Solving, tracking and stopping streaming linear inverse problems

OPEN ACCESS
IOP Publishing
Inverse Problems

Inverse Problems **40** (2024) 085003 (27pp) https://doi.org/10.1088/1361-6420/ad5583

Solving, tracking and stopping streaming linear inverse problems

Nathaniel Pritchard* **and Vivak Patel**

Department of Statistics, 1300 University Avenue, Madison, WI 53706, United States of America

E-mail: npritchard@wisc.edu and vivak.patel@wisc.edu

Received 30 January 2024; revised 17 May 2024
Accepted for publication 7 June 2024
Published 26 June 2024

Abstract

In large-scale applications including medical imaging, collocation differential equation solvers, and estimation with differential privacy, the underlying linear inverse problem can be reformulated as a streaming problem. In theory, the streaming problem can be effectively solved using memory-efficient, exponentially-converging streaming solvers. In special cases when the underlying linear inverse problem is finite-dimensional, streaming solvers can periodically evaluate the residual norm at a substantial computational cost. When the underlying system is infinite dimensional, streaming solver can only access noisy estimates of the residual. While such noisy estimates are computationally efficient, they are useful only when their accuracy is known. In this work, we rigorously develop a general family of computationally-practical residual estimators and their uncertainty sets for streaming solvers, and we demonstrate the accuracy of our methods on a number of large-scale linear problems. Thus, we further enable the practical use of streaming solvers for important classes of linear inverse problems.

Keywords: random sketching, consistent linear systems, randomized Kaczmarz, collocation problems, iterative methods, residual estimation

* Author to whom any correspondence should be addressed.

1. Introduction

In a myriad of scientific fields, such as medical imaging [1, 2], boundary element analysis [3], and differential privacy [4], large linear inverse problems can be reformulated as streaming problems that can be effectively solved using memory-efficient, exponentially-converging streaming solvers [5–7]. Specifically, a linear problem can be reformulated (and approximated) as determining at least one unknown, $z \in \mathbb{R}^n$, from a possibly infinite set, using elements of a random sequence $\{(\tilde{A}_k, \tilde{b}_k) : k \in \mathbb{N}\} \subset \mathbb{R}^{p\times n} \times \mathbb{R}^p$,[1] satisfying:

Assumption 1. For all $k \in \mathbb{N}$, there exists at least one z such that $\mathbb{P}(\tilde{A}_k z = \tilde{b}_k) = 1$; and

Assumption 2. $\{(\tilde{A}_k, \tilde{b}_k) : k \in \mathbb{N}\}$ are independent with distribution D.

Remark 1. Some examples of reformulating the linear inverse problem to a streaming model are detailed in appendix A.

When n and p are large enough that only one observation pair, $(\tilde{A}_k, \tilde{b}_k)$, can be stored in memory, the streaming problem can be solved using a streaming solver—a solver that makes use of $(\tilde{A}_k, \tilde{b}_k)$ to update its approximation of x^* and then discards the observation pair. A set of scalable streaming solvers that effectively confronts the complications of this scenario are known as Generalized Block Randomized Kaczmarz (GBRK) methods [7]. A GBRK method iteratively updates its current approximation by projecting it onto the hyperplanes specified by the observation pair, which allows it to make use of each observation pair and retain a low memory footprint. Importantly, GBRK methods also have (relatively) efficient, geometric convergence rates [5, 7–9][2].

However, a GBRK method's computational efficiency is undermined if it is stopped before, or stopped well-after, the desired solution accuracy is achieved. Thus, a GBRK method's iterates must be carefully tracked and stopped to realize its promise, yet neither tracking nor stopping is straightforward. In the special case where the linear inverse problem is a large linear system, a simple tracking and stopping approach could be to periodically compute the full residual [11], but this has two limitations: (1) an entire period of updates could be unecessarily computed before the method is stopped; and (2) in the setting of large linear systems, computing the entire residual is expensive (see numerical experiments in section 5). In the general case, a naïve tracking and stopping method is to use the residual at the current observation pair, but this has its own limitation: a point estimate is only useful if its accuracy is known. In other words, a naïve tracking and stopping method based on the residual at the current observation pair is only useful if we can estimate its uncertainty.

In this work, we rigorously develop a general family of computationally-efficient residual estimators and their uncertainty sets. Our family includes the simplest case of using just a single residual from an observation pair, and allows for using a moving window of previously computed, *dependent* residual estimates to reduce the size of the uncertainty set (see section 3). To analyze this dependent sequence, we develop a novel analytical technique for sub-Exponential distributions (see section 4), and show that the sub-Exponential model is appropriate for interesting streaming applications (see appendix A). From a practical perspective, we demonstrate the scalability of our methodology by tracking and stopping a streaming solver for a streaming collocation problem with one million fixed points (see section 5). In the context of non-streaming problems, we show that our methodology provides over a 500 times

[1] It is possible to let the copies of $(\tilde{A}_k, \tilde{b}_k)$ have different dimensions, but this would require introducing cumbersome notation at this point.

[2] If n is of moderate dimension, then more effective solvers can be generated at the cost of more storage [10].

improvement over tracking methods that periodically compute the full residual (see section 5). Thus, we further enable the practical use of streaming solvers for important classes of linear inverse problems.

Remark 2. While related, our current work substantially extends our previous methodology in [12] in a few key ways. First, in the previous work we required access to the full linear system at each iteration, in this work we no longer need to such accesses, instead we can access only a block of rows. Second, this work significantly extends the assumption about the distribution D. In the previous work D was a distribution of matrices that arise from applying Johnson-Lindenstrauss random matrices to the matrix A from the right. In this work, we extend this assumption to the case where D is the distribution of matrices satisfying a sub-Exponential property, which importantly allows for a D that generates random subsets of the matrix rows, a common choice of D for these methods [11, 13].

2. Notation

Throughout this paper we use, $\mathbb{E}[\cdot]$ to denote an expectation operation, and $\mathbb{P}(\cdot)$ to denote a probability measure. We let $\mathcal{F}_k$ denote the σ-algebra for $\{(\tilde{A}_1,\tilde{b}_1),\ldots,(\tilde{A}_k,\tilde{b}_k)\}$. Further, we use $\|\cdot\|_2$ to denote a two norm and $\|y\|_B = \sqrt{\langle y, By\rangle}$ to be a norm with respect to some symmetric positive definite matrix B. We discern between the quantities that we wish to estimate from the estimators by denoting the estimator with a $\hat{\cdot}$ above the quantity it is estimating.

3. Problem formulation & algorithm

Recall, we wish to solve (an approximation to) a linear inverse problem by determining an $x^* \in \mathbb{R}^n$ from a stream, $\{(\tilde{A}_k,\tilde{b}_k) : k \in \mathbb{N}\} \subset \mathbb{R}^{p\times n} \times \mathbb{R}^p$, satisfying assumptions 1 and 2; we use a GBRK method to determine x^* from this stream; and we need to efficiently track and stop this GBRK method. In this section, we describe our technique for accomplishing this task. To do this, we begin by discussing GBRK methods. In section 3.2, we present some additional, natural assumptions on the class of distributions for D to make tracking possible. Finally, in section 3.3, we introduce our methodology and discuss its salient properties.

3.1. GBRK

GBRK methods work by iteratively projecting a solution along hyperplanes formed by the row space of independent random observation pairs [5, 10, 14]. To be specific, given an initial point $x_0 \in \mathbb{R}^n$ and an inner product on $\mathbb{R}^n$ determined by a symmetric positive definite matrix $B \in \mathbb{R}^{n\times n}$ (usually just the identity) whose choice can correspond to other methods or whose value can be chosen to impact convergence rates [5], GBRK methods produce iterates recursively by

$$x_{k+1} = x_k - B^{-1}\tilde{A}_{k+1}^\top \left(\tilde{A}_{k+1}B^{-1}\tilde{A}_{k+1}^\top\right)^\dagger \left(\tilde{A}_{k+1}x_k - \tilde{b}_{k+1}\right). \tag{1}$$

GBRKs are known to have geometric convergence rates [5, 7, 8, 13, 15].

Remark 3. There are instances where in practice the inclusion of a relaxation parameter, $\phi \in (0,2]$, in the update could improve convergence [13]. In this case the update is

$$x_{k+1} = x_k - \phi B^{-1}\tilde{A}_{k+1}^\top \left(\tilde{A}_{k+1}B^{-1}\tilde{A}_{k+1}^\top\right)^\dagger \left(\tilde{A}_{k+1}x_k - \tilde{b}_{k+1}\right). \tag{2}$$

We suspect that the methods we present still work when $\phi \neq 1$; however, showing this theoretically requires the development of theoretical extensions to Meany's Inequality for scaled projections and is left as future work.

3.2. Sub-exponential distribution

In order to have an uncertainty set for the residual estimate, we need a model for the distribution of the residuals. To understand why, suppose $\tilde{A} = \tilde{b} \in \mathbb{R}$ and $\tilde{A}$ has Pareto distribution with shape parameter of 1. Then, this pair clearly satisfies assumption 1, and $\mathbb{E}[|\tilde{A}x - \tilde{b}|] = \infty$ for any $x \neq 1$. In this example, the (absolute) residual at all $x \neq 1$ has arbitrary variability, and, consequently, does not offer any reliable information about the system at any point besides 1 (even if we took an average over multiple independent copies of $(\tilde{A}, \tilde{b})$). Therefore, to avoid such pathological behavior, we will need to assume some control over the variability of the observations.

We use sub-Exponential distributions as a model for $(\tilde{A}_k, \tilde{b}_k)$, which is valid for interesting examples as demonstrated in appendix A. A sub-Exponential distribution for a random variable is defined in [16] as follows.

Definition 1. For a random variable Y, with $\mathbb{E}[Y] = \mu$, $Y - \mu$ follows a sub-Exponential, $\mathrm{SE}(\sigma, \omega)$, distribution with parameters σ and ω if for all $\delta \geqslant 0$

$$\mathbb{P}(|Y - \mu| > \delta) < 2e^{-\min\left\{\delta^2/\left(2\sigma^2\right), \delta/(2\omega)\right\}}. \tag{3}$$

Equivalently, a random variable Y is sub-Exponential, $\mathrm{SE}(\sigma, \omega)$, if

$$\mathbb{E}\left[e^{t(Y-\mu)}\right] \leqslant e^{\frac{t^2\sigma^2}{2}}, \tag{4}$$

when $|t| < 1/\omega$.

These sub-Exponential random variables continue to be sub-Exponential even when scaled by a constant, specifically we have the following lemma, which is a slight modification of the Bernstein inequality in [17].

Lemma 1. *If given a random variable Y with $\mathbb{E}[Y] = \mu$ such that $Y - \mu \sim SE(\sigma, \omega)$ then for constants c_1, c_2 where $c_2 \geqslant c_1 > 0$ it is the case that $c_1(Y - \mu) \sim SE(c_2\sigma, c_2\omega)$.*

The final key property is that all bounded random variables are sub-Exponential, which will be used to show that many example streaming problems are sub-Exponential. Specifically, if we adopt the convention for any positive constant, c, $c/0 = \infty$. Then, we have,

Lemma 2 ([16, example 2.4]). *If Y is a random variable with $\mathbb{E}[Y] = \mu \neq 0$ and $(Y - \mu)/\mu$ takes values in $[y_1, y_2]$, then*

$$\frac{Y - \mu}{\mu} \sim \mathrm{SE}\left(\frac{y_2 - y_1}{2}, 0\right). \tag{5}$$

With these facts established, we can now state our assumption for the variability of $(\tilde{A}_k, \tilde{b}_k)$.

Assumption 3. There exist $\sigma, \omega \geqslant 0$ such that, $\forall x \in \mathbb{R}^n, k \in \mathbb{N}$,

$$\begin{aligned} &\|\tilde{A}_k x - \tilde{b}_k\|_2^2 - \mathbb{E}\left[\|\tilde{A}_k x - \tilde{b}_k\|_2^2\right] \\ &\quad \sim \mathrm{SE}\left(\sigma\mathbb{E}\left[\|\tilde{A}_k x - \tilde{b}_k\|_2^2\right], \omega\mathbb{E}\left[\|\tilde{A}_k x - \tilde{b}_k\|_2^2\right]\right). \end{aligned} \tag{6}$$

As detailed in appendix A, assumptions 1–3 hold in a wide variety of situations including under the instances where the matrices are sketched according to the Johnson–Lindenstrauss transform, in the case when random subsets of a matrix are taken (Block Randomized Kaczmarz), and for a collocation problem evaluated at random points from a hyper-cube.

3.3. Streaming solver with tracking and stopping

Algorithm 1 describes our method for tracking and stopping streaming solvers applied to streaming reformulations of linear inverse problems whose observations satisfy assumptions 1–3. We highlight key aspects of the algorithm below.

Algorithm 1. Tracking and stopping for least squares.

Require: $B^{-1} \in \mathbb{R}^{n\times n}$ (usually identity matrix), $x_0 \in \mathbb{R}^n$.
Require: $\{(\tilde{A}_k, \tilde{b}_k)\} \subseteq \mathbb{R}^{p\times n} \times \mathbb{R}^p$ satisfying assumptions 1–3.
Require: Moving average window width $\lambda_1 \in \mathbb{N}$.
Require: $\alpha \in (0,1), \xi_I \in (0,1), \xi_{II} \in (0,1), \delta_I \in (0,1)$, $\delta_{II} > 1$, $\eta \geqslant 1$, $\upsilon > 0$.

