A}\|_F^2}.$$

The proof of Corollary 3 follows immediately from Theorem 2 and can be found in Section III-A.

1) Randomized Kaczmarz: If a single row is chosen at each iteration, with $\mathbf{W} = \mathbf{I}$ and $p_i = \frac{\|\mathbf{A}_i\|^2}{\|\mathbf{A}\|_F^2}$, then Algorithm 1 becomes the version of RK stated in [SV09]. In this case,

$$\|r^k\|_{\mathbf{W}}^2 = \|\mathbf{A}e^k\|^2 + \|r^\star\|^2. \tag{9}$$

Applying Theorem 2 leads to the following corollary, which recovers the error bound in Equation (5).

Corollary 4. *Suppose* $|\tau_k| = 1$, $\mathbf{W} = \mathbf{I}$ *and* $p_i = \frac{\|\mathbf{A}_i\|^2}{\|\mathbf{A}\|_F^2}$. *Then the expected error at each iteration of Algorithm 1 satisfies*

$$\begin{aligned}\mathbb{E}\left[\|e^{k+1}\|^2\right] &\leq \sigma_{\max}\left(\mathbf{I} - \frac{\mathbf{A}^\top\mathbf{A}}{\|\mathbf{A}\|_F^2}\right)\|e^k\|^2 + \frac{\|r^\star\|^2}{\|\mathbf{A}\|_F^2} \\ &= \left(1 - \frac{\sigma_{\min}^2(\mathbf{A})}{\|\mathbf{A}\|_F^2}\right)\|e^k\|^2 + \frac{\|r^\star\|^2}{\|\mathbf{A}\|_F^2}.\end{aligned}$$

A proof of Corollary 4 is included in Section III-B.

D. Consistent Systems

For consistent systems, Algorithm 1 converges to the solution $x^\star$ exponentially in expectation with the following guaranteed convergence rate.

Corollary 5. *Suppose* $\mathbf{P}$ *and* $\mathbf{W}$ *of Definition 2 are chosen such that* $\mathbf{P}\mathbf{W}\mathbf{D}^{-2} = \frac{\alpha}{\|\mathbf{A}\|_F^2}\mathbf{I}$ *for some constant* $\alpha > 0$. *Then the error at each iteration of Algorithm 1 satisfies*

$$\mathbb{E}\left[\|e^{k+1}\|^2\right] \leq \sigma_{\max}\left(\left(\mathbf{I} - \alpha\frac{\mathbf{A}^\top\mathbf{A}}{\|\mathbf{A}\|_F^2}\right)^2 + \frac{\mathbf{A}^\top}{\|\mathbf{A}\|_F}\left(\frac{\alpha}{|\tau_k|}\mathbf{W} - \frac{\alpha^2}{|\tau_k|}\frac{\mathbf{A}\mathbf{A}^\top}{\|\mathbf{A}\|_F^2}\right)\frac{\mathbf{A}}{\|\mathbf{A}\|_F}\right)\|e^k\|^2.$$

Corollary 5 can be derived from the proof of Theorem 2 with $r^\star = 0$.

For consistent systems and using uniform weights, Algorithm 1 becomes a subcase of the parallel sketch-and-project method of [RT17], which has a guaranteed convergence rate of

$$\mathbb{E}\left[\|e^{k+1}\|^2\right] \leq \left(1 - \frac{\sigma_{\min}^2(\mathbf{A})}{\frac{\|\mathbf{A}\|_F^2}{|\tau_k|} + \left(1 - \frac{1}{|\tau_k|}\right)\sigma_{\max}^2(\mathbf{A})}\right)\|e^k\|^2,$$

for relaxation parameter

$$\alpha^\star = \frac{1}{\frac{1}{|\tau_k|} + \left(1 - \frac{1}{|\tau_k|}\right)\frac{\sigma_{\max}^2(\mathbf{A})}{\|\mathbf{A}\|_F^2}}. \tag{10}$$

Here, the relaxation parameter $\alpha^\star$ is chosen to optimize the convergence rate guarantee. For a large number of threads, $|\tau_k|$, the optimal relaxation parameter becomes $\alpha^\star \approx \frac{\|\mathbf{A}\|_F^2}{\sigma_{\max}^2(\mathbf{A})}$.

III. Experiments

We present several experiments to demonstrate the convergence of Algorithm 1 under various conditions. In particular, we study the effects of the number of threads $|\tau_k|$, the relaxation parameter α, the weight matrix $\mathbf{W}$, and the probability matrix $\mathbf{P}$.

