Linearly Convergent Randomized Iterative Methods for Computing the Pseudoinverse

Robert Gower, Peter Richtárik

Linearly Convergent Randomized Iterative Methods for Computing the Pseudoinverse

Robert M. Gower* Peter Richtárik†

January 6, 2017

Abstract

We develop the first stochastic incremental method for calculating the Moore-Penrose pseudoinverse of a real matrix. By leveraging three alternative characterizations of pseudoinverse matrices, we design three methods for calculating the pseudoinverse: two general purpose methods and one specialized to symmetric matrices. The two general purpose methods are proven to converge linearly to the pseudoinverse of any given matrix. For calculating the pseudoinverse of full rank matrices we present two additional specialized methods which enjoy a faster convergence rate than the general purpose methods. We also indicate how to develop randomized methods for calculating approximate range space projections, a much needed tool in inexact Newton type methods or quadratic solvers when linear constraints are present. Finally, we present numerical experiments of our general purpose methods for calculating pseudoinverses and show that our methods greatly outperform the Newton-Schulz method on large dimensional matrices.

MSC classes: 15A09, 15B52, 15A24, 65F10, 65F08, 68W20, 65Y20, 65F20, 68Q25, 68W40, 90C20

ACM class: G.1.3

1 Introduction

Calculating the pseudoinverse matrix is a basic numerical linear algebra tool required throughout scientific computing; for example in neural networks [28], signal processing [26, 12] and image denoising [1]. Perhaps the most important application of approximate pseudoinverse matrices is in

*Inria - ENS, robert.gower@inria.fr

†School of Mathematics, University of Edinburgh. This author would like to acknowledge support from the EPSRC Grant EP/K02325X/1, *Accelerated Coordinate Descent Methods for Big Data Optimization* and the EPSRC Fellowship EP/N005538/1, *Randomized Algorithms for Extreme Convex Optimization.*

preconditioning; for instance, within the approximate inverse preconditioning[1] techniques [21, 15, 8, 4].

Currently, the pseudoinverse matrix is calculated using either the singular value decomposition or, when the dimensions of the matrix are large, using a Newton-Schulz type method [2, 3, 24]. However, neither of these aforementioned methods were designed with big data problems in mind, and when applied to matrices from machine learning, signal processing and image analysis, these classical methods can fail or take too much time.

In this paper we develop new fast stochastic incremental methods for calculating the pseudoinverse, capable of calculating an approximate pseudoinverse of truly large dimensional matrices. The problem of determining the pseudoinverse from stochastic measurements also serves as a model problem for determining an approximation to a high dimensional object from a few low dimensional measurements.

The new stochastic methods we present are part of a growing class of "sketch-and-project" methods [16], which have successfully been used for solving linear systems [19, 18], the distributed average consensus problem [22, 18] and inverting matrices [20].

We also envision that the new methods presented here for calculating pseudoinverse matrices will lead to the development of new quasi-Newton methods, much like the development of new randomized methods for inverting matrices [20] has lead to the development of new stochastic quasi-Newton methods [17].

1.1 The Moore-Penrose Pseudoinverse

The pseudoinverse of a real rectangular matrix $A \in \mathbb{R}^{m\times n}$ was first defined as the unique matrix $A^\dagger \in \mathbb{R}^{n\times m}$ that satisfies four particular matrix equations [25, 23]. However, for our purposes it will be more convenient to recall a definition using the singular value decomposition (SVD). Let $A = U\Sigma V^\top$ be the SVD of A, where $U \in \mathbb{R}^{m\times m}$ and $V \in \mathbb{R}^{n\times n}$ are orthogonal matrices and $\Sigma \in \mathbb{R}^{m\times n}$ is a diagonal matrix. The pseudoinverse $A^\dagger \in \mathbb{R}^{n\times m}$ is defined as $A^\dagger = V\Sigma^\dagger U^\top$, where $\Sigma^\dagger$ is the diagonal matrix with $\Sigma^\dagger_{ii} = 1/\Sigma_{ii}$ if $\Sigma_{ii} \neq 0$ and $\Sigma^\dagger_{ii} = 0$ otherwise. This immediately gives rise to a method for calculating the pseudoinverse via the SVD decomposition which costs $O(m^2n)$ floating point operations. When m and n are both large, this can be exacerbating and also unnecessary if one only needs a rough approximation of the pseudoinverse. Therefore, in this work we take a different approach.

In particular, it turns out that the pseudoinverse can alternatively be defined as the least-Frobenius-norm solution to any one of the three equations given in Lemma 1.

[1] A more accurate name for these techniques would be "approximate *pseudo*inverse preconditioning". This is because they form a preconditioner by approximately solving $\min_X \|AX - I\|$, with A not always guaranteed to be nonsingular, leading to the solution being the pseudoinverse.

Lemma 1 *The pseudoinverse matrix $A^\dagger$ is the least-Frobenius-norm solution of any of the three equations:*

$$(P1) \quad AXA = A, \qquad (P2) \quad A^\top = XAA^\top, \qquad \text{and} \qquad (P3) \quad A^\top = A^\top AX.$$

We use the three variational characterizations of the pseudoinverse given in Lemma 1 to design three different stochastic iterative methods for calculating the pseudoinverse. Based on (P2) and (P3), we propose two methods for calculating the pseudoinverse of any real matrix in Section 2. We exploit the symmetry in (P1) to propose a new randomized method for calculating the pseudoinverse of a symmetric matrix in Section 3.

In the next lemma we collect several basic properties of the pseudoinverse which we shall use often throughout the paper.