$k \leftarrow 0$
while $k == 0$ **or** $\hat{\rho}_k \geqslant \upsilon$ **or**

$$\sqrt{\hat{\iota}_k} \geqslant \min\left\{ \frac{\lambda\eta(1-\delta_I)^2\upsilon^2}{2\log(1/\xi_I)\sigma^2\sqrt{\hat{\iota}_k}(1+\log(\lambda))}, \frac{\lambda\eta\upsilon(1-\delta_I)}{2\log(1/\xi_I)\omega}, \right.$$
$$\left. \frac{\lambda\eta(\delta_{II}-1)^2\upsilon^2}{2\log(1/\xi_{II})\sigma^2\sqrt{\hat{\iota}_k}(1+\log(\lambda))}, \frac{\lambda\eta\upsilon(\delta_{II}-1)}{2\log(1/\xi_{II})\omega} \right\}$$

do
  # Iteration $k+1$ #
  Receive $\tilde{A}_{k+1}$ and $\tilde{b}_{k+1}$
  $\tilde{r}_{k+1} \leftarrow \tilde{A}_{k+1}x_k - \tilde{b}_{k+1}$
  if $\lambda = 1$ **then**
    $\hat{\rho}_{k+1}, \hat{\iota}_{k+1} \leftarrow \|\tilde{r}_{k+1}\|_2^2, \|\tilde{r}_{k+1}\|_2^4$
    if $\|\tilde{r}_{+1}\|_2^2 > \|\tilde{r}_k\|_2^2$ **then**
      $\lambda \leftarrow 2$
    end if
  else
    $\hat{\rho}_{k+1} \leftarrow \sum_{i=k-\lambda+2}^{k+1} \frac{\|\tilde{r}_i\|_2^2}{\lambda}$
    $\hat{\iota}_{k+1} \leftarrow \sum_{i=k-\lambda+2}^{k+1} \frac{\|\tilde{r}_i\|_2^4}{\lambda}$
    if $\lambda < \lambda_1$ **then**
      $\lambda \leftarrow \lambda + 1$
    end if
  end if
  Update the estimated $(1-\alpha)$-interval by computing:

$$\hat{\rho}_{k+1} \pm \begin{cases} \sqrt{2\log(2/\alpha)\frac{\sigma^2\hat{\iota}_{k+1}(1+\log(\lambda))}{\eta\lambda}} & \text{if } \log(2/\alpha) \leqslant \frac{\lambda\sigma^2(1+\log(\lambda))}{2\omega^2} \\ \frac{2\log(2/\alpha)\omega\sqrt{\hat{\iota}_{k+1}}}{\sqrt{\eta}\lambda} & \text{if } \log(2/\alpha) > \frac{\lambda\sigma^2(1+\log(\lambda))}{2\omega^2}. \end{cases}$$

  $x_{k+1} \leftarrow x_k - B^{-1}\tilde{A}_{k+1}^\top(\tilde{A}_{k+1}B^{-1}\tilde{A}_{k+1}^\top)^\dagger\tilde{r}_{k+1}$
  $k \leftarrow k+1$
end while
return x_k **and** estimated $(1-\alpha)$-interval

3.3.1. Progress tracking point estimate. We track the progress of the solver through the point estimate,

$$\hat{\rho}_k^{\lambda} = \sum_{i=k-\lambda+1}^{k} \frac{\|\tilde{A}_i x_{i-1} - \tilde{b}_i\|_2^2}{\lambda} = \sum_{i=k-\lambda+1}^{k} \frac{\|\tilde{r}_i\|_2^2}{\lambda}. \tag{7}$$

Note, $\tilde{r}_i$ are already computed for the GBRK update, and so our method only requires computing their norm and storing λ such scalars. In other words, $\hat{\rho}_k^{\lambda}$ requires a marginal computational and memory footprint relative to the overall procedure. When $\lambda = 1$, we see that this point estimate reduces simply to the observed residual at the current iterate. However, we have several benefits for considering $\lambda > 1$ as explained by (i)–(iii).

(i) When $\lambda > 1$, we shrink the uncertainty sets as we might expect with independent observations, but with an extra logarithmic term on λ because of the statistical dependency between consecutive residual estimates. This logarithmic cost for dependency is a small cost to pay and allows the method to obtain a strong shrinking of the uncertainty (see theorem 4).

(ii) The expected value of $\hat{\rho}_k^{\lambda}$ (relative to the observation pairs at a given update),

$$\rho_k^{\lambda} = \sum_{i=k-\lambda+1}^{k} \frac{\mathbb{E}\left[\|\tilde{A}_i x_{i-1} - \tilde{b}_i\|_2^2 | \mathcal{F}_{i-1}\right]}{\lambda}, \tag{8}$$

seems to better track the absolute error of the iterate (see section 5.1 for more details).

(iii) There seems to be many cases where the use of $\lambda > 1$ allows for improved stopping performance with relation to the absolute error. Specifically, if one set the stopping criterion to be the value of the tracking estimate then at iteration k' when $\|x_{k'} - x^*\|_2^2 \leqslant \upsilon$, the use of $\lambda > 1$ can limit the number of times an early stopping error can occur in the sense of observing $\rho_k^{\lambda} < \rho_{k'}^{\lambda}$ for $k < k'$, indicating that we stop before the desired quality of absolute error is achieved (see section 5.1 for more details).

This latter behavior motivates the need for quantifying the uncertainty of $\hat{\rho}_k^{\lambda}$ so that we can account for the probabilities of false positives and false negatives.

3.3.2. Progress tracking uncertainty set estimate. In line 18, a $(1-\alpha)$ uncertainty set for $\hat{\rho}_k^{\lambda}$ is computed, where $\alpha \in (0,1)$ is given by the user (and reflect the user's risk tolerance). This uncertainty set is an estimate of the true uncertainty set for $\hat{\rho}_k^{\lambda}$ derived in corollary 3. The true uncertainty set satisfies either

$$\mathbb{P}\left(\rho_k^{\lambda} \in \hat{\rho}_k^{\lambda} \pm \sqrt{2\log(2/\alpha)\,\frac{\sigma^2 \hat{\iota}_k\,(1+\log(\lambda))}{\lambda}}\right) \geqslant 1-\alpha, \tag{9}$$

when $\log(2/\alpha) \leqslant \frac{\lambda\sigma^2(1+\log(\lambda))}{2\omega^2}$; or

$$\mathbb{P}\left(\rho_k^{\lambda} \in \hat{\rho}_k^{\lambda} \pm \frac{2\log(2/\alpha)\,\omega\sqrt{\hat{\iota}_k}}{\lambda}\right) \geqslant 1-\alpha, \tag{10}$$

otherwise.

Compared to (9) and (10), line 18 has an additional η term. This term is included to adjust for conservativeness of the theoretical intervals. Choices of η for specific distributions can be found in appendix A.1.

3.3.3. Stopping criterion. After, the completion of the uncertainty quantification stage, the solution is updated in lines 19–20 in accordance with (2). Then on line 2 the stopping criterion is checked. Note, in a deterministic setting stopping is simple: stop when the tracking value falls below some threshold υ. When randomness is introduced, this criterion can incur two different errors. The first can be viewed as stopping too late, and it occurs when the tracking parameter value, $\rho_k^\lambda \leqslant \delta_I \upsilon$, while $\hat{\rho}_k^\lambda > \upsilon$, where δ_I is a user defined parameter that permits the specification of where the gap between $\hat{\rho}_k^\lambda$ and ρ_k^λ is large enough to be considered problematic. By using the condition on $\hat{\iota}_k^\lambda$ (defined on line 13), we approximately control the probability of this error at ξ_I (see corollary 4 and section 4.3).

The second error type can be viewed as stopping too early, and it occurs when the tracking parameter value, $\rho_k^\lambda \geqslant \delta_{II} \upsilon$, while $\hat{\rho}_k^\lambda < \upsilon$ (cf, figure 1(b)), Where δ_{II} is a user defined parameter that permits the specification of where the gap between $\hat{\rho}_k^\lambda$ and ρ_k^λ is great enough to be considered problematic. By choosing the right stopping criterion we can then control the probability of this error at ξ_{II}.

As we will show in corollary 4, the criterion dependent on $\hat{\iota}_k^\lambda$ in line 2 of the algorithm accomplishes this task. Once this condition is satisfied, then the deviations between ρ_k^λ and $\hat{\rho}_k^\lambda$ are reasonably well controlled; thus, it is safe to stop when $\hat{\rho}_k^\lambda < \upsilon$ (see corollary 4 and section 4.3).

3.3.4. Moving average logistics. During the early phase of the algorithm, a GBRK usually experiences a rapid convergence to a region of the solution. During this phase, the GBRK's estimated residual reflects this rapid convergence and warrants a small moving average window. At later iterations, the convergence slows down and most of the variability in the residuals comes from randomness. During this phase, the moving average windows should be larger. Lines 6–10 and 14–15 reflect this behavior by starting the moving window at $\lambda = 1$, and increasing it to λ once the phase change is detected. The difference between these phases is determined to be the iteration k' where $\|\tilde{r}_{k'}\|_2^2 > \|\tilde{r}_{k'-1}\|_2^2$, as seen on line 8, which yields good empirical behavior.

4. Consistency of estimators and uncertainty sets

Core to the establishment of the theoretical validity of algorithm 1 is proving the consistency of the estimator of, and reliability of the uncertainty sets for, $\hat{\rho}_k^\lambda$. To accomplish these two tasks, it is first necessary to show that the general form of (2) converges in all moments, which is novel in comparison to previous analyses of GBRKs. From there, we can then combine this convergence result with Chernoff bounds to derive the distribution around $\hat{\rho}_k^\lambda$, which can be used to show $\hat{\rho}_k^\lambda$'s consistency for ρ_k^λ. With the consistency of $\hat{\rho}_k^\lambda$ established, we use the distribution of $\hat{\rho}_k^\lambda$ to derive its uncertainty set. Unfortunately, the uncertainty set will rely on an uncomputable quantity, so we close the section by showing that the uncomputable quantity can be estimated by $\hat{\iota}_k^\lambda$ with a reasonable relative error.

4.1. Convergence of the residuals' moments

Our first goal is to prove that all the moments of the residual will converge to zero. To achieve this goal, we will show that the moments of the absolute error converge to zero. To this end, we will transform the iteration update, (2), into a more amenable form. As we will see, this more amenable form shows that the updates are a sequence of orthogonal projections. Using these orthogonal projections and theorem 1, we will show that, at a random iteration, a sufficient, random geometric reduction in the error will occur. Our final step will be to control the random iteration and the random reduction in the error. Once these pieces are in place, we will be able to conclude that the moments of the absolute error and, hence, the residual, decay (geometrically) to zero.

4.1.1. Transformation of variables. To avoid unnecessary considerations about inner products, we will begin with a transformation of the variables by a symmetric square root of B. In other words, (2) becomes

$$B^{1/2}x_{k+1} = B^{1/2}x_k - B^{-1/2}\tilde{A}_{k+1}^\top \left(\tilde{A}_{k+1}B^{-1}\tilde{A}_{k+1}^\top\right)^\dagger \left(\tilde{A}_{k+1}x_k - \tilde{b}_{k+1}\right). \tag{11}$$

To simplify this relationship further, it will be useful to introduce several important spaces. Let

$$\mathcal{H} = \{x \in \mathbb{R}^n, \forall k \in \mathbb{N} : \mathbb{P}\left(\tilde{A}_k x = \tilde{b}_k\right) = 1\}, \tag{12}$$

$$\mathcal{N} = \{x \in \mathbb{R}^n, \forall k \in \mathbb{N} : \mathbb{P}\left(\tilde{A}_k B^{-1/2} x = 0\right) = 1\}, \text{ and} \tag{13}$$

$$\mathcal{R} = \mathcal{N}^\perp. \tag{14}$$

Under assumption 1, $\mathcal{H} \neq \emptyset$ and denotes the set of all solutions to the linear inverse problem. Moreover, $\mathcal{N}$ represents the null space of the linear problem, which can be equivalently written as $\mathcal{N} = \text{null}(B^{-1/2}\mathbb{E}[\tilde{A}_k^\top \tilde{A}_k]B^{-1/2})$. From this characterization, $\mathcal{R}$ represents the row space of the linear problem and, equivalently, $\mathcal{R} = \text{row}(B^{-1/2}\mathbb{E}[\tilde{A}_k^\top \tilde{A}_k]B^{-1/2})$.

Using these spaces and noting there could be infinitely many values that satisfy assumption 1, we let x^* be the orthogonal projection of x_0 onto the set $\mathcal{H}$, and let $\beta_k = B^{1/2}(x_k - x^*)$. Then, (11) simplifies to

$$\beta_{k+1} = \beta_k - B^{-1/2}\tilde{A}_{k+1}^\top \left(\tilde{A}_{k+1}B^{-1}\tilde{A}_{k+1}^\top\right)^\dagger \tilde{A}_{k+1}B^{-1/2}\beta_k. \tag{15}$$

From (15), we observe that $B^{-1/2}\tilde{A}_{k+1}^\top(\tilde{A}_{k+1}B^{-1}\tilde{A}_{k+1}^\top)^\dagger\tilde{A}_{k+1}B^{-1/2}$ is an orthogonal projection onto $\text{row}(\tilde{A}_{k+1}B^{-1/2})$. As a result, we have an observation and a useful simplification. First,

Lemma 3. $\{\beta_k : k+1 \in \mathbb{N}\} \subset \mathcal{R}$.