A. Procedure

For each experiment, we run 100 independent trials each starting with the initial iterate $x^0 = 0$ and average the squared error norms $\|e^k\|^2$ across the trials. We sample $\mathbf{A}$ from 100×10 standard Gaussian matrices and least-squares solution $x^\star$ from 10-dimensional standard Gaussian vectors, normalized so that $\|x^\star\| = 1$. To form inconsistent systems, we generate the least-squares residual $r^\star$ as a Gaussian vector orthogonal to the range of $\mathbf{A}$, also normalized so that $\|r^\star\| = 1$. Finally, b is computed as $r^\star + \mathbf{A}x^\star$.

B. The Effect of the Number of Threads

In Figure 1, we see the effects of the number of threads $|\tau_k|$ on the approximation error of Algorithm 1 for different choices of the weight matrices $\mathbf{W}$ and probability matrices $\mathbf{P}$. In Figures 1a and 1b, $\mathbf{W}$ and $\mathbf{P}$ satisfy Assumption 1, while in Figure 1c they do not.

As the number of threads $|\tau_k|$ increases by a factor of ten, we see a corresponding decrease in the magnitude of the convergence horizon by approximately the same factor for Figures 1a and 1b. This result corroborates what we expect based on Theorem 2 and Corollary 3. For Figure 1c, we do not see the same consistent decrease in the magnitude of the convergence horizon. As $|\tau_k|$ increases, for weight matrices $\mathbf{W}$ and probability matrices $\mathbf{P}$ that do not satisfy Assumption 1, the iterates x^k approach a weighted least-squares solution instead of the desired least-squares solution $x^\star$ (see Section II-A).

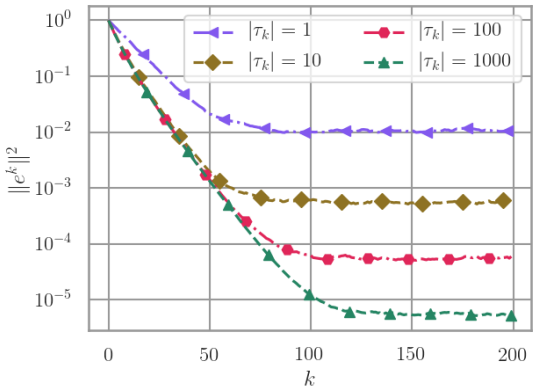
The rate of convergence in Figure 1 also improves as the number of threads $|\tau_k|$ increases. As $|\tau_k|$ increases, we see diminishing returns in the convergence rate. We expect this behavior based on the dependence on $\frac{1}{|\tau_k|}$ in Theorem 2 and Corollary 3.

C. The Effect of the Relaxation Parameter α

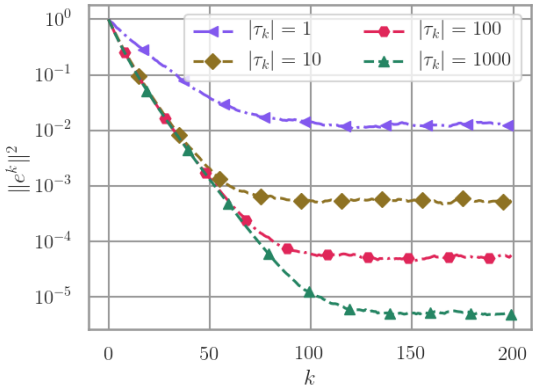
In Figure 2, we observe the effect on the convergence rate and convergence horizon as we vary the relaxation parameter α. From Theorem 2, we expect that the convergence horizon increases with α and indeed observe this experimentally. On the other hand, increasing α leads to improved convergence rates for α not too large. The squared norms of the errors behave similarly as α varies for both sets of weights and probabilities considered, each of which satisfy Assumption 1.

For larger values of the relaxation parameter α, the convergence rate for Algorithm 1 eventually decreases and the method can ultimately diverge. This behavior can be seen in Figure 3, which plots the number of iterations needed to converge to $\|e^k\|^2 < 10^{-10}$ for consistent Gaussian systems, various α, and various numbers of threads $|\tau_k|$. In terms of the number of iterations required, we find that there exists an optimal value for α, which increases with $|\tau_k|$. Comparing Figure 3a with Figure 3b, we observe a sharper minima in terms of α when using weights proportional to the squared row norms of $\mathbf{A}$ as opposed to uniform weights.