Lemma 2 *Any matrix $A \in \mathbb{R}^{m\times n}$ and its pseudoinverse $A^\dagger \in \mathbb{R}^{n\times m}$ satisfy the following identities:*

$$(1) \qquad (A^\dagger)^\top = (A^\top)^\dagger$$

$$(2) \qquad (AB)^\dagger = B^\dagger A^\dagger$$

$$(3) \qquad A^\top = A^\dagger AA^\top$$

$$(4) \qquad A^\top = A^\top AA^\dagger$$

$$(5) \qquad \mathbf{Null}\left(A^\dagger\right) = \mathbf{Null}\left(A^\top\right)$$

Note that in the identities above the pseudoinverse *behaves* like the inverse would, were it to exist. Because of (1), we will use $A^{\dagger\top}$ to denote $(A^\dagger)^\top$ or $(A^\top)^\dagger$. Lemma 2 is a direct consequence of the definition of the pseudoinverse; see [25, 23] for a proof based on the classical definition and [11] for a proof based on a definition of the pseudoinverse through projections (all of which are equivalent approaches).

1.2 Notation

We denote the Frobenius inner product and norm by

$$\langle X, Y\rangle \overset{\text{def}}{=} \mathbf{Tr}\left(X^\top Y\right) \qquad \text{and} \qquad \|X\| = \sqrt{\mathbf{Tr}\left(X^\top X\right)},$$

where X and Y are any compatible real matrices and $\mathbf{Tr}\left(X\right)$ denotes the trace of X. Since the trace is invariant under cyclic permutations, for matrices X, Y, Z and W of appropriate dimension, we have

$$\langle X, YZW\rangle = \mathbf{Tr}\left(X^\top YZW\right) = \mathbf{Tr}\left((WX^\top Y)Z\right) = \left\langle Y^\top XW^\top, Z\right\rangle. \tag{6}$$

By $\mathbf{Null}\left(A\right)$ and $\mathbf{Range}\left(A\right)$ we denote the null space and range space of A, respectively. For a positive semidefinite matrix G, let $\lambda^+_{\min}(G)$ denote the smallest nonzero eigenvalue of G.

2 Sketch-and-Project Methods Based on (P3) and (P2)

In view of property (P3) of Lemma 1, the pseudoinverse can be characterized as the solution to the constrained optimization problem

$$A^\dagger \stackrel{\text{def}}{=} \arg\min \frac{1}{2}\|X\|^2, \quad \text{subject to} \quad A^\top = A^\top AX. \tag{7}$$

We shall prove in Theorem 3 that the above variational characterization has the following equivalent dual formulation

$$A^\dagger = \arg_X \min_{X,\Gamma} \frac{1}{2}\|X - A^\dagger\|^2, \quad \text{subject to} \quad X = A^\top A\Gamma. \tag{8}$$

The dual formulation (8) appears to be rather impractical, since using (8) to calculate $A^\dagger$ requires projecting the unknown matrix $A^\dagger$ onto a particular affine matrix space. But duality reveals that (8) can be calculated by solving the primal formulation (7), which does not require knowing the pseudoinverse a priori. The dual formulation reveals that we should not search for $A^\dagger$ within the whole space $\mathbb{R}^{n\times m}$ but rather, the pseudoinverse is contained in the matrix space which forms the constraint in (8).

In the next section we build upon the characterization (7) to develop a new stochastic method for calculating the pseudoinverse.

2.1 The method

Starting from an iterate $X_k \in \mathbb{R}^{n\times m}$, we calculate the next iterate $X_{k+1} \in \mathbb{R}^{n\times m}$ by drawing a random matrix $S \in \mathbb{R}^{m\times\tau}$ from a fixed distribution $\mathcal{D}$ (we do not pose any restrictions on τ) and projecting X_k onto the sketch of (P3):

$$X_{k+1} \stackrel{\text{def}}{=} \arg\min \frac{1}{2}\|X - X_k\|^2, \quad \text{subject to} \quad S^\top A^\top = S^\top A^\top AX. \tag{9}$$

The dual formulation of (9) is given by

$$X_{k+1} \stackrel{\text{def}}{=} \arg_X \min_{X,\Gamma} \frac{1}{2}\|X - A^\dagger\|^2, \quad \text{subject to} \quad X = X_k + A^\top AS\Gamma. \tag{10}$$

The duality of these two formulations is established in the following theorem, along with an explicit solution to (9) that will be used to devise efficient implementations.

Theorem 3 *Solving* (7) *and* (9) *is equivalent to solving* (8) *and* (10)*, respectively. Furthermore, the explicit solution to* (9) *is given by*

$$\boxed{X_{k+1} = X_k - A^\top AS(S^\top A^\top AA^\top AS)^\dagger S^\top A^\top (AX_k - I)}. \tag{11}$$

Proof: We will first show, using Lagrangian duality, that (9) and (10) are equivalent. The Lagrangian of (9) is given by

$$\begin{aligned} L(X,\Gamma) &= \frac{1}{2}\|X-X_k\|^2 + \left\langle \Gamma, S^\top A^\top AX - S^\top A^\top \right\rangle \\ &= \frac{1}{2}\|X-X_k\|^2 + \left\langle A^\top AS\Gamma, X\right\rangle - \left\langle \Gamma, S^\top A^\top\right\rangle. \end{aligned} \tag{12}$$

Since (9) is a convex optimization problem, strong duality implies that

$$(8) \quad = \quad \max_{\Gamma}\min_{X} L(X,\Gamma) = \min_{X}\max_{\Gamma} L(X,\Gamma) = (7).$$

Differentiating (12) in X and setting to zero gives

$$X = X_k - A^\top AS\Gamma. \tag{13}$$

Left multiplying by $S^\top A^\top A$ and observing the constraint in (9) gives $S^\top A^\top = S^\top A^\top AX_k - S^\top A^\top AA^\top AS\Gamma$. Thus,

$$\mathbf{Range}\,(\Gamma) \subset (S^\top A^\top AA^\top AS)^\dagger S^\top A^\top (AX_k - I) + \mathbf{Null}\left(S^\top A^\top AA^\top AS\right). \tag{14}$$

From Lemma 13 with $G = AA^\top$ and $W = AS$ we have that $\mathbf{Null}\left(S^\top A^\top AA^\top AS\right) = \mathbf{Null}\left(A^\top AS\right)$. Consequently, left multiplying (14) by $A^\top AS$ gives

$$A^\top AS\Gamma \;=\; A^\top AS(S^\top A^\top AA^\top AS)^\dagger S^\top A^\top (AX_k - I). \tag{15}$$

The above combined with (13) gives (11).