Proof. By construction, $\mathcal{N} \subset \text{null}(\tilde{A}_k B^{-1/2})$ for all k with probability one. Hence, $\text{row}(\tilde{A}_k B^{-1/2}) \perp \mathcal{N}$ with probability one. Letting $\mathcal{P}_\mathcal{N}$ denote the orthogonal projection operator onto $\mathcal{N}$, if $\mathcal{P}_\mathcal{N}\beta_k = 0$ then $\mathcal{P}_\mathcal{N}\beta_{k+1} = 0$ by (15). Noting that from our definition of x^*, $\beta_0 = B^{-1/2}(x_0 - x^*)$ can only be in $\mathcal{N}$ when $x_0 = x^*$. By construction, since $\beta_0 \notin \mathcal{N} \Rightarrow \beta_0 \in \mathcal{N}^\perp$; thus, $\mathcal{P}_\mathcal{N}\beta_0 = 0$. The result follows by induction. □

Second, if we let Q_{k+1} be a matrix with orthonormal columns that form a basis for $\text{row}(\tilde{A}_{k+1}B^{-1/2})$, then (15) becomes

$$\beta_{k+1} = \beta_k - Q_{k+1}Q_{k+1}^\top \beta_k. \tag{16}$$

4.1.2. Geometric reduction in error. Now, let $\tau_0 = 0$ and let τ_1 be the first iteration such that

$$\text{col}\,(Q_1) + \text{col}\,(Q_2) + \ldots + \text{col}\,(Q_{\tau_1}) = \mathcal{R}, \tag{17}$$

otherwise let τ_1 be infinite. When τ_1 is finite, we can use the following extension of Meany's lemma proposed in [10, theorem 4.1] about the convergence of sequences of orthogonal projections.

Theorem 1 ([10], theorem 4.1). *Let $q_1, \ldots, q_k$ be unit vectors in $\mathbb{R}^n$ for some $k \in \mathbb{N}$. Let $\mathcal{V} = span[q_1, \ldots, q_k]$. Let $\mathcal{W}$ denote all matrices W, where the columns of W are the vectors $\{w_1, \ldots, w_k\} \subset \{q_1, \ldots, q_k\}$ that are a maximal linearly independent subset. Then*

$$\sup_{y \in \mathcal{V}, \|y\|_2 = 1} \|Qy\|_2 \leqslant \sqrt{1 - \min_{W \in \mathcal{W}} det\,(W^\top W)}, \tag{18}$$

where $Q = (I - q_k q_k^\top) \ldots (I - q_1 q_1^\top)$.

By combining theorem 1 with our observation that (15) is a sequence of orthogonal projections, we obtain the following lemma.

Lemma 4. *Let $x_0 \in \mathbb{R}^n$. Let $\{x_k\}$ be generated according to (2) for $(\tilde{A}_k, \tilde{b}_k)$ from a distribution D satisfying assumption 1. Let x^* be the orthogonal projection of x_0 onto $\mathcal{H}$. On the event, $\{\tau_1 < \infty\}$, there exists a $\gamma_1 \in (0,1)$ that is a function of $\{Q_1, \ldots, Q_{\tau_1}\}$ such that*

$$\|x_{\tau_1} - x^*\|_B \leqslant \gamma_1 \|x_0 - x^*\|_B. \tag{19}$$

Proof. By our definition of β_k, we need only prove that $\exists \gamma_1 \in (0,1)$ such that $\|\beta_{\tau_1}\|_2 \leqslant \gamma_1 \|\beta_0\|_2$. To prove this, let $q_{k,1}, \ldots, q_{k,p}$ denote the columns of Q_k. Then, by (16),

$$\beta_{\tau_1} = \left[\prod_{k=1}^{\tau_1} \left(\prod_{j=1}^{p} (I - q_{k,j} q_{k,j}^\top) \right) \right] \beta_0. \tag{20}$$

Since $\beta_0 \in \mathcal{R}$ by lemma 3. theorem 1 implies that there exists a $\gamma_1 \in (0,1)$ that is a function of $\{q_{1,1}, q_{1,2}, \ldots, q_{\tau_1, p-1}, q_{\tau_1, p}\}$ such that $\|\beta_{\tau_1}\|_2 \leqslant \gamma_1 \|\beta_0\|_2$.

□

This result tells us that after seeing enough of the matrix at iteration τ_1, we will see a decrease in the norm of the absolute error by a random amount, γ_1. It is therefore reasonable to wonder how the algorithm behaves if we continue updating the solution beyond τ_1.

By continuing beyond iteration τ_1, we can iterate on this argument in the following manner. When $\{\tau_\ell < \infty\}$, define $\tau_{\ell+1}$ to be the first iteration after τ_ℓ such that

$$\text{col}\,(Q_{\tau_\ell + 1}) + \text{col}\,(Q_{\tau_\ell + 2}) + \ldots + \text{col}\,\left(Q_{\tau_{\ell+1}}\right) = \mathcal{R}, \tag{21}$$

otherwise let $\tau_{\ell+1}$ be infinite. Then, we have the following straightforward corollary.

Corollary 1. *Let $x_0 \in \mathbb{R}^n$. Let $\{x_k\}$ be generated according to (2) for $(\tilde{A}_k, \tilde{b}_k)$ from a distribution D satisfying assumption 1. Let x^* be the orthogonal projection of x_0 onto $\mathcal{H}$. On the event, $\cap_{\ell=1}^{L} \{\tau_\ell < \infty\}$, there exists $\gamma_\ell \in (0,1)$ that is a function of $\{Q_{\tau_{\ell-1}+1}, \ldots, Q_{\tau_\ell}\}$ for $\ell = 1, \ldots, L$, such that*

$$\|x_{\tau_L} - x^*\|_B \leqslant \left(\prod_{\ell=1}^{L} \gamma_\ell \right) \|x_0 - x^*\|_B. \tag{22}$$

4.1.3. Control of the random rate and random iteration. Of course, corollary 1 does not imply that the absolute error converges to zero. In fact, corollary 1 has two points of failure. First, it may happen that $\gamma_\ell \to 1$ as $\ell \to \infty$ with some nonzero probability; that is, we have no control over the random rate of convergence. This issue is addressed by the following result, which relies on the independence of $\{(\tilde{A}_k, \tilde{b}_k)\}$.

To obtain this results we rely on the following key theorem from [18, theorem 4.1.3].

Theorem 2 ([18], theorem 4.1.3). *Let $Z_1, Z_2, \ldots$ be i.i.d. random variables, $\mathcal{Z}_n = \sigma(Z_1, \ldots, Z_k)$ and τ be a stopping time with $\mathbb{P}(\tau < \infty) > 0$. Conditioned on $\{\tau < \infty\}$, $\{Z_{\tau+k}, k > 1\}$ is independent of $\mathcal{Z}_n$ and has the same distribution as $Z_1, \ldots, Z_k$.*

We now can use theorem 2 to show that $\gamma_\ell \to 1$ does not occur.

Lemma 5. *Let $x_0 \in \mathbb{R}^n$. Let $\{x_k\}$ be generated according to (2) for $\{(\tilde{A}_k, \tilde{b}_k)\}$ from a distribution D satisfying assumptions 1, 2. Let x^* be the orthogonal projection of x_0 onto $\mathcal{H}$. Then, whenever they exist, $\{\tau_\ell - \tau_{\ell-1} : \ell \in \mathbb{N}\}$ are independent and identically distributed; and $\{\gamma_\ell : \ell \in \mathbb{N}\}$ are independent and identically distributed.*

Proof. When τ_ℓ is finite, by theorem 2, $\{Q_{\tau_\ell+1}, \ldots, Q_{\tau_\ell+k}\}$ given τ_ℓ are independent of $\{Q_1, \ldots, Q_{\tau_\ell}\}$ and are identically distributed to $\{Q_1, \ldots, Q_k\}$ for all k. Therefore, $\tau_\ell - \tau_{\ell-1}$ are independent and identically distributed, as are γ_ℓ. □

The second point of failure in corollary 1, as alluded to in lemma 5, is the existence of $\{\tau_\ell\}$. Specifically, on $\{\tau_\ell = \infty\}$, corollary 1 will no longer supply a rate of improvement in the absolute error. Therefore, we must show that $\{\tau_\ell < \infty\}$ occurs with probability one, which is the content of the next result.

Lemma 6. *Let $x_0 \in \mathbb{R}^n$. Let $\{x_k\}$ be generated according to (2) for $\{(\tilde{A}_k, \tilde{b}_k)\}$ from streaming distribution D, satisfying assumptions 1–3. Let x^* be the orthogonal projection of x_0 onto $\mathcal{H}$.*

Then $\mathbb{P}(\tau_\ell < \infty) = 1$ for every $\ell \in \mathbb{N}$. Moreover, $\exists \pi \in (0, 1]$ such that, for all $\ell \in \mathbb{N}$ and $k \geqslant \operatorname{rank}(\mathcal{R})$,

$$\mathbb{P}(\tau_\ell - \tau_{\ell-1} = k) \leqslant \binom{k-1}{\operatorname{rank}(\mathcal{R}) - 1} (1-\pi)^{k-\operatorname{rank}(\mathcal{R})} \pi^{\operatorname{rank}(\mathcal{R})}. \tag{23}$$

Proof. Given that $\{Q_k : k \in \mathbb{N}\}$ are independent and identically distributed, we will show that the probability that $\operatorname{col}(Q_1) + \ldots + \operatorname{col}(Q_{k+1})$ grows in dimension relative to $\operatorname{col}(Q_1) + \ldots + \operatorname{col}(Q_k)$, when $\dim(\operatorname{col}(Q_1) + \ldots + \operatorname{col}(Q_k)) < \operatorname{rank}(\mathcal{R})$ is some $\pi \in (0, 1]$. As a result, the probability that the dimension increases $\operatorname{rank}(\mathcal{R})$ times in k iterations (with $k \geqslant \operatorname{rank}(\mathcal{R})$) is dominated by a negative binomial distribution. In other words, for $k \geqslant \operatorname{rank}(\mathcal{R})$,

$$\mathbb{P}(\tau_1 = k) \leqslant \binom{k-1}{\operatorname{rank}(\mathcal{R}) - 1} (1-\pi)^{k-\operatorname{rank}(\mathcal{R})} \pi^{\operatorname{rank}(\mathcal{R})}. \tag{24}$$

This implies τ_1 is finite with probability one, and the result for $\tau_\ell - \tau_{\ell-1}$ follows by lemma 5.

Thus, it only remains to show that the probability that the dimension grows is bounded from below by $\pi' \in (0, 1]$. To do so, we need only show that $\exists \pi' > 0$ such that for any $z \in \mathcal{R}$, $\mathbb{P}(\|Q_1^\top z\|_2^2 > 0) \geqslant \pi'$. Since π' represents a lower bound on the probability that any vector in $\mathcal{R}$ falls in the column space of a Q_i, it implies if we have some vector $z' \in \mathcal{R}$ but $z' \notin \operatorname{col}(Q_1) + \ldots + \operatorname{col}(Q_k)$ then we know $z' \in \operatorname{col}(Q_{k+1})$ with probability at least π', which implies that $\dim(\operatorname{col}(Q_1) + \ldots + \operatorname{col}(Q_{k+1}))$ will increase with at least the same probability. If we now recall that $\operatorname{col}(Q_1) = \operatorname{row}(\tilde{A}_1 B^{-1/2}) \subset \mathcal{R}$, then we can note that $\mathbb{P}(\|Q_1^\top z\|_2^2 > 0) =$

$\mathbb{P}(\|\tilde{A}_k B^{-1/2} z\|_2^2 > 0)$, and thus if we observe enough stream blocks, the sum of the spaces will span all of $\mathcal{R}$. We prove the existence of π' by noting from definition 1 that for any $\delta \in (0,1)$,

$$\mathbb{P}\left(\|\tilde{A}_k B^{-1/2} z\|_2^2 > 0\right) \geqslant \mathbb{P}\left(\|\tilde{A}_k B^{-1/2} z\|_2^2 \geqslant (1-\delta)\,\mathbb{E}\left[\|\tilde{A}_k B^{-1/2} z\|_2^2\right]\right) \tag{25}$$

$$\geqslant 1 - \mathbb{P}\left(\frac{\|\tilde{A}_k B^{-1/2} z\|_2^2 - \mathbb{E}\left[\|\tilde{A}_k B^{-1/2} z\|_2^2\right]}{\mathbb{E}\left[\|\tilde{A}_k B^{-1/2} z\|_2^2\right]} < -\delta\right) \tag{26}$$

$$\geqslant 1 - \exp\left(-\min\{\delta^2/\left(2\sigma^2\right), \delta/\left(2\omega\right)\}\right), \tag{27}$$

where (26) comes from assumption 3.