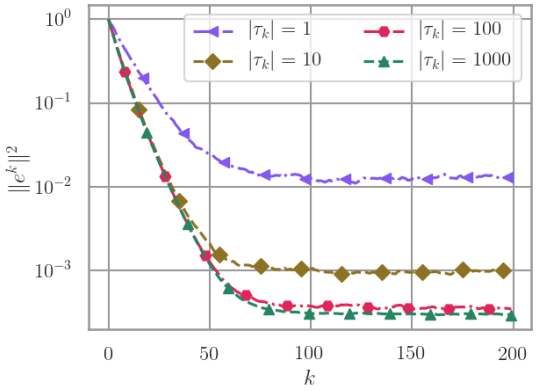
For uniform weights, an approximate optimal relaxation parameter $\alpha^\star$ for Algorithm 1 is calculated in [RT17]. This formula for $\alpha^\star$ is given in Equation (10). Table I provides values of $\alpha^\star$ for different numbers of threads $|\tau_k|$. These values are computed using the matrices from the experiment whose results are shown



(a) Uniform weights $w_i = 1$ and probabilities proportional to squared row norms $p_i = \frac{\|\mathbf{A}_i\|^2}{\|\mathbf{A}\|_F^2}$.

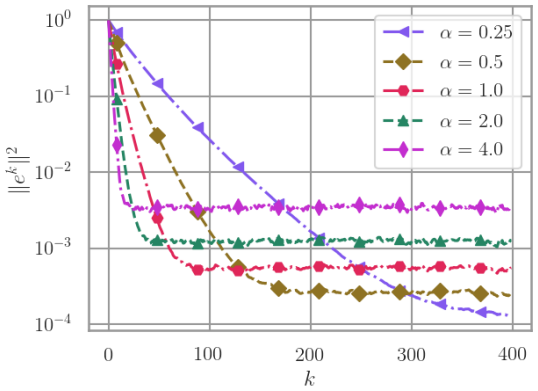


(b) Weights proportional to squared row norms $w_i = m\frac{\|\mathbf{A}_i\|^2}{\|\mathbf{A}\|_F^2}$ and uniform probabilities $p_i = \frac{1}{m}$.

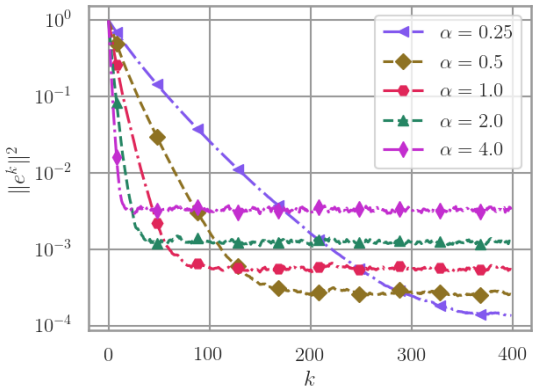


(c) Uniform weights $w_i = 1$ and uniform probabilities $p_i = \frac{1}{m}$.

Fig. 1: The effect of the number of threads on the average squared error norm vs iteration for Algorithm 1 applied to inconsistent systems. The weights w_i and probabilities p_i in (a) and (b) satisfy Assumption 1, while in (c) they do not.

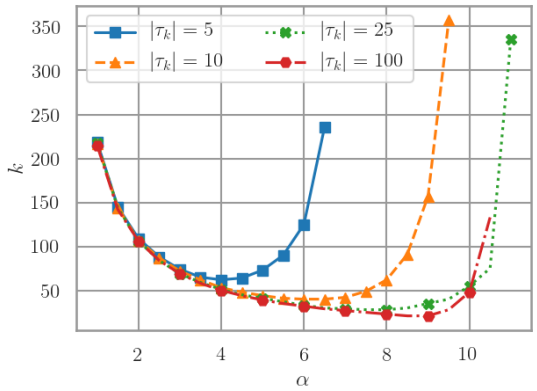


(a) Uniform weights $w_i = \alpha$, probabilities proportional to squared row norms $p_i = \frac{\|\mathbf{A}_i\|^2}{\|\mathbf{A}\|_F^2}$, and number of threads $|\tau_k| = 10$.