To derive the dual of (9), first substitute (13) into (12)

$$\begin{aligned} L(X,\Gamma) &= \frac{1}{2}\|X-X_k\|^2 - \langle X - X_k, X\rangle - \left\langle \Gamma, S^\top A^\top AA^\dagger\right\rangle \\ &= -\frac{1}{2}\|X-X_k\|^2 - \langle X-X_k, X_k\rangle - \left\langle A^\top AS\Gamma, A^\dagger\right\rangle \\ &= -\frac{1}{2}\|X-X_k\|^2 + \left\langle X-X_k, A^\dagger - X_k\right\rangle \pm \frac{1}{2}\|A^\dagger - X_k\|^2 \\ &= -\frac{1}{2}\|X - X_k - (A^\dagger - X_k)\|^2 + \frac{1}{2}\|A^\dagger - X_k\|^2. \end{aligned} \tag{16}$$

Calculating the argument that maximizes the above, subject to the constraint (13), is equivalent to solving (10). Thus (9) and (10) are dual to one another and consequently equivalent.

Finally, to see that (8) is the dual of (7), note that by substituting $X_k = 0$ and $S = I$ into (9) and (10) gives (7) and (8), respectively. Furthermore, when $S = I$ in (11) we have that

$$\begin{aligned} X_{k+1} &\overset{(11)+\text{Lemma 2}}{=} X_k - A^\top A(A^\top AA^\top A)^\dagger A^\top A(X_k - A^\dagger) \\ &\overset{P1}{=} X_k - (X_k - A^\dagger) = A^\dagger, \end{aligned}$$

consequently the pseudoinverse is indeed the solution to (7) and (8). ∎

The bottleneck in computing (11) is performing the matrix-matrix product $S^\top A$, which costs $O(\tau mn)$ arithmetic operations. Since we allow τ to be any positive integer, even $\tau = 1$, the iterations (11) can be very cheap to compute. Furthermore, method (9) converges linearly (in $L2$) under very weak assumptions on the distribution $\mathcal{D}$, as we show in the next section.

2.2 Convergence

Since the iterates (11) are defined by a projection process, as we shall see, proving convergence is rather straightforward. Indeed, we will now prove that the iterates (11) converge in $L2$ to the pseudoinverse; that is, the expected norm of $X_k - A^\dagger$ converges to zero. Furthermore, we have a precise expression for the rate at which the iterates converge.

The proofs of convergence of all our methods follow the same machinery. We first start by proving an invariance property of the iterates; namely, that all the iterates reside in a particular affine matrix subspace. We then show that $X_k - A^\dagger$ converges to zero within the said matrix subspace.

Lemma 4 *If* $\mathbf{Range}\left(X_0\right) \subset \mathbf{Range}\left(A^\top A\right)$, *then the iterates* (11) *are such that* $\mathbf{Range}\left(X_k - A^\dagger\right) \subset \mathbf{Range}\left(A^\top A\right)$ *for all* k.

Proof: Using induction and the constraint in (10) we have that $\mathbf{Range}\left(X_k\right) \subset \mathbf{Range}\left(A^\top A\right)$ for all k. The result now follows from $\mathbf{Range}\left(A^\dagger\right) \subset \mathbf{Range}\left(A^\top A\right)$ as can be seen in (8). ∎

Theorem 5 *Let* $X_0 \in \mathbb{R}^{n\times n}$ *with* $\mathbf{Range}\left(X_0\right) \subset \mathbf{Range}\left(A^\top A\right)$ *and let* $H_S \overset{def}{=} S(S^\top A^\top A A^\top A S)^\dagger S^\top$. *The expected iterates* (11) *evolve according to*

$$\mathbf{E}\left[X_{k+1} - A^\dagger\right] = \mathbf{E}\left[I - A^\top A H_S A^\top A\right]\mathbf{E}\left[X_k - A^\dagger\right]. \tag{17}$$

Furthermore, if $\mathbf{E}\left[H_S\right]$ *is finite and positive definite then*

$$\mathbf{E}\left[\|X_k - A^\dagger\|^2\right] \leq \rho^k \|X_0 - A^\dagger\|^2, \tag{18}$$

where

$$\rho = 1 - \lambda_{\min}^+\left(A^\top A \mathbf{E}\left[H_S\right] A^\top A\right). \tag{19}$$