Note, for any $\delta \in (0,1)$, the last term is strictly larger than 0. Hence, we can fix a $\delta \in (0,1)$ and set the resulting value to π'. □

4.1.4. Convergence of the moments. We now put these pieces together to conclude as follows.

Theorem 3. *Let $x_0 \in \mathbb{R}^n$. Let $\{x_k\}$ be generated according to (2) for $\{(\tilde{A}_k, \tilde{b}_k)\}$ from distribution D and satisfying assumptions 1–3. Let x^* be the orthogonal projection of x_0 onto $\mathcal{H}$. Then, for any $d \in \mathbb{N}$, $\mathbb{E}[\|\tilde{A}_{k+1} x_k - \tilde{b}_{k+1}\|_2^d] \to 0$ and $\mathbb{E}[\|x_k - x^*\|_B^d] \to 0$ as $k \to \infty$. Additionally, for any $\ell \in \mathbb{N}$, we have*

$$\mathbb{E}\left[\|x_{\tau_\ell} - x^*\|_B^d\right] \leqslant \mathbb{E}\left[\gamma_1^d\right]^\ell \|x_0 - x^*\|_B^d. \tag{28}$$

Proof. It is enough to show that $\mathbb{E}[\|x_k - x^*\|_B^d] \to 0$ as $k \to \infty$. By (16), the absolute error is a non-increasing sequence. Thus, we need only show that a subsequence converges to zero. By corollary 1, lemmas 5 and 6,

$$\mathbb{E}\left[\|x_{\tau_\ell} - x^*\|_B^d\right] \leqslant \mathbb{E}\left[\gamma_1^d\right]^\ell \|x_0 - x^*\|_B^d, \tag{29}$$

for all $\ell \in \mathbb{N}$, where $\mathbb{E}[\gamma_1^d] < 1$. Therefore, as $\ell \to \infty$, the conclusion follows. □

4.2. Distribution and consistency of estimators

Using the convergence of the moments established in theorem 3, we now determine the distribution of $\hat{\rho}_k^\lambda$. Determining the distribution of $\hat{\rho}_k^\lambda$ would be easy if the terms in the sum of the norm squared of the residuals, $\|\tilde{r}_i\|_2^2 = \|\tilde{A}_i x_{i-1} - \tilde{b}_i\|_2^2$, were independent, as assumption 3 would imply that $\hat{\rho}_k^\lambda$ is sub-Exponential. Unfortunately, the terms composing $\hat{\rho}_k^\lambda$ are dependent. To handle this dependence, we innovate the following proof and conclude, $\hat{\rho}_k^\lambda$ is sub-Exponentially distributed with a variance that is only a logarithmic term worse than the independent case.

Since we are concerned with the consistency of the estimators we need to first understand the distribution conditioning on $\mathcal{F}_{k-\lambda}$, the σ-algebra generated by $\{(\tilde{A}_j, \tilde{b}_j) : j = 1, \ldots k-\lambda\}$, of $\|\tilde{r}_i\|_2^2 - \mathbb{E}[\|\tilde{r}_i\|_2^2 | \mathcal{F}_{i-1}] | \mathcal{F}_{k-\lambda}$ for any $i > k-\lambda$.

Lemma 7. *Suppose the setting of theorem 3 holds. Then, for any $i > k-\lambda$,*

$$\|\tilde{r}_i\|_2^2 - \mathbb{E}\left[\|\tilde{r}_i\|_2^2 | \mathcal{F}_{i-1}\right] \Big| \mathcal{F}_{k-\lambda} \sim \boldsymbol{SE}\left(\sigma M_{k-\lambda}, \omega M_{k-\lambda}\right), \tag{30}$$

where $M_{k-\lambda} = \mathbb{E}[\|\tilde{A}_{k-\lambda} B^{-1/2}\|_2^2] \|x_{k-\lambda-1} - x^*\|_B^2$.

Proof. By (16),

$$\mathbb{E}\left[\|\tilde{r}_i\|_2^2|\mathcal{F}_{i-1}\right] \leqslant \mathbb{E}\left[\|\tilde{A}_i B^{-1/2}\|_2^2\right] \|x_{i-1}-x^*\|_B^2 \leqslant M_{k-\lambda} \tag{31}$$

Using (31) and assumption 3, for any $\delta > 0$,

$$\begin{aligned}
&\mathbb{P}\left(\left|\|\tilde{r}_i\|_2^2 - \mathbb{E}\left[\|\tilde{r}_i\|_2^2|\mathcal{F}_{i-1}\right]\right| > \delta \middle| \mathcal{F}_{k-\lambda}\right) \\
&< 2\exp\left(-\min\left\{\frac{\delta^2}{2\sigma^2\mathbb{E}\left[\|\tilde{r}_i\|_2^2|\mathcal{F}_{i-1}\right]^2}, \frac{\delta}{2\omega\mathbb{E}\left[\|\tilde{r}_i\|_2^2|\mathcal{F}_{i-1}\right]}\right\}\right) &(32)\\
&< 2\exp\left(-\min\left\{\frac{\delta^2}{2\sigma^2 M_{k-\lambda}^2}, \frac{\delta}{2\omega M_{k-\lambda}}\right\}\right). &(33)
\end{aligned}$$

Taking expectations of the probabilities with respect to $\mathcal{F}_{k-\lambda}$ in the previous display equation and noting that $M_{k-\lambda}$ is measurable with respect to $\mathcal{F}_{k-\lambda}$, we conclude. □

Using this lemma, we now characterize the distribution of $\hat{\rho}_k^\lambda - \rho_k^\lambda$ to be sub-Exponential.

Theorem 4. *Suppose the setting of theorem 3 holds. Then,*

$$\hat{\rho}_k^\lambda - \rho_k^\lambda \Big| \mathcal{F}_{k-\lambda} \sim \mathbf{SE}\left(\sigma M_{k-\lambda}\sqrt{\frac{(1+\log(\lambda))}{\lambda}}, \frac{\omega M_{k-\lambda}}{\lambda}\right). \tag{34}$$

Proof. Using lemma 7 and induction, we will prove, for $|t| \leqslant \lambda/(\omega M_{k-\lambda})$,

$$\begin{aligned}
&\mathbb{E}\left[e^{t\left(\hat{\rho}_k^\lambda - \rho_k^\lambda\right)} \middle| \mathcal{F}_{k-\lambda}\right] &(35)\\
&= \mathbb{E}\left[\prod_{i=k-\lambda+1}^{k} \exp\left\{\frac{t}{\lambda}\left(\|\tilde{r}_i\|_2^2 - \mathbb{E}\left[\|\tilde{r}_i\|_2^2|\mathcal{F}_{i-1}\right]\right)\right\} \middle| \mathcal{F}_{k-\lambda}\right] &(36)\\
&\leqslant \exp\left(\frac{t^2\sigma^2 M_{k-\lambda}^2}{2\lambda}\sum_{j=1}^{\lambda}\frac{1}{j}\right). &(37)
\end{aligned}$$

We can then use a logarithm to bound the summation, which yields the conclusion.

The base case of $\lambda = 1$ follows trivially from lemma 7. Now assume that the result holds up to $\lambda - 1$. Then,

$$\begin{aligned}
&\mathbb{E}\left[\prod_{i=k-\lambda+1}^{k} \exp\left\{\frac{t}{\lambda}\left(\|\tilde{r}_i\|_2^2 - \mathbb{E}\left[\|\tilde{r}_i\|_2^2|\mathcal{F}_{i-1}\right]\right)\right\} \middle| \mathcal{F}_{k-\lambda}\right] &(38)\\
&= \mathbb{E}\left[\mathbb{E}\left[\prod_{i=k-\lambda+1}^{k} \exp\left\{\frac{t}{\lambda}\left(\|\tilde{r}_i\|_2^2 - \mathbb{E}\left[\|\tilde{r}_i\|_2^2|\mathcal{F}_{i-1}\right]\right)\right\} \middle| \mathcal{F}_{k-1}\right] \middle| \mathcal{F}_{k-\lambda}\right] &(39)\\
&= \mathbb{E}\left[\mathbb{E}\left[\exp\left\{\frac{t}{\lambda}\left(\|\tilde{r}_k\|_2^2 - \mathbb{E}\left[\|\tilde{r}_k\|_2^2|\mathcal{F}_{k-1}\right]\right)\right\} \middle| \mathcal{F}_{k-1}\right]\right.\\
&\left.\times \prod_{i=k-\lambda+1}^{k-1} \exp\left\{\frac{t}{\lambda}\left(\|\tilde{r}_i\|_2^2 - \mathbb{E}\left[\|\tilde{r}_i\|_2^2|\mathcal{F}_{i-1}\right]\right)\right\} \middle| \mathcal{F}_{k-\lambda}\right] &(40)
\end{aligned}$$

$$\leqslant \mathbb{E}\left[\exp\left\{\frac{t^2\sigma^2 M_{k-\lambda}^2}{2\lambda^2}\right\}\right. \tag{41}$$
$$\left.\times \prod_{i=k-\lambda+1}^{k-1} \exp\left\{\frac{t}{\lambda}\left(\|\tilde{r}_i\|_2^2 - \mathbb{E}\left[\|\tilde{r}_i\|_2^2|\mathcal{F}_{i-1}\right]\right)\right\}\middle|\mathcal{F}_{k-\lambda}\right],$$

where we have made use of lemma 7 in the final line. Now, applying Hölder's inequality and the induction hypothesis,

$$\mathbb{E}\left[\exp\left\{\frac{t^2\sigma^2 M_{k-\lambda}^2}{2\lambda^2}\right\}\right. \tag{42}$$
$$\left.\times \prod_{i=k-\lambda+1}^{k-1} \exp\left\{\frac{t}{\lambda}\left(\|\tilde{r}_i\|_2^2 - \mathbb{E}\left[\|\tilde{r}_i\|_2^2|\mathcal{F}_{i-1}\right]\right)\right\}\middle|\mathcal{F}_{k-\lambda}\right]$$
$$\leqslant \mathbb{E}\left[\exp\left\{\frac{t^2\sigma^2 M_{k-\lambda}^2}{2\lambda}\right\}\middle|\mathcal{F}_{k-\lambda}\right]^{\frac{1}{\lambda}} \tag{43}$$
$$\times \mathbb{E}\left[\prod_{i=k-\lambda+1}^{k-1} \exp\left\{\frac{t}{\lambda-1}\left(\|\tilde{r}_i\|_2^2 - \mathbb{E}\left[\|\tilde{r}_i\|_2^2|\mathcal{F}_{i-1}\right]\right)\right\}\middle|\mathcal{F}_{k-\lambda}\right]^{\frac{\lambda-1}{\lambda}}$$
$$\leqslant \mathbb{E}\left[\exp\left\{\frac{t^2\sigma^2 M_{k-\lambda}^2}{2\lambda}\right\}\middle|\mathcal{F}_{k-\lambda}\right]^{\frac{1}{\lambda}} \exp\left\{\frac{t^2\sigma^2 M_{k-\lambda}^2}{2(\lambda-1)}\sum_{j=1}^{\lambda-1}\frac{1}{j}\right\}^{\frac{\lambda-1}{\lambda}}. \tag{44}$$

Since $M_{k-\lambda}$ is measurable with respect to $\mathcal{F}_{k-\lambda}$, we have shown our desired inequality. □

From this distribution we can easily conclude the following corollary about the consistency of $\hat{\rho}_k^\lambda$.

Corollary 2. *Suppose the setting of theorem 3 holds. Then, for any* $\lambda \in \mathbb{N}$ *and* $\forall \delta > 0$, $\lim_{k\to\infty} \mathbb{P}(|\hat{\rho}_k^\lambda - \rho_k^\lambda| > \delta) = 0$. *That is,* $\hat{\rho}_k^\lambda$ *is a consistent estimator of* ρ_k^λ.

Proof. We begin by noting that

$$\lim_{k\to\infty} \mathbb{P}\left(|\hat{\rho}_k^\lambda - \rho_k^\lambda| > \delta\right) = \lim_{k\to\infty} \mathbb{E}\left[\mathbb{P}\left(|\hat{\rho}_k^\lambda - \rho_k^\lambda| > \delta|\mathcal{F}_{k-\lambda}\right)\right] \tag{45}$$
$$\leqslant \lim_{k\to\infty} \mathbb{E}\left[2e^{-\min\left\{\frac{\lambda\delta^2}{2\sigma^2 M_{k-\lambda}^2(1+\log(\lambda))}, \frac{\delta}{2\omega M_{k-\lambda}}\right\}}\right] \tag{46}$$
$$= \mathbb{E}\left[\lim_{k\to\infty} 2e^{-\min\left\{\frac{\lambda\delta^2}{2\sigma^2 M_{k-\lambda}^2(1+\log(\lambda))}, \frac{\delta}{2\omega M_{k-\lambda}}\right\}}\right] \tag{47}$$
$$= 0, \tag{48}$$

where (45) comes from the tower property for conditional expectation; (46) comes from theorem 4 and definition 1, (47) comes from $M_{k-\lambda}$ being a monotonically decreasing sequence implying that the term inside the expectation is dominated by $2\exp(-\min\{\frac{\lambda\delta^2}{2\sigma^2 M_0^2(1+\log(\lambda))}, \frac{\delta}{2\omega M_0}\})$, an integrable function. This domination by an integrable function allows us to use the dominated convergence theorem to switch the limit and the integral. Equation (48) then arises from theorem 3 and the definition of $M_{k-\lambda}$. □

With the validity of our estimators established and the derivation of the distribution of $\hat{\rho}_k^\lambda$, we are now able to derive the uncertainty set and stopping criterion in the following two corollaries, whose proofs are straightforward.