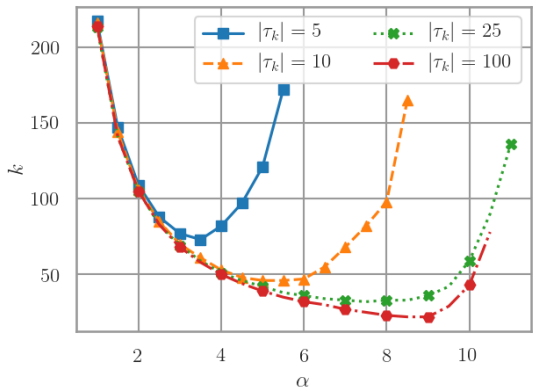


(b) Weights proportional to squared row norms $w_i = \alpha m \frac{\|\mathbf{A}_i\|^2}{\|\mathbf{A}\|_F^2}$, uniform probabilities $p_i = \frac{1}{m}$, and number of threads $|\tau_k| = 10$.

Fig. 2: The effect of the relaxation parameter α on the average squared error norm vs iteration for Algorithm 1 applied to inconsistent systems.



(a) Uniform weights $w_i = \alpha$ and probabilities proportional to squared row norms $p_i = \frac{\|\mathbf{A}_i\|^2}{\|\mathbf{A}\|_F^2}$.



(b) Weights proportional to squared row norms $w_i = \alpha m \frac{\|\mathbf{A}_i\|^2}{\|\mathbf{A}\|_F^2}$ and uniform probabilities $p_i = \frac{1}{m}$.

Fig. 3: Number of iterations needed for Algorithm 1 to achieve $\|e^k\|^2 < 10^{-10}$ on consistent systems for various choices of relaxation parameter α.

in Figure 3a. Comparing the $\alpha^\star$ of Table I with the α that minimize the curves in Figure 3a, we find that these values generally underestimate the optimal α that we observe experimentally.

TABLE I: Average optimal $\alpha^\star$ from Equation (10) for matrices $\mathbf{A}$ used in Figure 3a.

$\lvert\tau_k\rvert$	5	10	25	100
$\alpha^\star$	3.06	4.12	5.21	6.00

IV. Conclusion

We prove a general error bound for RK with averaging given in Algorithm 1 in terms of the number of threads $|\tau_k|$ and a relaxation parameter α. We find a natural coupling between the probability matrix $\mathbf{P}$ and the weight matrix $\mathbf{W}$ that leads to a reduced convergence horizon. We demonstrate that for uniform weights, i.e. $\mathbf{W} \propto \mathbf{I}$, the rate of convergence and convergence horizon for Algorithm 1 improve both in theory and practice as $|\tau_k|$ increases. For consistent systems with uniform weights, we recover existing convergence results. We also recover existing convergence results when a single thread is used, $|\tau_k| = 1$.

V. Acknowledgements

Molitor and Needell were partially supported by NSF CAREER grant #1348721 and NSF BIGDATA #1740325. Moorman was supported by NSF grant DGE-1829071. Tu was supported by DARPA under agreement number FA8750-18-2-0066.

REFERENCES

[CZT12] Yong Cai, Yang Zhao, and Yuchao Tang. Exponential convergence of a randomized Kaczmarz algorithm with relaxation. In Ford Lumban Gaol and Quang Vinh Nguyen, editors, *Proceedings of the 2011 2nd International Congress on Computer Applications and Computational Science*, pages 467–473, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.

[GG05] D. Gordon and R. Gordon. Component-averaged row projections: A robust, block-parallel scheme for sparse linear systems. *SIAM Journal on Scientific Computing*, 27(3):1092–1117, 2005.

[GR15] Robert Mansel Gower and Peter Richtárik. Randomized iterative methods for linear systems. *SIAM Journal on Matrix Analysis and Applications*, 36(4):1660–1690, 2015.

[HN90a] M Hanke and W Niethammer. On the use of small relaxation parameters in Kaczmarz method. *Zeitschrift fur Angewandte Mathematik und Mechanik*, 70(6):T575–T576, 1990.

[HN90b] Martin Hanke and Wilhelm Niethammer. On the acceleration of Kaczmarz's method for inconsistent linear systems. *Linear Algebra and its Applications*, 130:83–98, 1990.

[Kac37] M S Kaczmarz. Angenäherte auflösung von systemen linearer gleichungen. *Bulletin International de l'Académie Polonaise des Sciences et des Lettres. Classe des Sciences Mathématiques et Naturelles. Série A, Sciences Mathématiques*, 35:355–357, 1937.