Proof: Let $R_k \overset{\text{def}}{=} X_k - A^\dagger$ and $Z \overset{\text{def}}{=} A^\top A S(S^\top (A^\top A)^2 S)^\dagger S^\top A^\top A = A^\top A H_S A^\top A$. Subtracting $A^\dagger$ from both sides of (11) we have

$$R_{k+1} = (I - Z)R_k. \tag{20}$$

Taking expectation conditioned on R_k gives $\mathbf{E}\left[R_{k+1} \,|\, R_k\right] = (I - \mathbf{E}\left[Z\right])R_k$. Taking expectation again gives (17). Using the properties of pseudoinverse, it is easy to show that Z is a projection matrix and thus $I - Z$ is also a projection matrix[2]. Taking norm squared and then expectation conditioned on R_k in (20) gives

$$\begin{aligned}
\mathbf{E}\left[\|R_{k+1}\|^2 \,|\, R_k\right] &= \mathbf{E}\left[\langle (I-Z)R_k, (I-Z)R_k\rangle \,|\, R_k\right] \\
&\overset{(I-Z) \text{ is a proj.}}{=} \mathbf{E}\left[\langle (I-Z)R_k, R_k\rangle \,|\, R_k\right] \\
&= \|R_k\|^2 - \langle \mathbf{E}\left[Z\right] R_k, R_k\rangle. \qquad (21)
\end{aligned}$$

[2]See Lemma 2.2 in [19] for an analogous proof

From Lemma 4 we have that there exists W_k such that $R_k = A^\top A W_k$. Therefore,

$$\begin{aligned}\langle \mathbf{E}\left[Z\right] R_k, R_k\rangle &\overset{\text{Lemma 4}}{=} \left\langle A^\top A \mathbf{E}\left[Z\right] A^\top A W_k, W_k\right\rangle \\ &= \left\langle (A^\top A)(A^\top A) H_S (A^\top A)(A^\top A) W_k, W_k\right\rangle \\ &\overset{\text{Lemma 15}}{\geq} \lambda_{\min}^+\left(A^\top A H_S A^\top A\right)\left\langle (A^\top A) W_k, (A^\top A) W_k\right\rangle \\ &= \lambda_{\min}^+\left(A^\top A H_S A^\top A\right)\|R_k\|^2 = (1-\rho)\|R_k\|^2. \end{aligned} \tag{22}$$

Taking expectation in (21) we have

$$\mathbf{E}\left[\|R_{k+1}\|^2\right] = \mathbf{E}\left[\|R_k\|^2\right] - \mathbf{E}\left[\langle \mathbf{E}\left[Z\right] R_k, R_k\rangle\right] \overset{(22)}{\leq} \rho\, \mathbf{E}\left[\|R_k\|^2\right].$$

It remains now to unroll the above recurrence to arrive at (18). ∎

With a precise expression for the convergence rate (19) opens up the possibility of tuning the distribution of S so that the resulting has a faster convergence. Next we give an instantiation of the method (11) and indicate how one can choose the distribution of S to accelerate the method. We refer to methods based on (11) as the `SATAX` methods, inspired on the constraint in (9) whose right hand side almost spells out `SATAX`. Later in Section 5 we perform experiments on variants of the `SATAX` method.

2.3 Discrete examples

Though our framework and Theorem 3 allows for $\mathcal{D}$ to be a continuous distribution, for illustration purposes here we focus our attention on developing examples where $\mathcal{D}$ is a discrete distribution.

For any discrete distribution $\mathcal{D}$ the random matrix $S \sim \mathcal{D}$ will have a finite number of possible outcomes. Fix r as the number of outcomes and let $\tau \in \mathbb{N}$ and $S = S_i \in \mathbb{R}^{m\times\tau}$ with probability $p_i > 0$ for $i = 1, \ldots, r$. Let $\mathbb{S} \overset{\text{def}}{=} [S_1, \ldots, S_r] \in \mathbb{R}^{m\times r\tau}$. If

$$p_i = \frac{\mathbf{Tr}\left(S_i^\top (A^\top A)^2 S_i\right)}{\mathbf{Tr}\left(\mathbb{S}^\top (A^\top A)^2 \mathbb{S}\right)},$$

then by Lemma 17 with $G = A^\top A$ as proven in the Appendix, the rate of convergence in Theorem 3 is given by

$$\rho = 1 - \lambda_{\min}^+(\mathbf{E}\left[Z\right]) \leq 1 - \frac{\lambda_{\min}^+(\mathbb{S}^\top (A^\top A)^2 \mathbb{S})}{\mathbf{Tr}\left(\mathbb{S}^\top (A^\top A)^2 \mathbb{S}\right)}. \tag{23}$$

The number $\lambda_{\min}^+(\mathbb{S}^\top (A^\top A)^2 \mathbb{S}) / \mathbf{Tr}\left(\mathbb{S}^\top (A^\top A)^2 \mathbb{S}\right)$ is known as the scaled condition number of $\mathbb{S}^\top (A^\top A)^2 \mathbb{S}$ and it is the same condition number on which the rate of convergence of the randomized Kaczmarz method depends [27]. This rate (23) suggests that we should choose $\mathbb{S}$ so that $\mathbb{S}^\top (A^\top A)^2 \mathbb{S}$ has a concentrated spectrum and consequently, the scaled condition number is minimized. Ideally, we would want $\mathbb{S} = (A^\top A)^\dagger$, but this is not possible in practice, though it does inspire the following heuristic choice. If we choose $\mathbb{S} = X_k X_k^\top$, then as $X_k \to A^\dagger$ we have that

$\mathbb{S} \to A^\dagger A^{\dagger\top} = (A^\top A)^\dagger$. Though through experiments we have identified that choosing the sketch matrix so that $\mathbb{S} = X_k$ resulted in the best performance. This observation, together with other empirical observations, has lead us to suggest two alternative sketching strategies:

Uniform τ–batch sampling: We say S is a uniform τ–batch sampling if $\mathbf{P}(S = I_{:C}) = 1/\binom{n}{|C|}$ where $C \in \{1, \ldots, n\}$ is a random subset with $|C| = \tau$ chosen uniformly at random and $I_{:C}$ denotes the column concatenation of the columns of the identity matrix indexed by C.