Corollary 3. *An uncertainty set of level* $1-\alpha$ *for* ρ_k^λ *is*

$$\hat{\rho}_k^\lambda \pm \epsilon, \tag{49}$$

where

$$\epsilon = \begin{cases} \sqrt{2\log(2/\alpha)\frac{\sigma^2 M_{k-\lambda}^2(1+\log(\lambda))}{\lambda}} & \text{if } \log(2/\alpha) \leqslant \frac{\lambda\sigma^2(1+\log(\lambda))}{2\omega^2} \\ \frac{2\log(2/\alpha)\omega M_{k-\lambda}}{\lambda} & \text{if } \log(2/\alpha) > \frac{\lambda\sigma^2(1+\log(\lambda))}{2\omega^2}. \end{cases} \tag{50}$$

Corollary 4. *Given* υ, δ_I, ξ_I, δ_{II}, ξ_{II} *as defined in point section 3.3 and a sampling matrix satisfying 3, then:*

$$M_{k-\lambda} \leqslant \min\left\{ \frac{\lambda\upsilon^2(1-\delta_I)^2}{2\log(1/\xi_I)\sigma^2 M_{k-\lambda}(1+\log(\lambda))}, \frac{\lambda\upsilon(1-\delta_I)}{2\log(1/\xi_I)\omega} \right\} \tag{51}$$

$$\Rightarrow \mathbb{P}\left[\hat{\rho}_k^\lambda > \upsilon, \rho_k^\lambda \leqslant \delta_I\upsilon \middle| \mathcal{F}_{k-\lambda} \right] < \xi_I; \textit{and} \tag{52}$$

$$M_{k-\lambda} \leqslant \min\left\{ \frac{\lambda\upsilon^2(\delta_{II}-1)^2}{2\log(1/\xi_{II})\sigma^2 M_{k-\lambda}(1+\log(\lambda))}, \frac{\lambda\upsilon(\delta_{II}-1)}{2\log(1/\xi_{II})\omega} \right\} \tag{53}$$

$$\Rightarrow \mathbb{P}\left[\hat{\rho}_k^\lambda \leqslant \upsilon, \rho_k > \delta_{II}\upsilon \middle| \mathcal{F}_{k-\lambda} \right] < \xi_{II}. \tag{54}$$

4.3. Estimating the uncertainty set and stopping criterion

Corollaries 3 and 4 provide a well-controlled uncertainty set and stopping criterion, yet require knowing $M_{k-\lambda}^2$, which is usually not available. As stated before, corollaries 3 and 4 can be operationalized by replacing $M_{k-\lambda}^2$ with $\hat{\iota}_k^\lambda$. Of course, $M_{k-\lambda}^2$ and $\hat{\iota}_k^\lambda$ must coincide in some sense in order for this estimation to be valid. Indeed, by theorem 3, both $M_{k-\lambda}^2$ and $\hat{\iota}_k^\lambda$ converge to zero as $k \to \infty$, which allows us to estimate $M_{k-\lambda}^2$ with $\hat{\iota}_k^\lambda$ to generate consistent estimators. However, we could also estimate $M_{k-\lambda}^2$ by 0 to generate consistent estimators, but these would be uninformative during finite time. Therefore, we must establish that estimating $M_{k-\lambda}^2$ by $\hat{\iota}_k^\lambda$ is also appropriate within some finite time. To do this we establish that the relative error between $M_{k-\lambda}^2$ and $\hat{\iota}_k^\lambda$ is controlled by a constant (in probability).

Accomplishing this task requires the following lemma, which gives a probability bound between $\hat{\iota}_k^\lambda$ and a specific intermediate quantity.

Lemma 8. *Under the conditions of theorem 3,* $\forall \delta > 0$

$$\mathbb{P}\left(\left|\frac{\hat{\iota}_k^\lambda - \frac{\sum_{i=k-\lambda+1}^{k}\left(\mathbb{E}\left[\|\tilde{r}_i\|_2^2|\mathcal{F}_{i-1}\right]\right)^2}{\lambda}}{M_{k-\lambda}^2}\right| > \delta \middle| \mathcal{F}_{k-\lambda}\right)$$
$$\leqslant (1+\lambda)\exp\left(-\min\left\{\frac{\delta^2\lambda}{2\left(2+\sqrt{\delta\lambda}\right)^2\sigma^2\left(1+\log\left(\lambda\right)\right)}, \frac{\delta\lambda}{2\omega\left(2+\sqrt{\delta\lambda}\right)}\right\}\right). \tag{55}$$

Proof. Using the definition of $\hat{\iota}_k^\lambda$ we have

$$\begin{aligned}
&\mathbb{P}\left(\left|\frac{\hat{\iota}_k^\lambda - \frac{\sum_{i=k-\lambda+1}^{k}\left(\mathbb{E}\left[\|\tilde{r}_i\|_2^2|\mathcal{F}_{i-1}\right]\right)^2}{\lambda}}{M_{k-\lambda}^2}\right| > \delta \middle| \mathcal{F}_{k-\lambda}\right)\\
&\quad= \mathbb{P}\left(\left|\sum_{i=k-\lambda+1}^{k}\frac{\|\tilde{r}_i\|_2^4 - \left(\mathbb{E}\left[\|\tilde{r}_i\|_2^2|\mathcal{F}_{i-1}\right]\right)^2}{\lambda M_{k-\lambda}^2}\right| > \delta \middle| \mathcal{F}_{k-\lambda}\right)\\
&\quad\leqslant \mathbb{P}\left(\sum_{i=k-\lambda+1}^{k}\left|\frac{\|\tilde{r}_i\|_2^4 - \left(\mathbb{E}\left[\|\tilde{r}_i\|_2^2|\mathcal{F}_{i-1}\right]\right)^2}{\lambda M_{k-\lambda}^2}\right| > \delta \middle| \mathcal{F}_{k-\lambda}\right)\\
&\quad\leqslant \mathbb{P}\left(\sum_{i=k-\lambda+1}^{k}\left|\frac{\|\tilde{r}_i\|_2^2 - \mathbb{E}\left[\|\tilde{r}_i\|_2^2|\mathcal{F}_{i-1}\right]}{\lambda M_{k-\lambda}}\right|\right.\\
&\qquad\left.\times\left|\frac{\|\tilde{r}_i\|_2^2 + \mathbb{E}\left[\|\tilde{r}_i\|_2^2|\mathcal{F}_{i-1}\right]}{M_{k-\lambda}}\right| > \delta \middle| \mathcal{F}_{k-\lambda}\right).
\end{aligned} \tag{56}$$

Then by defining a variable, $G > 2$, to partition (56) into disjoint sets and using the definition of measure,

$$\begin{aligned}
&\mathbb{P}\left(\sum_{i=k-\lambda+1}^{k}\left|\frac{\|\tilde{r}_i\|_2^2 - \mathbb{E}\left[\|\tilde{r}_i\|_2^2|\mathcal{F}_{i-1}\right]}{\lambda M_{k-\lambda}}\right|\left|\frac{\|\tilde{r}_i\|_2^2 + \mathbb{E}\left[\|\tilde{r}_i\|_2^2|\mathcal{F}_{i-1}\right]}{M_{k-\lambda}}\right| > \delta \middle| \mathcal{F}_{k-\lambda}\right)\\
&\quad= \mathbb{P}\left(\sum_{i=k-\lambda+1}^{k}\left|\frac{\|\tilde{r}_i\|_2^2 - \mathbb{E}\left[\|\tilde{r}_i\|_2^2|\mathcal{F}_{i-1}\right]}{\lambda M_{k-\lambda}}\right|\left|\frac{\|\tilde{r}_i\|_2^2 + \mathbb{E}\left[\|\tilde{r}_i\|_2^2|\mathcal{F}_{i-1}\right]}{M_{k-\lambda}}\right| > \delta,\right.\\
&\qquad\left.\bigcap_{i=k-\lambda+1}^{k}\left\{\left|\frac{\|\tilde{r}_i\|_2^2 + \mathbb{E}\left[\|\tilde{r}_i\|_2^2|\mathcal{F}_{i-1}\right]}{M_{k-\lambda}}\right| \leqslant G\right\} \middle| \mathcal{F}_{k-\lambda}\right)\\
&\qquad+ \mathbb{P}\left(\sum_{i=k-\lambda+1}^{k}\left|\frac{\|\tilde{r}_i\|_2^2 - \mathbb{E}\left[\|\tilde{r}_i\|_2^2|\mathcal{F}_{i-1}\right]}{\lambda M_{k-\lambda}}\right|\left|\frac{\|\tilde{r}_i\|_2^2 + \mathbb{E}\left[\|\tilde{r}_i\|_2^2|\mathcal{F}_{i-1}\right]}{M_{k-\lambda}}\right| > \delta,\right.\\
&\qquad\left.\bigcup_{i=k-\lambda+1}^{k}\left\{\left|\frac{\|\tilde{r}_i\|_2^2 + \mathbb{E}\left[\|\tilde{r}_i\|_2^2|\mathcal{F}_{i-1}\right]}{M_{k-\lambda}}\right| > G\right\} \middle| \mathcal{F}_{k-\lambda}\right)\\
&\quad\leqslant \mathbb{P}\left(\sum_{i=k-\lambda+1}^{k}\left|\frac{\|\tilde{r}_i\|_2^2 - \mathbb{E}\left[\|\tilde{r}_i\|_2^2|\mathcal{F}_{i-1}\right]}{\lambda M_{k-\lambda}}\right| > \frac{\delta}{G} \middle| \mathcal{F}_{k-\lambda}\right)\\
&\qquad+ \mathbb{P}\left(\bigcup_{i=k-\lambda+1}^{k}\left\{\left|\frac{\|\tilde{r}_i\|_2^2 + \mathbb{E}\left[\|\tilde{r}_i\|_2^2|\mathcal{F}_{i-1}\right]}{M_{k-\lambda}}\right| > G\right\} \middle| \mathcal{F}_{k-\lambda}\right).
\end{aligned} \tag{57}$$

From here we will present the bounds for the left and right terms of (57) separately. For the left term of (57) we use (35) and when $t \leqslant \lambda/\omega$ have

$$\mathbb{P}\left(\sum_{i=k-\lambda+1}^{k}\left|\frac{\|\tilde{r}_i\|_2^2-\mathbb{E}\left[\|\tilde{r}_i\|_2^2|\mathcal{F}_{i-1}\right]}{\lambda M_{k-\lambda}}\right|>\frac{\delta}{G}\middle|\mathcal{F}_{k-\lambda}\right) \tag{58}$$

$$\leqslant \exp\left(\frac{t^2\sigma^2\left(1+\log\left(\lambda\right)\right)}{2\lambda}-\frac{\delta t}{G}\right) \tag{59}$$

$$\leqslant \exp\left(-\min\left\{\frac{\delta^2\lambda}{2G^2\sigma^2\left(1+\log\left(\lambda\right)\right)},\frac{\lambda\delta}{2\omega G}\right\}\right), \tag{60}$$

where (60) comes from minimizing (58) in terms of t under the constraint $t \leqslant \lambda/\omega$. We next address the right side of (57) for which we have

$$\mathbb{P}\left(\bigcup_{i=k-\lambda+1}^{k}\left\{\left|\frac{\|\tilde{r}_i\|_2^2+\mathbb{E}\left[\|\tilde{r}_i\|_2^2|\mathcal{F}_{i-1}\right]}{M_{k-\lambda}}\right|>G\right\}\middle|\mathcal{F}_{k-\lambda}\right)$$

$$=\mathbb{P}\left(\bigcup_{i=k-\lambda+1}^{k}\left\{\left|\frac{\|\tilde{r}_i\|_2^2-\mathbb{E}\left[\|\tilde{r}_i\|_2^2|\mathcal{F}_{i-1}\right]+2\mathbb{E}\left[\|\tilde{r}_i\|_2^2|\mathcal{F}_{i-1}\right]}{M_{k-\lambda}}\right|>G\right\}\middle|\mathcal{F}_{k-\lambda}\right)$$

$$\leqslant\mathbb{P}\left(\bigcup_{i=k-\lambda+1}^{k}\left\{\left|\frac{\|\tilde{r}_i\|_2^2-\mathbb{E}\left[\|\tilde{r}_i\|_2^2|\mathcal{F}_{i-1}\right]}{M_{k-\lambda}}\right|+2>G\right\}\middle|\mathcal{F}_{k-\lambda}\right) \tag{61}$$

$$\leqslant\sum_{i=k-\lambda+1}^{k}\mathbb{P}\left(\left|\frac{\|\tilde{r}_i\|_2^2-\mathbb{E}\left[\|\tilde{r}_i\|_2^2|\mathcal{F}_{i-1}\right]}{M_{k-\lambda}}\right|>G-2\middle|\mathcal{F}_{k-\lambda}\right) \tag{62}$$

$$\leqslant\lambda\exp\left(\frac{t^2\sigma^2}{2}-t\left(G-2\right)\right) \tag{63}$$

$$\leqslant\lambda\exp\left(-\min\left\{\frac{\left(G-2\right)^2}{2\sigma^2},\frac{G-2}{2\omega}\right\}\right), \tag{64}$$

where (61) comes from (31), (63) comes from the (35) and (64) comes from minimizing (63) in terms of t under the constraint $t \leqslant 1/\omega$. Putting both parts together gives us,

$$\mathbb{P}\left(\left|\frac{\hat{\iota}_k^\lambda-\iota_k^\lambda}{M_{k-\lambda}^2}\right|>\delta\middle|\mathcal{F}_{k-\lambda}\right)$$

$$\leqslant\inf_{G>2}\exp\left(-\min\left\{\frac{\delta^2\lambda}{2G^2\sigma^2\left(1+\log\left(\lambda\right)\right)},\frac{\delta\lambda}{2\omega G}\right\}\right)$$

$$+\lambda\exp\left(-\min\left\{\frac{\left(G-2\right)^2}{2\sigma^2},\frac{G-2}{2\omega}\right\}\right).$$

We can then observe that when $G \geqslant 2+\sqrt{\delta\lambda}$ it is the case that

$$\exp\left(-\min\left\{\frac{\delta^2\lambda}{2G^2\sigma^2\left(1+\log\left(\lambda\right)\right)},\frac{\delta\lambda}{2\omega G}\right\}\right)$$

$$\geqslant\exp\left(-\min\left\{\frac{\left(G-2\right)^2}{2\sigma^2},\frac{G-2}{2\omega}\right\}\right).$$

Thus, if we let $Y = 2 + \sqrt{\delta\lambda}$ we can upper bound the right-hand side of (65) in the following manner,

$$\begin{aligned}
&\inf_{G>2} \exp\left(-\min\left\{\frac{\delta^2\lambda}{2G^2\sigma^2\left(1+\log\left(\lambda\right)\right)}, \frac{\delta\lambda}{2\omega G}\right\}\right) \\
&\quad + \lambda\exp\left(-\min\left\{\frac{\left(G-2\right)^2}{2\sigma^2}, \frac{G-2}{2\omega}\right\}\right) \\
&\leqslant \inf_{G>Y}\left(1+\lambda\right)\exp\left(-\min\left\{\frac{\delta^2\lambda}{2G^2\sigma^2\left(1+\log\left(\lambda\right)\right)}, \frac{\delta\lambda}{2\omega G}\right\}\right) \\
&\leqslant \left(1+\lambda\right)\exp\left(-\min\left\{\frac{\delta^2\lambda}{2\left(2+\sqrt{\delta\lambda}\right)^2\sigma^2\left(1+\log\left(\lambda\right)\right)}, \frac{\delta\lambda}{2\omega\left(2+\sqrt{\delta\lambda}\right)}\right\}\right).
\end{aligned}$$

□

We now present our bound on the relative error between $\hat{\iota}_k^\lambda$ and $M_{k-\lambda}^2$.