[LWS14] Ji Liu, Stephen J Wright, and Sridhar Srikrishna. An asynchronous parallel randomized Kaczmarz algorithm. *arXiv:1401.4780*, 2014.

[Nee10] Deanna Needell. Randomized Kaczmarz solver for noisy linear systems. *BIT Numerical Mathematics*, 50(2):395–403, 2010.

[NRRW11] Feng Niu, Benjamin Recht, Christopher Ré, and Stephen J. Wright. HOGWILD!: A lock-free approach to parallelizing stochastic gradient descent. In *Neural Information Processing Systems*, 2011.

[NT12] Deanna Needell and Joel A. Tropp. Paved with good intentions: Analysis of a randomized block Kaczmarz method. *Linear Algebra and Its Applications*, 441(August):199–221, 2012.

[RT17] Peter Richtárik and Martin Takáč. Stochastic reformulations of linear systems: Algorithms and convergence theory. *arXiv e-prints*, page arXiv:1706.01108, June 2017.

[SV09] Thomas Strohmer and Roman Vershynin. A randomized Kaczmarz algorithm with exponential convergence. *Journal of Fourier Analysis and Applications*, 15(2):262–278, 2009.

[ZF13] Anastasios Zouzias and Nikolaos M Freris. Randomized extended Kaczmarz for solving least squares. *SIAM Journal on Matrix Analysis and Applications*, 34(2):773–793, 2013.

APPENDIX I
PROOF OF LEMMA 1

$$\mathbb{E}\left[\mathbf{M}_k\right] = \mathbb{E}\left[\sum_{i\in\tau_k} \frac{w_i}{|\tau_k|}\frac{\mathbf{I}_i^\top \mathbf{I}_i}{\|\mathbf{A}_i\|^2}\right] = \mathbb{E}\left[w_i \frac{\mathbf{I}_i^\top \mathbf{I}_i}{\|\mathbf{A}_i\|^2}\right]$$
$$= \sum_{i=1}^{m} p_i w_i \frac{\mathbf{I}_i^\top \mathbf{I}_i}{\|\mathbf{A}_i\|^2} = \mathbf{P}\mathbf{W}\mathbf{D}^{-2}.$$

$$\mathbb{E}\left[\mathbf{M}_k^\top \mathbf{A}\mathbf{A}^\top \mathbf{M}_k\right]$$
$$= |\tau_k|\,\mathbb{E}\left[\left(\frac{w_i}{|\tau_k|}\frac{\mathbf{I}_i^\top \mathbf{A}_i}{\|\mathbf{A}_i\|^2}\right)\left(\frac{w_i}{|\tau_k|}\frac{\mathbf{A}_i^\top \mathbf{I}_i}{\|\mathbf{A}_i\|^2}\right)\right]$$
$$+ (|\tau_k|^2 - |\tau_k|)\mathbb{E}\left[\frac{w_i}{|\tau_k|}\frac{\mathbf{I}_i^\top \mathbf{A}_i}{\|\mathbf{A}_i\|^2}\right]\mathbb{E}\left[\frac{w_j}{|\tau_k|}\frac{\mathbf{A}_j^\top \mathbf{I}_j}{\|\mathbf{A}_j\|^2}\right]$$
$$= \frac{1}{|\tau_k|}\mathbb{E}\left[w_i^2 \frac{\mathbf{I}_i^\top \mathbf{I}_i}{\|\mathbf{A}_i\|^2}\right]$$
$$+ \left(1 - \frac{1}{|\tau_k|}\right)\mathbf{P}\mathbf{W}\mathbf{D}^{-2}\mathbf{A}\mathbf{A}^\top\mathbf{P}\mathbf{W}\mathbf{D}^{-2}$$
$$= \frac{1}{|\tau_k|}\mathbf{P}\mathbf{W}^2\mathbf{D}^{-2}$$
$$+ \left(1 - \frac{1}{|\tau_k|}\right)\mathbf{P}\mathbf{W}\mathbf{D}^{-2}\mathbf{A}\mathbf{A}^\top\mathbf{P}\mathbf{W}\mathbf{D}^{-2}.$$