Adaptive sketching: Fix the iteration count k and consider the current iterate X_k. We say that S is an adaptive sketching if $\mathbf{P}(S = X_k I_{:C})$ where $I_{:C}$ is a uniform τ–batch sampling.

When using a uniform τ–batch sampling together with the `SATAX` method, we refer to the resulting method as the `SATAX_uni`. We use `SATAX_ada` when referring to the method that uses the adaptive sketching. We benchmark both these methods later in Section 5.

2.4 A sketch-and-project method based on (P2)

Yet another characterization of the pseudoinverse is given by the solution to the constrained optimization problem based on property (P2):

$$A^\dagger \overset{\text{def}}{=} \arg\min \frac{1}{2}\|X\|^2, \quad \text{subject to} \quad A^\top = XAA^\top. \tag{24}$$

which has the following equivalent dual formulation

$$A^\dagger \overset{\text{def}}{=} \arg_X \min_{X,\Gamma} \frac{1}{2}\|X - A^\dagger\|^2, \quad \text{subject to} \quad X = \Gamma^\top AA^\top. \tag{25}$$

Transposing the constraint in (24) gives $A = AA^\top X^\top$. Since the Frobenius norm is invariant to transposing the argument, we have that by setting $Y = X^\top$ in (24) we get

$$A^{\dagger\top} \overset{\text{def}}{=} \arg\min \frac{1}{2}\|Y\|^2, \quad \text{subject to} \quad A = AA^\top Y. \tag{26}$$

It is now clear to see that (26) is equivalent to (7) where each occurrence of A has been swapped for $A^\top$. Because of this simple mapping from (26) to (7) we refrain from developing methods based on (26) (even though these methods are different).

3 A Sketch-and-Project Method Based on (P1)

Now we turn out attention to designing a method based on (P1). In contrast with the development in the previous section, here we make explicit use of the symmetry present in (P1). In particular, we introduce a novel sketching technique which we call *symmetric sketch*. As we shall see, if A is symmetric, our method (31) maintains the symmetry of iterates if started from a symmetric matrix X_0. Throughout this section we assume that $A \in \mathbb{R}^{n\times n}$ is a symmetric matrix.

The final variational characterization of the pseudoinverse from Lemma 1, based on (P1), is

$$A^\dagger = \arg\min \frac{1}{2}\|X\|^2, \quad \text{subject to} \quad AXA = A. \tag{27}$$

As before, we have the following equivalent dual formulation

$$A^\dagger = \arg_X \min_{\Gamma,X} \frac{1}{2}\|A^\dagger - X\|^2 \quad \text{subject to} \quad X = A\Gamma A. \tag{28}$$

In Section 3.1 we describe our method. In Theorem 6 we prove that these two formulations are equivalent and also show that the iterates of our method are symmetric. This is in contrast with techniques such as the block BFGS update and other methods designed for calculating the inverse of a matrix in [20], where symmetry has to be imposed on the iterates through an explicit constraint.

Calculating approximations of the pseudoinverse of a symmetric matrix is particularly relevant when designing variable metric methods in optimization, where one wishes to maintain an approximate of the (pseudo)inverse of the Hessian matrix. In contrast to the symmetric methods for calculating the inverse presented in [20], which can be readily interpreted as extensions of known quasi-Newton methods, the method presented in this section appears not to be related to any Broyden quasi-Newton method [5], nor the SR1 update. This naturally leads to the question: how would a quasi-Newton method based on (31) fair? We leave this question to future research.

3.1 The method

Similarly to the methods developed in Section 2, we define an iterative method by projecting onto a sketch of (27). In this case, however, we use the *symmetric sketch*. Specifically, we calculate the next iterate X_{k+1} via

$$X_{k+1} \overset{\text{def}}{=} \arg\min \frac{1}{2}\|X - X_k\|^2, \quad \text{subject to} \quad S^\top AXAS = S^\top AS, \tag{29}$$

where $S \in \mathbb{R}^{m\times\tau}$ is drawn from $\mathcal{D}$. The dual formulation is given by

$$X_{k+1} = \arg_X \min_{\Gamma,X} \frac{1}{2}\|X - A^\dagger\|^2 \quad \text{subject to} \quad X = X_k + AS\Gamma S^\top A. \tag{30}$$

This symmetric sketch makes its debut in this work, since it has not been used in any of the previous works developing sketch-and-project sketching methods [19, 18, 20].

Theorem 6 *Solving* (27) *and* (29) *is equivalent to solving* (28) *and* (30)*, respectively. Furthermore, the explicit solution to* (29) *is*

$$\boxed{X_{k+1} = X_k + AS(S^\top A^2 S)^\dagger S^\top (A - AX_kA)S(S^\top A^2 S)^\dagger S^\top A.} \tag{31}$$

Proof: Let

$$E \stackrel{\text{def}}{=} X - X_k \quad \text{and} \quad B \stackrel{\text{def}}{=} S^\top (A - AX_kA)S. \tag{32}$$

Using the above renaming we have that (29) is equivalent to solving

$$\min \frac{1}{2}\|E\|^2, \quad \text{subject to} \quad S^\top AEAS = B. \tag{33}$$

The Lagrangian of (33) is given by

$$\begin{aligned} L(E,\Gamma) &= \frac{1}{2}\|E\|^2 + \left\langle \Gamma, S^\top AEAS - B \right\rangle \\ &\stackrel{(6)}{=} \frac{1}{2}\|E\|^2 + \left\langle AS\Gamma S^\top A, E \right\rangle - \langle \Gamma, B \rangle . \end{aligned} \tag{34}$$