Theorem 5. *Under the conditions of theorem 3, for $\delta > 0$, $M_{k-\lambda}^2$ as described in theorem 4,*

$$\mathbb{P}\left(\left|\frac{M_{k-\lambda}^2 - \hat{\iota}_k^\lambda}{M_{k-\lambda}^2}\right| > 1+\delta, M_{k-\lambda}^2 \neq 0 \middle| \mathcal{F}_{k-\lambda}\right) \tag{65}$$

$$\leqslant \left(1+\lambda\right)\exp\left(-\min\left\{\frac{\delta^2\lambda}{2\left(2+\sqrt{\delta\lambda}\right)^2\sigma^2\left(1+\log\left(\lambda\right)\right)}, \frac{\delta\lambda}{2\omega\left(2+\sqrt{\delta\lambda}\right)}\right\}\right). \tag{66}$$

Proof. First,

$$\left|\frac{M_{k-\lambda}^2 - \hat{\iota}_k^\lambda}{M_{k-\lambda}^2}\right| \leqslant \left|\frac{M_{k-\lambda}^2 - \frac{\sum_{i=k-\lambda+1}^k \left(\mathbb{E}\left[\|\tilde{r}_i\|_2^2 | \mathcal{F}_{i-1}\right]\right)^2}{\lambda}}{M_{k-\lambda}^2}\right| \tag{67}$$

$$+ \left|\frac{\frac{\sum_{i=k-\lambda+1}^k \left(\mathbb{E}\left[\|\tilde{r}_i\|_2^2 | \mathcal{F}_{i-1}\right]\right)^2}{\lambda} - \hat{\iota}_k^\lambda}{M_{k-\lambda}^2}\right|$$

$$\leqslant 1 + \left|\frac{\frac{\sum_{i=k-\lambda+1}^k \left(\mathbb{E}\left[\|\tilde{r}_i\|_2^2 | \mathcal{F}_{i-1}\right]\right)^2}{\lambda} - \hat{\iota}_k^\lambda}{M_{k-\lambda}^2}\right|. \tag{68}$$

We now apply the bound in lemma 8 to conclude. □

Owing to lemma 5, the relative error between $\hat{\iota}_k^\lambda$ and $M_{k-\lambda}^2$ is reasonably well controlled for practical purposes. As a result, we can use $\hat{\iota}_k^\lambda$ as a plug-in estimator for $M_{k-\lambda}^2$ for the uncertainty set, (49), to produce the estimated uncertainty set suggested in Line 18 of algorithm 1. Additionally, we can use $\hat{\iota}_k^\lambda$ as a plug-in estimator for $M_{k-\lambda}^2$ for the stopping condition controls in (51) and (53) to produce the estimated stopping criterion in Line 2 of algorithm 1.

5. Experimental results

We conduct three experiments. Our first experiment compares the performance of ρ_k^λ at tracking the error when $\lambda = 1$ and $\lambda > 1$. Our second experiment is solving a collocation problem reformulated as a streaming problem (see appendix A.3). We close this section with a timing comparison between algorithm 1> and periodic residual calculation methods on a 125 000 by 125 000 system.

5.1. Benefits of $\lambda > 1$

To understand why it is valuable to track progress with $\lambda > 1$ we perform an experiment where we run a GBRK with $B = I$ and random row subset samples for 1000 iterations on a Phillips matrix from the *MatrixDepot* package in Julia [19]. Then we compare the values of (7) with $\lambda = 1$ and $\lambda = 100$ to the absolute error at the corresponding iteration, which is displayed in figure 1(a). This shows how closely the two tracking metrics align with the quantity we care about tracking, the absolute error. Then, to demonstrate how stopping using $\lambda = 1$ can be problematic, we perform the same experiment on 15 other sample matrix systems. For each system, we record the iteration, k' when the absolute error is less than 1 and record the tracking metrics value at that iteration giving us $\hat{\rho}_{k'}^1$ and $\hat{\rho}_{k'}^{100}$. Then we compute the number of early stopping occurrences by counting the number of times $\hat{\rho}_k^1 < \hat{\rho}_{k'}^1$ for $k < k'$ and the number of times $\hat{\rho}_k^{100} < \hat{\rho}_{k'}^{100}$ for $k < k'$. These counts tell us how often using each method for stopping decisions results in us failing to achieve our objective. These results are shown in figure 1(b). As a reminder to the reader, we re-state points (ii) and (iii) from section 3.3.1 and remark how this experiment elucidates these points.

(ii) The expected value of $\hat{\rho}_k^\lambda$ (relative to the observation pairs at a given update),

$$\rho_k^\lambda = \sum_{i=k-\lambda+1}^{k} \frac{\mathbb{E}\left[\|\tilde{A}_i x_{i-1} - \tilde{b}_i\|_2^2 | \mathcal{F}_{i-1}\right]}{\lambda}, \tag{8}$$

seems to better track the absolute error of the iterate. Indeed, figure 1(a) plots the behavior ρ_k^1 and ρ_k^{100} against the absolute error of the iterate for a GBRK method applied to a linear system generated from the Phillips matrix in Matrix Depot [19], which shows that the estimator with $\lambda > 1$ corresponds better to the absolute error in comparison to the case when $\lambda = 1$.

(iii) There seems to be many cases where the use of $\lambda > 1$ allows for improved stopping performance with relation to the absolute error. Specifically, consider applying a GBRK method to 15 linear systems generated from matrices in Matrix Depot [19]. Figure 1(b) plots the number of times ρ_k^1 and ρ_k^{100} fall below the residual threshold but $k < k'$, which is an imperfect, but informative proxy for determining if the GBRK is stopped too soon. We see that, for nearly half of the systems, ρ_k^1 would have stopped early, whereas ρ_k^{100} does not.

5.2. Collocation problem

To demonstrate that algorithm 1 works on large-scale streaming problems, we turn to the collocation problem laid out in appendix A.3. In this problem, we verify that the uncertainty set for

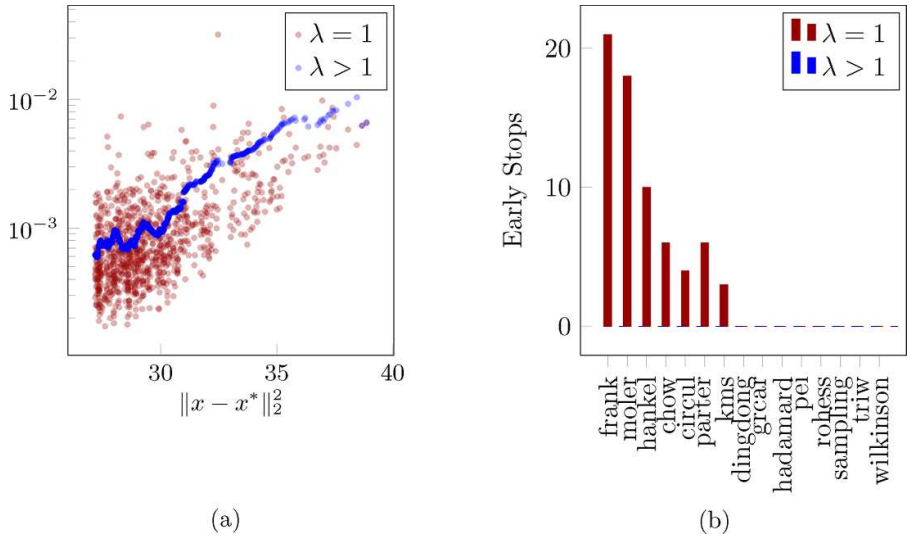

Figure 1. (a) Plot of the ρ_k^λ when $\lambda = 1$ (red) and $\lambda > 1$ (blue) at different values of absolute error for a Phillips matrix. (b) Graph showing the number of iterates with a ρ_k^λ less than the ρ_κ^λ where $\kappa = \min\{k : \|x_k - x^*\|_2 < 1\}$ for $\lambda = 1$ (red) and $\lambda > 1$ (blue).

$\hat{\rho}_k^\lambda$ contains ρ_k^λ at appropriate rates and that the stopping criterion correctly controls stopping errors. To accomplish this task we perform the following experiment.

Experiment 1. Given the gap between fixed grid points $\epsilon = 1/99$, the number of sample rows $p = 20$, the moving average width $\lambda_1 \in \{100, 300\}$, the uncertainty set parameter α, and the stopping criterion parameters $(\upsilon, \delta_I, \delta_{II}, \xi_I, \xi_{II}) = (400, .9, 1.1, .01, .01)$ we:

(i) Generate an equally spaced grid on the unit cube with the gap between grid points being ϵ. This equates to there being $(1 + 1/\epsilon)^3$ fixed points on the cube.
(ii) Generate $(\tilde{A}_k, \tilde{b}_k)$ by generating 20 random coordinates as explained in appendix A.3.
(iii) Run algorithm 1 using $(\tilde{A}_k, \tilde{b}_k)$'s and save at each iteration the x_k, $\hat{\rho}_k^\lambda$, $\hat{\iota}_k^\lambda$, $\|r_k\|_2^2$, and the width of moving average λ.
(iv) At each x_k approximate $\mathbb{E}[\|\tilde{r}_k\|_2^2 | \mathcal{F}_{k-1}]$ using a Monte Carlo simulation with 100 samples.
(iv) Approximate ρ_k^λ at the k^{th} iteration by taking a moving average of the $\hat{\mathbb{E}}[\|\tilde{r}_k\|_2^2 | \mathcal{F}_{k-1}]$s with the same width as $\hat{\rho}_k^\lambda$ at iteration k.

Practically, our value of σ^2 in (A.14) (i.e. $V/2$) is roughly 9 trillion in this experiment, which is far too conservative to be useful[3]. Thus, instead of using $V/2$ for σ^2, we estimate it by computing

$$\text{Variance}\left(\frac{|\hat{\mathbb{E}}\left[\|\tilde{r}_k\|_2^2 | \mathcal{F}_{k-1}\right] - \|r_k\|_2^2|}{\hat{\mathbb{E}}\left[\|\tilde{r}_k\|_2^2 | \mathcal{F}_{k-1}\right]}, k = 1, \ldots, 125\right), \tag{69}$$

with $\|r_k\|_2^2$ coming from step (iii) and $\hat{\mathbb{E}}[\|\tilde{r}_k\|_2^2 | \mathcal{F}_{k-1}]$ coming from step (v) of experiment 1. This estimator is more optimistic than $V/2$, yet seems to be appropriate. It should be noted that

[3] Based on different choices of ϵ it does not appear the $V/2$ ever really accurately reflects the variance.

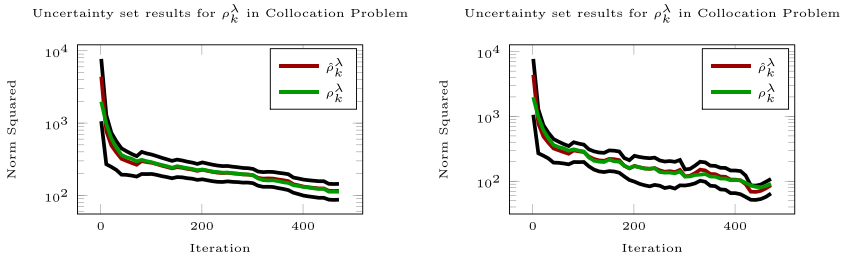

Figure 2. In the left plot, the black lines represent the upper and lower bounds of the uncertainty set, the green line represents ρ_k^λ, the red line represents $\hat{\rho}_k^\lambda$ for $\lambda = 300$ and the right plot represents the same values for $\lambda = 100$.

Table 1. Interval failure rates at different moving average window widths.

Width	10	50	100	200	300
Failure rate	0.0148	0.0148	0.006	0.004	0.004

Table 2. Interval failure rates with different σ^2 estimates and moving average window widths.