APPENDIX II
PROOF OF THEOREM 2

We prove Theorem 2 starting from from the error update in Equation (7). Expanding the squared error norm,

$$\|e^{k+1}\|^2 = \|(\mathbf{I} - \mathbf{A}^\top \mathbf{M}_k \mathbf{A})e^k + \mathbf{A}^\top \mathbf{M}_k r^\star\|^2$$
$$= \|(\mathbf{I} - \mathbf{A}^\top \mathbf{M}_k \mathbf{A})e^k\|^2$$
$$+ 2\langle(\mathbf{I} - \mathbf{A}^\top \mathbf{M}_k \mathbf{A})e^k, \mathbf{A}^\top \mathbf{M}_k r^\star\rangle$$
$$+ \|\mathbf{A}^\top \mathbf{M}_k r^\star\|^2.$$

Upon taking expections, the middle term simplifies since $\mathbf{A}^\top \mathbb{E}\left[\mathbf{M}_k\right] r^\star = 0$ by Assumption 1. Thus,

$$\mathbb{E}\left[\|e^{k+1}\|^2\right] = \mathbb{E}\left[\|(\mathbf{I} - \mathbf{A}^\top \mathbf{M}_k \mathbf{A})e^k\|^2\right]$$
$$- 2\mathbb{E}\left[\langle \mathbf{A}^\top \mathbf{M}_k \mathbf{A}e^k, \mathbf{A}^\top \mathbf{M}_k r^\star\rangle\right] \quad (11)$$
$$+ \mathbb{E}\left[\|\mathbf{A}^\top \mathbf{M}_k r^\star\|^2\right].$$

Making use of Lemma 1 to take the expectation of the first term in Equation (11),

$$\mathbb{E}\left[\|(\mathbf{I} - \mathbf{A}^\top \mathbf{M}_k \mathbf{A})e^k\|^2\right]$$
$$= \Big\langle e^k, (\mathbf{I} - 2\mathbf{A}^\top \mathbb{E}\left[\mathbf{M}_k\right]\mathbf{A}$$
$$+ \mathbf{A}^\top \mathbb{E}\left[\mathbf{M}_k^\top \mathbf{A}\mathbf{A}^\top \mathbf{M}_k\right]\mathbf{A})e^k\Big\rangle$$
$$= \left\langle e^k, \left(\mathbf{I} - 2\alpha\frac{\mathbf{A}^\top\mathbf{A}}{\|\mathbf{A}\|_F^2} + \frac{\alpha}{|\tau_k|}\frac{\mathbf{A}^\top\mathbf{W}\mathbf{A}}{\|\mathbf{A}\|_F^2}\right.\right.$$
$$\left.\left. + \alpha^2\left(1 - \frac{1}{|\tau_k|}\right)\left(\frac{\mathbf{A}^\top\mathbf{A}}{\|\mathbf{A}\|_F^2}\right)^2\right)e^k\right\rangle$$
$$= \left\langle e^k, \left(\left(\mathbf{I} - \alpha\frac{\mathbf{A}^\top\mathbf{A}}{\|\mathbf{A}\|_F^2}\right)^2\right.\right.$$
$$\left.\left. + \frac{\mathbf{A}^\top}{\|\mathbf{A}\|_F^2}\left(\frac{\alpha}{|\tau_k|}\mathbf{W} - \frac{\alpha^2}{|\tau_k|}\frac{\mathbf{A}\mathbf{A}^\top}{\|\mathbf{A}\|_F^2}\right)\mathbf{A}\right)e^k\right\rangle.$$

Since $\mathbf{A}^\top r^\star = 0$, for the second term,

$$\begin{aligned}2\mathbb{E}\left[\langle \mathbf{A}^\top \mathbf{M}_k \mathbf{A} e^k, \mathbf{A}^\top \mathbf{M}_k r^\star \rangle\right] &= 2\langle \mathbf{A}e^k, \mathbb{E}\left[\mathbf{M}_k^\top \mathbf{A}\mathbf{A}^\top \mathbf{M}_k\right] r^\star \rangle \\ &= 2\frac{\alpha}{|\tau_k|\,\|\mathbf{A}\|_F^2}\langle \mathbf{A}e^k, \mathbf{W}r^\star\rangle.\end{aligned}$$

Similarly, for the last term,

$$\mathbb{E}\left[\|\mathbf{A}^\top \mathbf{M}_k r^\star\|^2\right] = \frac{\alpha}{|\tau_k|}\frac{\|r^\star\|_{\mathbf{W}}^2}{\|\mathbf{A}\|_F^2}.$$