Differentiating in E and setting the derivative to zero gives

$$E = -AS\Gamma S^\top A. \tag{35}$$

Left and right multiplying by $S^\top A$ and AS, respectively, and using the constraint in (33) gives

$$B = -(S^\top A^2 S)\Gamma S^\top A^2 S. \tag{36}$$

The equation (36) is equivalent to solving in Γ the following system

$$(S^\top A^2 S)Y = -B \tag{37}$$

$$(S^\top A^2 S)\Gamma^\top = Y^\top. \tag{38}$$

The solution to (37) is given by any Y such that

$$\begin{aligned} \mathbf{Range}\,(Y) &\subset -(S^\top A^2 S)^\dagger B + \mathbf{Null}\left(S^\top A^2 S\right) \\ &\stackrel{\text{Lemma 13}}{\subset} -(S^\top A^2 S)^\dagger B + \mathbf{Null}\,(AS)\,, \end{aligned} \tag{39}$$

where we applied Lemma 13 with $G = I$ and $W = AS$. The solution to (38) is given by any Γ that satisfies

$$\begin{aligned} \mathbf{Range}\left(\Gamma^\top\right) &\subset (S^\top A^2 S)^\dagger Y^\top + \mathbf{Null}\left(S^\top A^2 S\right) \\ &\stackrel{\text{Lemma 13}}{\subset} (S^\top A^2 S)^\dagger Y^\top + \mathbf{Null}\,(AS)\,. \end{aligned} \tag{40}$$

Transposing the above, substituting (39), left and right multiplying by AS and $S^\top A$ respectively gives

$$\begin{aligned} \mathbf{Range}\left(AS\Gamma S^\top A\right) &\subset AS\left(\mathbf{Null}\,(AS) - (S^\top A^2 S)^\dagger B\right)(S^\top A^2 S)^\dagger S^\top A \\ &\quad + AS\mathbf{Null}\,(AS)^\top S^\top A \end{aligned} \tag{41}$$

$$= -AS(S^\top A^2 S)^\dagger B(S^\top A^2 S)^\dagger S^\top A, \tag{42}$$

where in the last step we used the fact that $(AS\mathbf{Null}\,(AS)^\top S^\top A)^\top = AS\mathbf{Null}\,(AS)\,S^\top A = 0$. Inserting (42) into (35) gives $E = AS(S^\top A^2 S)^\dagger B(S^\top A^2 S)^\dagger S^\top A$. Substituting in the definition of E and B we have (31).

For the dual problem, using (35) and substituting (46) into (34) gives

$$\begin{aligned} L(E,\Gamma) &= \frac{1}{2}\|E\|^2 - \|E\|^2 - \left\langle \Gamma, S^\top A(A^\dagger - X_k)AS \right\rangle \\ &= -\frac{1}{2}\|E\|^2 - \left\langle AS\Gamma S^\top A, X_k - A^\dagger \right\rangle \\ &\overset{(35)}{=} -\frac{1}{2}\|E\|^2 + \left\langle E, X_k - A^\dagger \right\rangle \pm \frac{1}{2}\|X_k - A^\dagger\|^2 \\ &= -\frac{1}{2}\|E - (A^\dagger - X_k)\|^2 + \frac{1}{2}\|X_k - A^\dagger\|^2. \end{aligned}$$

Substituting $E = X - X_k$, maximizing in Γ and minimizing in X while observing the constraint (35), we arrive at (30).

Furthermore, substituting $X_k = 0$ and $S = I$ in (29) and (30) gives (27) and (28), respectively, thus (27) and (28) are indeed equivalent dual formulations. Finally, substituting $X_k = 0$ and $S = I$ in (31) and using properties P1 and P2, it is not hard to see that (31) is equal to $A^\dagger$, and thus (27) and (28) are indeed alternative characterizations of the pseudoinverse. ∎

One of the insights given by the dual formulation (30) is that the resulting method is monotonic, that is, the error $\|X_{k+1} - A^\dagger\|$ must be a decreasing sequence. Inspired on the constraint in (29), we refer to the class of methods defined by (29) as the `SAXAS` methods.

3.2 Convergence

Proving the convergence of the iterates (31) follows the same machinery as the convergence proof in Section 2.2. But different from Section 2.2 the resulting convergence rate ρ may be equal to one $\rho = 1$. We determine discrete distributions for S such that $\rho < 1$ in Section 3.3.

The first step of proving convergence is the following invariance result.

Lemma 7 *If $X_0 = AWA$ for some matrix W, then for each $k \geq 0$ there exists matrix $Q_k \in \mathbb{R}^{n\times m}$ such that the iterates* (31) *satisfy $X_k - A^\dagger = AQ_kA$.*

Proof: Using induction and the constraint in (30) we have that $X_{k+1} = AW_{k+1}A$ where $W_{k+1} = W_k + S\Gamma S^\top$. Furthermore, from the constraint in (30), we have that there exists Γ such that $A^\dagger = A\Gamma A$. Thus $X_{k+1} - A^\dagger = AQ_{k+1}A$ with $Q_{k+1} = W_{k+1} - \Gamma$. ∎