	$k=5$		$k=125$		$k=474$	
Dataset	$\lambda=10$	$\lambda=300$	$\lambda=10$	$\lambda=300$	$\lambda=10$	$\lambda=300$
σ^2	0.172	0.172	0.079	0.079	0.088	0.088
Failure rate	0.0148	0.004	0.0148	0.004	0.0148	0.004

the choice of using the first 125 iterations for the variance estimation is somewhat arbitrary and many other choices of the number iterations could be made to produce approximately the same variance estimate as can be seen in table 2.

The results of this experiment can be observed in figure 2 for two different choices of λ_1 one being $\lambda_1 = 100$ and one being $\lambda_1 = 300$. As should be clear from figure 2 the 95% uncertainty sets correctly cover ρ_k^λ at 99.6% of the iterations for $\lambda_1 = 300$ and 99.4% for $\lambda_1 = 100$, which is still conservative despite us not using the larger estimate of σ^2 from (A.14). This rate coverage failure also does not change much when the width is set to be some other value as seen in table 1. Additionally, when the stopping criterion is satisfied there are no instances where the two types of stopping errors occur. This indicates that even at large scales algorithm 1 still performs well.

5.3. Timing comparison with naive methods

We now demonstrate the computational benefits of our method over periodic calculation of the full residual on a 125 000 by 125 000 system. We performed this experiment using a single thread of a Xeon E5-2680 v3 @ 2.50GHz with 32 GB of memory. For all the GBRKs we chose $B = I$ as there is not currently a good method for alternative choices of B *a priori*, although this would be an interesting piece of future work. The results of the experiment can be found in table 3.

Table 3. Iterations achieved in four-day run on a collocation problem with a $51 \times 51 \times 51$ grid collocation points.

Method	Iterations	Time per iteration
Algorithm 1	7819	44.2 s
Full Residual every GBRK update	16	22, 109 s
Full Residual every 1000 GBRK updates	4162	83.03 s

When comparing the performance of algorithm 1 to other methods of tracking, we can see substantial benefits of algorithm 1 compared to alternatives. We do this by calculating the number of iterations completed in a four day window by three methods: algorithm 1, calculating the full residual after each GBRK update, and calculating the full residual after every 1000 GBRK updates.

6. Conclusion

To address the issue of effectively tracking and stopping the progress of a streaming solver, we have presented a computationally efficient estimator and uncertainty set for a moving average of the residuals. We then rigorously demonstrated the effectiveness of this estimator using only assumptions about the streams having a set of consistent solutions and the norm of the residuals of the streams being sub-Exponential. We additionally show that the assumptions used to make these conclusions are relatively weak and apply to a large class of common problems. Moreover, we verified our methodology by successfully applying it to a large-scale collocation problem and demonstrated its computational benefits over alternative methods.

Data availability statement

The data that support the findings of this study are openly available at the following URL/DOI: https://github.com/npritch928/tracking_kazmarz_data.

Acknowledgments

This material is based upon work supported by the National Science Foundation (NSF) under Awards 2309445 and 2309446.

Appendix. Areas of Applications

With the theory for algorithm 1 established, we now move on to presenting common areas of application that satisfy assumptions 1 and 3 and for which it is sensible apply algorithm 1. The situations that will be discussed are Johnson-Lindenstrauss matrix sketches, Block Randomized Kaczmarz, and the collocation problem arising from boundary element analysis.

A.1. Johnson-Lindenstrauss matrix sketching

With the pervasiveness of large scale linear system in areas such as Gaussian Process modeling [20], optimization [21], and network analysis [22], random algorithms have been shown to

Table A1. Values of C and ω in definition 2 for common sampling methods.

	C	ω
Gaussian Matrix [31]	1.1	.47
Achlioptas [23]	1.16	.46
FJLT [24]	0.83	.7

Table A2. Conservative choices of the contraction parameter, η, by sketching method.[a]

Sampling method	Achlioptas [23]	FJLT [24]	Gaussian [31]
η	26	188	26

[a] By conservative, we mean that these are choices of η that are the largest integer such that the coverage failure rate of the interval is as large as possible, but still less than the designed rate.

accelerate the speed of achieving good approximate solutions. One popular form of random transformation is the use of a random matrix that satisfies the Johnson-Lindenstrauss property [23–25], whose definition follows.

Definition 2. A matrix $S \in \mathbb{R}^{m\times p}$ satisfies the Johnson–Lindenstrauss property if there exists constants $C, \omega > 0$ s.t. for all $\delta \geqslant 0$ and for any $x \in \mathbb{R}^m$,

$$\mathbb{P}\left(\left|\|S^\top x\|_2^2 - \|x\|_2^2\right| > \delta\|x\|_2^2\right) \quad < 2e^{-\min\left\{\left(Cp\delta^2\right)/2, \delta/(2\omega)\right\}}. \tag{A.1}$$

Remark 4. Matrices S satisfying definition 2 are known as sketching matrices and have broad applications in numerical linear algebra [26, 27]. Examples of such matrices include the Gaussian matrix [28], the Achlioptas sparse sampling matrix [23], and the Fast Johnson-Lindenstrauss transform [24]. These methods will be the focus of experiment section, and thus we have included the values C and ω for when $\delta = 1$ in table A1.

By applying a matrix satisfying definition 2 to a consistent linear system with coefficient matrix $A \in \mathbb{R}^{m\times n}$ and constant vector $b \in \mathbb{R}^m$, and we denote $\tilde{A} = S^\top A$ and $\tilde{b} = S^\top b$. Then, we readily see that $(\tilde{A}, \tilde{b})$ satisfy assumptions 3 and 1 [29]. This type of process, in addition to allowing faster approximations of large consistent systems, has the potential for use in large network systems where one may compute an Adjacency-based, or Graph-Laplacian-based, measure such as Katz centrality [22], and still maintain the privacy of the individual according to a particular differential privacy standard [4, 30]. In this case, one could use realizations a of Johnson-Lindenstrauss sketch and generate $(\tilde{A}_k, \tilde{b}_k)$, which can then be used in the GBRK framework to compute the desired metric without violating the privacy of the individuals.

A.2. Computed tomography (Block Randomized Kaczmarz)

A common area of use for GBRK is in computed tomography, which aims to non-invasively generate the image of the inside of an object [32]. It accomplishes this by collecting the change of intensity of an electromagnetic wave as it is passed through an object at multiple different angles around the object [32]. The ability to non-invasively form images of the inside of an object has made tomography widely useful in numerous fields including meteorology [33] and medicine [32].

Tomography measures the change in intensity of an electromagnetic wave along different paths, L. By noting that different materials absorb different amounts of energy, at each coordinate (z,y) we can define the function representing this energy absorption, $f(z, y)$. This

function can be related to the change in intensity along a particular path L through the equation $\log(I_0/I_1) = \int_L f(z,y)\mathrm{d}L$ [34]. It was shown that $f(z,y)$ could be determined using an infinite number of these path observations [35]. As collecting an infinite number of these path observations is impossible, tomographic techniques are developed to approximately reconstruct this function under a finite number of observations [34].

One common set of techniques are known as algebraic approaches and work by dividing an image into $N = n^2$ pixels. For each of these pixels it is assumed that the energy absorption of the material of the object is the same, i.e. $f(z,y)$ is constant at each pixel. Then for each of the m paths L_i, we observe a change of intensity, b_i, and the length path through each pixel, a_{ij} [34]. With this information we know that $b_i = \sum_{j=1}^{N} a_{ij}x_j$, where $x_j = f(z,y)$ at pixel j. Combining all path observations gives us the linear system, $Ax = b$, which is typically nonsymmetric and consistent with a very large dimension [34].

Often this system is solved using some variation of the Kaczmarz algorithm [34], in what is also referred to as the Algebraic Reconstruction Technique [1, 2]. Under this setting the blocks $\{(\tilde{A}_k, \tilde{b}_k) : k \in \mathbb{N}\}$, are independent, identically distributed subsets of the path equations for the tomographic linear system, whose coefficient matrix is given by $A \in \mathbb{R}^{m\times N}$ and whose constant vector is $b \in \mathbb{R}^m$.

By updating a solution with randomly selected row blocks, as is done when applying Block Randomized Kaczmarz in the tomographic setting, this model satisfies assumption 3. The most general form of these Kaczmarz updates allows for each row to be weighted by its own respective constant and each block, $\mathcal{J}$, have its own probability of selection given by $P_{\mathcal{J}}$ [13]. This is stated formally now.

Proposition 1. *Let $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$ such that $\exists x^* \in \mathbb{R}^n$ where $Ax^* = b$. Let $\mathcal{J}$ be a random subset of $\{1,\dots,m\}$ and let $P_i > 0$ be the probability $i \in \mathcal{J}$. Let $w_i > 0$ for $i = 1,\dots,m$. Define $S_{\mathcal{J}} \in \mathbb{R}^{m\times p}$ to be the matrix whose columns are scalings of standard basis elements indexed by $\mathcal{J}$, where the scaling for the column corresponding to index $i \in \mathcal{J}$ is w_i. Let $(\tilde{A}_{\mathcal{J}}, \tilde{b}_{\mathcal{J}}) = (S_{\mathcal{J}}^{\top}A, S_{\mathcal{J}}^{\top}b)$. Then, for any $x \in \mathbb{R}^n$ such that $Ax \neq b$,*

$$\frac{\|\tilde{A}_{\mathcal{J}}x - \tilde{b}_{\mathcal{J}}\|_2^2 - \mathbb{E}\left[\|\tilde{A}_{\mathcal{J}}x - \tilde{b}_{\mathcal{J}}\|_2^2\right]}{\mathbb{E}\left[\|\tilde{A}_{\mathcal{J}}x - \tilde{b}_{\mathcal{J}}\|_2^2\right]} \sim SE\left(\frac{\max_i w_i^2}{2\left(\min_i P_i w_i^2\right)}, 0\right). \tag{A.2}$$

Proof. By lemma 2, all bounded random variables are sub-Exponential with $\omega = 0$. To show the block residuals are bounded in this instance, we find a lower bound on $\mathbb{E}[\|\tilde{A}_{\mathcal{J}}x - \tilde{b}_{\mathcal{J}}\|_2^2]$ and an upper bound on $\|\tilde{A}_{\mathcal{J}}x - \tilde{b}_{\mathcal{J}}\|_2^2$. For the lower bound, let $a_i \in \mathbb{R}^n$ denote the i^{th} row of A, and let $\mathbb{P}(i \in \mathcal{J}) = P_i$. This allows us to write

$$\mathbb{E}\left[\|\tilde{A}_{\mathcal{J}}x - \tilde{b}_{\mathcal{J}}\|_2^2\right] = \sum_{i=1}^{m} P_i w_i^2 \left[a_i^{\top}(x - x^*)\right]^2 \tag{A.3}$$

$$\geqslant \left(\min_i P_i w_i^2\right) \sum_{i=1}^{m} \left[a_i^{\top}(x - x^*)\right]^2. \tag{A.4}$$

For the upper bound on $\|\tilde{A}_{\mathcal{J}}x - \tilde{b}_{\mathcal{J}}\|_2^2$ we note that

$$\sum_{i\in\mathcal{J}} w_i^2 \left[a_i^{\top}(x - x^*)\right]^2 \leqslant \left(\max_i w_i^2\right) \sum_{i=1}^{m} \left[a_i^{\top}(x - x^*)\right]^2. \tag{A.5}$$

Combing these lower and upper bounds, we conclude that for any x such that $Ax \neq b$,

$$-1 \leqslant \frac{\|\tilde{A}_{\mathcal{J}}x - \tilde{b}_{\mathcal{J}}\|_2^2 - \mathbb{E}\left[\|\tilde{A}_{\mathcal{J}}x - \tilde{b}_{\mathcal{J}}\|_2^2\right]}{\mathbb{E}\left[\|\tilde{A}_{\mathcal{J}}x - \tilde{b}_{\mathcal{J}}\|_2^2\right]} \leqslant \frac{\max_i w_i^2}{\min_i P_i w_i^2} - 1. \tag{A.6}$$

Giving us that the relative error is bounded. We then obtain the result by applying lemma 2. □

In the case where the rows are selected with equal probability and equal weights, this bound is equivalent to the variance bound being m/p. As this grows with the dimension of the matrix, concerns may arise relating to the tightness of this bound. While a tighter bound is not possible for general matrices, we can tighten this bound depending on the block condition numbers in the following proposition.

Proposition 2. *Let the conditions in proposition 1 hold. Further, if we let $s_{\max_p}, s_{\min_p}$ be the largest and smallest singular value of all row blocks of a matrix A containing p rows. Then, for any $x \in \mathbb{R}^n$ such that $Ax \neq b$,*

$$\frac{\|\tilde{A}_{\mathcal{J}}x - \tilde{b}_{\mathcal{J}}\|_2^2 - \mathbb{E}\left[\|\tilde{A}_{\mathcal{J}}x - \tilde{b}_{\mathcal{J}}\|_2^2\right]}{\mathbb{E}\left[\|\tilde{A}_{\mathcal{J}}x - \tilde{b}_{\mathcal{J}}\|_2^2\right]} \sim \mathrm{SE}\left(\frac{s_{\max_p}^2}{2s_{\min_p}^2}, 0\right). \tag{A.7}$$

Proof. We note that

$$\frac{\|\tilde{A}_{\mathcal{J}}x - \tilde{b}_{\mathcal{J}}\|_2^2}{\mathbb{E}\left[\|\tilde{A}_{\mathcal{J}}x - \tilde{b}_{\mathcal{J}}\|_2^2\right]} \leqslant \frac{s_{\max_p}^2}{s_{\min_p}^2}. \tag{A.8}$$

This allows us to conclude that

$$-1 \leqslant \frac{\|\tilde{A}_{\mathcal{J}}x - \tilde{b}_{\mathcal{J}}\|_2^2 - \mathbb{E}\left[\|\tilde{A}_{\mathcal{J}}x - \tilde{b}_{\mathcal{J}}\|_2^2\right]}{\mathbb{E}\left[\|\tilde{A}_{\mathcal{J}}x - \tilde{b}_{\mathcal{J}}\|_2^2\right]} \leqslant \frac{s_{\max_p}^2}{s_{\min_p}^2} - 1. \tag{A.9}$$

The result then follow from lemma 2. □

A.3. Collocation problem with randomly sampled coordinates from a fixed set

Our final example of a problem that satisfies assumption 3 is the collocation problem in boundary element analysis. We will describe this problem in detail for a specific setting that will be used in our largest numerical experiment (see section 5.2); however, the results from this subsection will hold for more general problems as well.