Combining these in Equation (11),

$$\begin{aligned}&\mathbb{E}\left[\|e^{k+1}\|^2\right] \\ &= \left\langle e^k, \left(\mathbf{I} - \alpha\frac{\mathbf{A}^\top\mathbf{A}}{\|\mathbf{A}\|_F^2}\right)^2 e^k\right\rangle \\ &\quad + \left\langle e^k, \frac{\mathbf{A}^\top}{\|\mathbf{A}\|_F^2}\left(\frac{\alpha}{|\tau_k|}\mathbf{W} - \frac{\alpha^2}{|\tau_k|}\frac{\mathbf{A}\mathbf{A}^\top}{\|\mathbf{A}\|_F^2}\right)\mathbf{A}e^k\right\rangle \\ &\quad - 2\frac{\alpha}{|\tau_k|}\frac{\langle \mathbf{A}e^k, \mathbf{W}r^\star\rangle}{\|\mathbf{A}\|_F^2} + \frac{\alpha}{|\tau_k|}\frac{\|r^\star\|_{\mathbf{W}}^2}{\|\mathbf{A}\|_F^2} \\ &= \left\langle e^k, \left(\left(\mathbf{I} - \alpha\frac{\mathbf{A}^\top\mathbf{A}}{\|\mathbf{A}\|_F^2}\right)^2 - \frac{\alpha^2}{|\tau_k|}\left(\frac{\mathbf{A}^\top\mathbf{A}}{\|\mathbf{A}\|_F^2}\right)^2\right) e^k\right\rangle \\ &\quad + \frac{\alpha}{|\tau_k|}\frac{\|r^k\|_{\mathbf{W}}^2}{\|\mathbf{A}\|_F^2} \\ &\leq \sigma_{\max}\left(\left(\mathbf{I} - \alpha\frac{\mathbf{A}^\top\mathbf{A}}{\|\mathbf{A}\|_F^2}\right)^2 - \frac{\alpha^2}{|\tau_k|}\left(\frac{\mathbf{A}^\top\mathbf{A}}{\|\mathbf{A}\|_F^2}\right)^2\right)\|e^k\|^2 \\ &\quad + \frac{\alpha}{|\tau_k|}\frac{\|r^k\|_{\mathbf{W}}^2}{\|\mathbf{A}\|_F^2}.\end{aligned}$$

APPENDIX III
COROLLARY PROOFS

We provide proofs for the corollaries of Section II, which follow from Theorem 2.

A. Proof of Corollary 3

Suppose $p_i = \frac{\|\mathbf{A}_i\|^2}{\|\mathbf{A}\|_F^2}$ and $\mathbf{W} = \alpha\mathbf{I}$. From the proof of Theorem 2,

$$\begin{aligned}&\mathbb{E}\left[\|e^{k+1}\|^2\right] \\ &= \left\langle e^k, \left(\left(\mathbf{I} - \alpha\frac{\mathbf{A}^\top\mathbf{A}}{\|\mathbf{A}\|_F^2}\right)^2 - \frac{\alpha^2}{|\tau_k|}\left(\frac{\mathbf{A}^\top\mathbf{A}}{\|\mathbf{A}\|_F^2}\right)^2\right) e^k\right\rangle \\ &\quad + \frac{\alpha}{|\tau_k|}\frac{\|r^k\|_{\mathbf{W}}^2}{\|\mathbf{A}\|_F^2}.\end{aligned}$$

In this case, since $\mathbf{A}^\top r^\star = 0$, $\langle \mathbf{A}e^k, r^\star\rangle = 0$ and

$$\begin{aligned}\|r^k\|_{\mathbf{W}}^2 &= \alpha\|\mathbf{A}e^k\|^2 + 2\alpha\langle \mathbf{A}e^k, r^\star\rangle + \alpha\|r^\star\|^2 \\ &= \alpha\langle e^k, \mathbf{A}^\top\mathbf{A}e^k\rangle + \alpha\|r^\star\|^2.\end{aligned}$$