Theorem 8 *If $X_0 = AWA$ then the iterates* (11) *converge according to*

$$\mathbf{E}\left[\|X_k - A^\dagger\|^2\right] \quad \leq \quad \rho^k \|X_0 - A^\dagger\|^2, \tag{43}$$

where

$$\rho \quad \overset{def}{=} \quad 1 - \inf_{\substack{R = AQA, Q \in \mathbb{R}^{n\times n} \\ \|R\|^2 = 1}} \langle \mathbf{E}\left[ZRZ\right], R\rangle, \tag{44}$$

and

$$Z \overset{def}{=} AS(S^\top A^2 S)^\dagger S^\top A. \tag{45}$$

Proof: Let $R_k = X_k - A^\dagger$. Using

$$S^\top(A - AX_kA)S \overset{(P1)}{=} S^\top A(A^\dagger - X_k)AS = S^\top AR_kAS, \tag{46}$$

and subtracting $A^\dagger$ from both sides of (31) gives

$$R_{k+1} = R_k - ZR_kZ. \tag{47}$$

Applying the properties of the pseudoinverse, it can be shown that Z is a projection matrix, whence $Z^2 = Z$. Taking norms and expectation conditioned on R_k on both sides gives

$$\begin{aligned} \mathbf{E}\left[\|R_{k+1}\|^2 \mid R_k\right] &= \mathbf{E}\left[\langle R_k - ZR_kZ, R_k - ZR_kZ\rangle \mid R_k\right] \\ &\overset{Z \text{ is a proj.}}{=} \|R_k\|^2 - \langle \mathbf{E}\left[ZR_kZ\right], R_k\rangle. \end{aligned} \tag{48}$$

By Lemma 7 and (48) we have that

$$\begin{aligned} \mathbf{E}\left[\|R_{k+1}\|^2 \mid R_k\right] &\overset{(48)}{=} \|R_k\|^2 - \left\langle \mathbf{E}\left[Z\frac{R_k}{\|R_k\|}Z\right], \frac{R_k}{\|R_k\|}\right\rangle \|R_k\|^2. \\ &\overset{\text{Lemma 7+(44)}}{\leq} \rho\|R_k\|^2. \end{aligned} \tag{49}$$

It remains to take expectations again, apply the tower property, and unroll the recurrence. ∎

The method described in (29) is particularly well suited to calculating an approximation to the pseudoinverse of symmetric matrices, since symmetry is preserved by the method.

Lemma 9 (Symmetry invariance) *If $X_0 = X_0^\top$ and $A = A^\top$ then the iterates* (31) *are symmetric.*

Proof: The constraint in (30) and induction shows that $X_k = X_k^\top$ holds for any k. ∎

3.3 The rate of convergence

It is not immediately obvious that (19) is a valid rate. That is, is it the case that $0 \leq \rho \leq 1$? We give an affirmative answer to this in Lemma 11. Subsequently, in Lemma 12 we establish necessary and sufficient conditions on discrete distribution $\mathcal{D}$ to characterize when $\rho < 1$. Consequently, under these conditions a linear convergence rate is guaranteed.

To establish the next results we make use of vectorization and the Kronecker product so that we can leverage on classic results in linear algebra. For convenience, we state several well known properties and equalities involving Kronecker products in the following lemma. But first, the Kronecker product of matrices $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{p\times q}$ is defined as

$$(A \otimes B)_{p(r-1)+i,q(s-1)+j} = a_{rs}b_{ij}. \tag{50}$$

Let $\vec{A} \in \mathbb{R}^{nm}$ denote the vector obtained by stacking the columns of the matrix A on top of one another.

Lemma 10 (Properties Kronecher products) *For matrices A, B and C of compatible dimensions we have that*

1. $\overrightarrow{ABC} = (C^\top \otimes A)\vec{B}$,
2. $(C \otimes B)^\top = C^\top \otimes B^\top$.
3. *If A and B are symmetric positive semidefinite then $B \otimes A$ is symmetric positive semidefinite.*
4. *Since both vectorization and expectation are linear operators, if Z is a random matrix then $\overrightarrow{\mathbf{E}\left[Z\right]} = \mathbf{E}\left[\vec{Z}\right]$.*

Lemma 11 *The rate* (44) *satisfies $0 \leq \rho \leq 1$. Furthermore, if*

$$\{R : \mathbf{E}\left[ZRZ\right] = 0\} \subset \{R : ARA = 0\}, \tag{51}$$

then

$$\rho \leq 1 - \lambda_{\min}^{+}(\mathbf{E}\left[Z \otimes Z\right]) < 1, \tag{52}$$

and the iterates (11) *converge.*

Proof: Since Z is positive semidefinite we have that

$$\langle ZRZ, R\rangle = \mathbf{Tr}\left(R^\top ZRZ\right) = \mathbf{Tr}\left(Z^{1/2}R^\top ZRZ^{1/2}\right) \geq 0.$$

Taking expectation in the above gives that $\rho \leq 1$. Furthermore, since Z is a projection matrix,

$$\begin{aligned}
\langle ZRZ, R\rangle &= \mathbf{Tr}\left(R^\top ZRZ\right) \\
&\leq \mathbf{Tr}\left(R^\top ZR\right)\underbrace{\lambda_{\max}(Z)}_{=1} \\
&= \mathbf{Tr}\left(ZRR^\top\right) \\
&\leq \mathbf{Tr}\left(RR^\top\right)\lambda_{\max}(Z) = \|R\|^2.
\end{aligned}$$

Dividing by $\|R\|^2$ and taking expectation over Z gives

$$\frac{\langle \mathbf{E}\left[ZRZ\right], R\rangle}{\|R\|^2} \leq 1. \tag{53}$$

Thus, for any $R \neq 0$, we have that

$$\rho \overset{(44)}{\geq} 1 - \frac{\langle \mathbf{E}\left[ZRZ\right], R\rangle}{\|R\|^2} \overset{(53)}{\geq} 0,$$

which concludes the proof that $0 \leq \rho \leq 1$.