In this problem, we look to approximate the unknown solution, $u(t)$, with $t = (t^{(1)}, t^{(2)}, t^{(3)}) \in [0,1]^3$, that satisfies

$$\Delta\left(u\left(t\right)\right) = \frac{-7\pi^2}{2}\sin\left(\pi t^{(1)}\right)\sin\left(\frac{\pi t^{(2)}}{2}\right)\sin\left(\frac{3\pi t^{(3)}}{2}\right), t \in [0,1]^3, \tag{A.10}$$

$$u\left(t\right) = \sin\left(\pi t^{(1)}\right)\sin\left(\pi t^{(2)}/2\right)\sin\left(3\pi t^{(3)}/2\right), t \in \partial\left[0,1\right]^3, \tag{A.11}$$

where $t \in \partial[0,1]^3$ represents the t being on boundary of the cube; and Δ is the Laplace operator.

To make this approximation, we wish to find the coefficients, x_j, of the linear combination of Quadric Radial Basis functions, $\phi(t, \chi_j) = \sqrt{\|t - \chi_j\|_2^2 + 1}$, and their Laplacians, evaluated at a fixed and finite set of control points $\{\chi_j\}$ equally spaced throughout the unit cube.

Since both the Laplacian and the boundary condition are linear operators, we can represent (A.10) as a streaming linear system, whose streams consist of coordinates $t_i \in [0,1]^3$ selected uniformly at random from either the interior control points (with probability 2/3), the face control points (with probability 1/6), or the edge control points (with probability 1/6).

With the sampling points t_i and control points χ_j, we form a linear system defined by

$$\tilde{A}^{(i,j)} = \begin{cases} \phi_j(t_i,\chi_j) & \text{if } t_i \in \partial[0,1]^3 \\ \Delta\phi_j(t_i,\chi_j) & \text{if } t_i \in (0,1)^3 \end{cases}, \tag{A.12}$$

$$\tilde{b}^{(i)} = \begin{cases} \sin\left(\pi t_i^{(1)}\right)\sin\left(\pi t_i^{(2)}/2\right)\sin\left(3\pi t_i^{(3)}/2\right) & \text{if } t_i \in \partial[0,1]^3 \\ \frac{-7\pi^2}{2}\sin\left(\pi t_i^{(1)}\right)\sin\left(\frac{\pi t_i^{(2)}}{2}\right)\sin\left(\frac{3\pi t_i^{(3)}}{2}\right) & \text{if } t_i \in (0,1)^3. \end{cases} \tag{A.13}$$

As the set of sampling and control points are the same, the resulting linear system is consistent as A is nonsingular [36, theorem 2.2]. This method is designed such that the blocks are generated at each iteration based on the coordinates sampled meaning the full system never has to be stored. The sampling method proposed is one approach to ensure that update contains information from all of possible constraints (the faces, the edges, and the interior). This particular problem also satisfy assumption 3.

Proposition 3. *Let $\{t_1,\ldots,t_p\}$ be chosen independently, as described above. Let the entries of $\tilde{A}_k$ be defined according to (A.12) and those of $\tilde{b}_k$ be defined according to (A.13). Then, if we let $V = 9N/\sigma_{\min}(\mathbb{E}[\tilde{A}_k^\top \tilde{A}_k])$, where $\sigma_{\min}$ is the smallest non-zero singular value and N is the number of columns in $\tilde{A}_k$,*

$$\frac{\|\tilde{A}_k x - \tilde{b}_k\|_2^2 - \mathbb{E}\left[\|\tilde{A}_k x - \tilde{b}_k\|_2^2\right]}{\mathbb{E}\left[\|\tilde{A}_k x - \tilde{b}_k\|_2^2\right]} \sim \mathrm{SE}\left(\frac{V}{2}, 0\right). \tag{A.14}$$

Proof. Let $x^* \in \mathbb{R}^n$ denote any vector such that $\mathbb{P}(\tilde{A}_k x^* = \tilde{b}_k) = 1$. For any $x \in \mathbb{R}^n$ such that $\mathbb{P}(\tilde{A}_k x = \tilde{b}_k) < 1$, we can decompose $x - x^*$ into $u \in \mathrm{row}(\mathbb{E}[\tilde{A}_k^\top \tilde{A}_k])$ and $v \in \mathrm{null}(\mathbb{E}[\tilde{A}_k^\top \tilde{A}_k])$. By construction, $u \neq 0$. Moreover, $v \in \mathrm{null}(\tilde{A}_k^\top \tilde{A}_k)$ with probability one (otherwise we would have a contradiction with $v \in \mathrm{null}(\mathbb{E}[\tilde{A}_k^\top \tilde{A}_k])$).

Therefore,

$$0 \leqslant \frac{\|\tilde{A}_k x - \tilde{b}_k\|_2^2}{\mathbb{E}\left[\|\tilde{A}_k x - \tilde{b}_k\|_2^2\right]} = \frac{\|\tilde{A}_k u\|_2^2}{u^\top \mathbb{E}\left[\tilde{A}_k^\top \tilde{A}_k\right] u} \leqslant \frac{\|\tilde{A}_k\|_2^2}{\sigma_{\min}\left(\mathbb{E}\left[\tilde{A}_k^\top \tilde{A}_k\right]\right)}, \tag{A.15}$$

where $\sigma_{\min}$ denotes the smallest non-zero singular value of the given matrix. Using (A.12), we can find the maximum possible value of $\tilde{A}_k$ on the unit cube is 3; thus, using the equivalence between the infinity and two norms $\|\tilde{A}_k\|_2^2 \leqslant 9N$. So we can set $V = \frac{9N}{\sigma_{\min}(\mathbb{E}[\tilde{A}_k^\top \tilde{A}_k])}$.

Hence,

$$-1 \leqslant \frac{\|\tilde{A}_k x - \tilde{b}_k\|_2^2}{\mathbb{E}\left[\|\tilde{A}_k x - \tilde{b}_k\|_2^2\right]} - 1 \leqslant V - 1. \tag{A.16}$$

Applying lemma 2 gives the conclusion. □

It should be emphasized that this result holds true for any right-hand side provided that it is bounded over the domain of the collocation problem. To generalize this result to different right-hand sides one can replace the 9 in V with the maximum of the right-hand side over the desired domain. Additionally, it should be noted that in practice $V/2$ is far too loose of a bound

on to σ and should be replaced with estimates from more practical techniques, such as those presented in section 5.2.

ORCID iDs

Nathaniel Pritchard https://orcid.org/0000-0003-0304-1847
Vivak Patel https://orcid.org/0000-0003-4130-0897

References

[1] Birk M, Dapp R, Ruiter N V and Becker J 2014 GPU-based iterative transmission reconstruction in 3D ultrasound computer tomography *J. Parallel Distrib. Comput.* **74** 1730–43
[2] Kak A C and Slaney M 2001 *Principles of Computerized Tomographic Imaging* (Society for Industrial and Applied Mathematics)
[3] Brebbia C A, Domínguez J and Lambros Tassoulas J 1989 Boundary elements: an introductory course *Boundary Elements: an Introductory Course* (Computational Mechanics Publications)
[4] Upadhyay J 2014 Differentially private linear algebra in the streaming model *CoRR* (arXiv:1409.5414)
[5] Gower R M and Richtárik P 2015 Randomized iterative methods for linear systems *SIAM J. Matrix Anal. Appl.* **36** 1660–90
[6] Clarkson K L and Woodruff D P 2009 Numerical linear algebra in the streaming model *Proc. 41st Annual ACM Symp. on Theory of Computing* (*STOC'09*) (Association for Computing Machinery) pp 205–14
[7] Patel V, Jahangoshahi M and Adrian Maldonado D 2023 Randomized block adaptive linear system solvers *SIAM J. Matrix Anal. Appl.* **44** 1349–69
[8] Dereziński M and Rebrova E 2024 Sharp analysis of sketch-and-project methods via a connection to randomized singular value decomposition *SIAM J. Math. Data Sci.* **6** 127–53
[9] Richtárik P and Takáč M 2017 Stochastic reformulations of linear systems: algorithms and convergence theory *SIAM J. Matrix Anal. Appl.* **41** 06
[10] Patel V, Jahangoshahi M and Maldonado D A 2021 An implicit representation and iterative solution of randomly sketched linear systems *SIAM J. Matrix Anal. Appl.* **42** 800–31
[11] Needell D and Tropp. J A 2014 Paved with good intentions: analysis of a randomized block Kaczmarz method *Linear Algebr. Appl.* **441** 199–221
[12] Pritchard N and Patel V 2023 Towards practical large-scale randomized iterative least squares solvers through uncertainty quantification *SIAM/ASA J. Uncertain. Quantification* **11** 996–1024
[13] Necoara I 2019 Faster randomized block Kaczmarz algorithms *SIAM J. Matrix Anal. Appl.* **40** 1425–52
[14] Haddock J, Needell D, Rebrova E and Swartworth W 2022 Quantile-based iterative methods for corrupted systems of linear equations *SIAM J. Matrix Anal. Appl.* **43** 605–37
[15] Strohmer T and Vershynin R 2007 A randomized Kaczmarz algorithm with exponential convergence *J. Fourier Anal. Appl.* **15** 262–78
[16] Wainwright M J 2019 *High-Dimensional Statistics: a Non-Asymptotic Viewpoint Cambridge Series in Statistical and Probabilistic Mathematics* (Cambridge University Press)
[17] Vershynin R 2018 *High-Dimensional Probability: an Introduction With Applications in Data Science* (*Cambridge Series in Statistical and Probabilistic Mathematics*) (Cambridge University Press)
[18] Durrett R 2013 *Probability: Theory and Examples* 3rd edn (Cambridge University Press) (https://doi.org/10.1017/CBO9780511779398)
[19] Zhang W and Higham N 2016 Matrix depot: an extensible test matrix collection for julia *PeerJ Comput. Sci.* **2** e58
[20] Edward Rasmussen C and Williams C K I 2005 *Gaussian Processes for Machine Learning* (*Adaptive Computation and Machine Learning*) (The MIT Press)
[21] Dembo R S, Eisenstat S C and Steihaug T 1982 Inexact newton methods *SIAM J. Numer. Anal.* **19** 400–8
[22] Zhan J, Gurung S and Phani Krishna Parsa S 2017 Identification of top-K nodes in large networks using Katz centrality *J. Big Data* **4** 16

[23] Dimitris Achlioptas 2003 Database-friendly random projections: Johnson-lindenstrauss with binary coins *J. Comput. Syst. Sci.* **66** 671–87
[24] Ailon N and Chazelle. B 2009 The fast Johnson–Lindenstrauss transform and approximate nearest neighbors *SIAM J. Comput.* **39** 302–22
[25] Johnson W and Lindenstrauss J 1984 Extensions of lipschitz maps into a hilbert space *Contemp. Math.* **26** 189–206
[26] Drineas P and Mahoney M W 2016 Randnla: randomized numerical linear algebra *Commun. ACM* **59** 80–90
[27] Martinsson P-G and Tropp J A 2020 Randomized numerical linear algebra: foundations and algorithms *Acta Numer.* **29** 403–572
[28] Indyk P and Motwani R 1998 Approximate nearest neighbors: towards removing the curse of dimensionality *STOC '98: Proc. of the thirtieth annual ACM symp. on Theory of computing (Dallas, Texas, USA, May 24 - 26, 1998)* pp 604–13
[29] Gupta S and Gupta A 2002 An elementary proof of a theorem of Johnson and Lindenstrauss *Random Struct. Algor.* **22** 1–65
[30] Blocki J, Blum A, Datta A and Sheffet O 2012 The Johnson-Lindenstrauss transform itself preserves differential privacy *2013 IEEE 54th Annual Symp. on Foundations of Computer Science* (IEEE Computer Society) pp 410–9
[31] Dasgupta S and Gupta A 2003 An elementary proof of a theorem of Johnson and Lindenstrauss *Random Struct. Algorithms* **22** 60–65
[32] Kalender W A 2006 X-ray computed tomography *Phys. Med. Biol.* **51** R29
[33] Cantatore A and Müller P 2011 Introduction to computed tomography *Kgs. Lyngby* (*DTU Mechanical Engineering*)
[34] Pleszczyński M 2021 Implementation of the computer tomography parallel algorithms with the incomplete set of data *PeerJ Comput. Sci.* **7** e339
[35] Radon J 1986 On the determination of functions from their integrals along certain manifolds *IEEE Trans. Med. Imaging* **5** 170 –6
[36] Buhmann M D 2003 *Radial Basis Functions: Theory and Implementations* vol 12 (Cambridge University Press)