Combining the inner products,

$$\begin{aligned}&\mathbb{E}\left[\|e^{k+1}\|^2\right] \\ &= \left\langle e^k, \left(\left(\mathbf{I} - \alpha\frac{\mathbf{A}^\top\mathbf{A}}{\|\mathbf{A}\|_F^2}\right)^2 \right.\right. \\ &\quad \left.\left. + \frac{\alpha^2}{|\tau_k|}\left(\mathbf{I} - \frac{\mathbf{A}^\top\mathbf{A}}{\|\mathbf{A}\|_F^2}\right)\frac{\mathbf{A}^\top\mathbf{A}}{\|\mathbf{A}\|_F^2}\right) e^k\right\rangle \\ &\quad + \frac{\alpha^2\|r^\star\|^2}{|\tau_k|\|\mathbf{A}\|_F^2} \\ &\leq \sigma_{\max}\left(\left(\mathbf{I} - \alpha\frac{\mathbf{A}^\top\mathbf{A}}{\|\mathbf{A}\|_F^2}\right)^2 \right. \\ &\quad \left. + \frac{\alpha^2}{|\tau_k|}\left(\mathbf{I} - \frac{\mathbf{A}^\top\mathbf{A}}{\|\mathbf{A}\|_F^2}\right)\frac{\mathbf{A}^\top\mathbf{A}}{\|\mathbf{A}\|_F^2}\right)\|e^k\|^2 + \frac{\alpha^2\|r^\star\|^2}{|\tau_k|\|\mathbf{A}\|_F^2}.\end{aligned}$$

B. Proof of Corollary 4

Suppose $|\tau_k| = 1$, $\mathbf{W} = \mathbf{I}$ and $p_i = \frac{\|\mathbf{A}_i\|^2}{\|\mathbf{A}\|_F^2}$.

$$\begin{aligned}\mathbb{E}\left[\|e^{k+1}\|^2\right] &\leq \sigma_{\max}\left(\mathbf{I} - \frac{\mathbf{A}^\top\mathbf{A}}{\|\mathbf{A}\|_F^2}\right)\|e^k\|^2 + \frac{\|r^\star\|^2}{\|\mathbf{A}\|_F^2} \\ &= \left(1 - \frac{\sigma_{\min}^2(\mathbf{A})}{\|\mathbf{A}\|_F^2}\right)\|e^k\|^2 + \frac{\|r^\star\|^2}{\|\mathbf{A}\|_F^2}.\end{aligned}$$

From the proof of Theorem 2,

$$\begin{aligned}&\mathbb{E}\left[\|e^{k+1}\|^2\right] \\ &= \left\langle e^k, \left(\left(\mathbf{I} - \frac{\mathbf{A}^\top\mathbf{A}}{\|\mathbf{A}\|_F^2}\right)^2 - \left(\frac{\mathbf{A}^\top\mathbf{A}}{\|\mathbf{A}\|_F^2}\right)^2\right) e^k\right\rangle \\ &\quad + \frac{\|r^k\|^2}{\|\mathbf{A}\|_F^2}.\end{aligned}$$

Decomposing r^k,

$$\begin{aligned}\|r^k\|^2 &= \|\mathbf{A}e^k\|^2 + \|r^\star\|^2 \\ &= \langle e^k, \mathbf{A}^\top\mathbf{A}e^k\rangle + \|r^\star\|^2.\end{aligned}$$

Combining the inner products,

$$\begin{aligned}&\mathbb{E}\left[\|e^{k+1}\|^2\right] \\ &= \left\langle e^k, \left(\left(\mathbf{I} - \frac{\mathbf{A}^\top\mathbf{A}}{\|\mathbf{A}\|_F^2}\right)^2 \right.\right. \\ &\quad \left.\left. - \left(\frac{\mathbf{A}^\top\mathbf{A}}{\|\mathbf{A}\|_F^2}\right)^2 + \frac{\mathbf{A}^\top\mathbf{A}}{\|\mathbf{A}\|_F^2}\right) e^k\right\rangle + \frac{\|r^\star\|^2}{\|\mathbf{A}\|_F^2} \\ &= \left\langle e^k, \left(\mathbf{I} - \frac{\mathbf{A}^\top\mathbf{A}}{\|\mathbf{A}\|_F^2}\right) e^k\right\rangle + \frac{\|r^\star\|^2}{\|\mathbf{A}\|_F^2} \\ &\leq \sigma_{\max}\left(\mathbf{I} - \frac{\mathbf{A}^\top\mathbf{A}}{\|\mathbf{A}\|_F^2}\right)\|e^k\|^2 + \frac{\|r^\star\|^2}{\|\mathbf{A}\|_F^2} \\ &= \left(1 - \frac{\sigma_{\min}^2(\mathbf{A})}{\|\mathbf{A}\|_F^2}\right)\|e^k\|^2 + \frac{\|r^\star\|^2}{\|\mathbf{A}\|_F^2}.\end{aligned}$$