After vectorizing and using item 1 of Lemma 10, the condition (51) is equivalent to

$$\{\vec{R} \,:\, \mathbf{E}\left[Z \otimes Z\right] \vec{R} = 0\} = \mathbf{Null}\left(\mathbf{E}\left[Z \otimes Z\right]\right) \overset{(51)}{\subset} \{\vec{R} \,:\, (A \otimes A)\vec{R} = 0\} = \mathbf{Null}\left(A \otimes A\right). \tag{54}$$

Since Z is symmetric positive semidefinite, item 3 of Lemma 10 states that the matrix $Z \otimes Z$, and consequently $\mathbf{E}\left[Z \otimes Z\right]$, are symmetric positive semidefinite. Thus taking orthogonal complements in (54) we have

$$\mathbf{Range}\left(A \otimes A\right) \subset \mathbf{Range}\left(\mathbf{E}\left[Z \otimes Z\right]\right). \tag{55}$$

Therefore, using vectorization we have

$$\begin{aligned} \inf_{\substack{R = AQA,\, Q \in \mathbb{R}^{n\times n} \\ \|R\|^2 = 1}} \left\langle \mathbf{E}\left[ZRZ\right], R\right\rangle &= \inf_{\substack{R \in \mathbf{Range}\left(A \otimes A\right) \\ \|R\|^2 = 1}} \left\langle \mathbf{E}\left[Z \otimes Z\right] \vec{R}, \vec{R}\right\rangle_2 \\ &\overset{(55)}{\geq} \inf_{\substack{R \in \mathbf{Range}\left(\mathbf{E}\left[Z \otimes Z\right]\right) \\ \|R\|^2 = 1}} \left\langle \mathbf{E}\left[Z \otimes Z\right] \vec{R}, \vec{R}\right\rangle_2 \\ &\overset{(57)}{=} \lambda_{\min}^{+}(\mathbf{E}\left[Z \otimes Z\right]) > 1, \end{aligned} \tag{56}$$

where we have used that for any G positive semi-definite we have

$$\lambda_{\min}^{+}(G) = \inf_{\substack{x \,\in\, \mathbf{Null}\left(G\right)^{\perp} \\ \|x\|_2 = 1}} \left\langle Gx, x\right\rangle. \tag{57}$$

Combining (56) with (44) gives the desired result (52). ∎

3.3.1 Characterization of $\rho < 1$ for discrete distributions

The following lemma gives a practical characterization of the condition (51) for discrete distributions.

Lemma 12 *Let S be a random matrix with a discrete distribution such that $\mathbb{P}(S = S_i) = p_i > 0$, where $S_i \in \mathbb{R}^{n\times q_i}$ for $i = 1, \ldots, r$. Let*

$$\mathbb{S} \overset{def}{=} \begin{pmatrix} S_1^\top \otimes S_1^\top \\ \vdots \\ S_r^\top \otimes S_r^\top \end{pmatrix} \in \mathbb{R}^{\sum_{i=1}^r q_i^2 \times n}. \tag{58}$$

Then the iterates (31) *converge according to Theorem 6 with a rate $\rho < 1$ if*

$$\bigcap_{i=1}^{r} \{R \,:\, S_i^\top ARAS_i = 0\} \subset \{R \,:\, ARA = 0\}. \tag{59}$$

Equivalently, condition (59) *holds if and only if*

$$\mathbf{Null}\left(\mathbb{S}\left(A \otimes A\right)\right) \subset \mathbf{Null}\left(A \otimes A\right). \tag{60}$$

Proof: We show that (51) and (59) are equivalent, therefore convergence of the iterates (31) with $\rho < 1$ is guaranteed by Lemma 11. First, note once more that $\mathbf{Null}\left(\mathbf{E}\left[Z \otimes Z\right]\right) = \{\vec{R} \,:\, \mathbf{E}\left[ZRZ\right] = 0\}$. Let $Z_i \overset{\text{def}}{=} AS_i(S_i^\top A^2 S_i)^\dagger S_i^\top A$ and note that Z_i is a symmetric positive semidefinite matrix. Using the distribution of S we have that $\vec{R} \in \mathbf{Null}\left(\mathbf{E}\left[Z \otimes Z\right]\right)$ is equivalent to

$$\mathbf{E}\left[ZRZ\right] \;=\; \sum_{i=1}^{r} p_i Z_i R Z_i = 0. \tag{61}$$

Since Z_i is symmetric positive semidefinite by Lemma 10 items 3 and 4 we have that $\mathbf{E}\left[Z_i \otimes Z_i\right]$ is positive definite, consequently

$$\begin{aligned}
\mathbf{Null}\left(\mathbf{E}\left[Z \otimes Z\right]\right) &= \{\vec{R} \,:\, \sum_{i=1}^{r} p_i (Z_i \otimes Z_i)\vec{R} = 0\} \\
&= \{\vec{R} \,:\, (Z_i \otimes Z_i)\vec{R} = 0, \quad \text{for } i = 1, \ldots, r\} \\
&= \bigcap_{i=1}^{r} \mathbf{Null}\left(Z_i \otimes Z_i\right).
\end{aligned} \tag{62}$$

Fix an index $i \in \{1, \ldots, r\}$. The remainder of the proof is now dedicated to showing that $\mathbf{Null}\left(Z_i \otimes Z_i\right) = \{\vec{R} \,:\, S_i^\top ARAS_i = 0\}$. To this end, we collect some facts. Given that

$$\mathbf{Null}\left((S_i^\top A^2 S_i)^\dagger\right) \overset{(5)}{=} \mathbf{Null}\left(S_i^\top A^2 S_i\right) \overset{\text{Lemma 13}}{=} \mathbf{Null}\left(AS_i\right), \quad \text{for } i = 1, \ldots, r,$$

we can apply Lemma 13 once again with $G = (S_i^\top A^2 S_i)^\dagger$ and $W = S_i^\top A$ which shows that

$$\mathbf{Null}\left(Z_i\right) = \mathbf{Null}\left(AS_i(S_i^\top A^2 S_i)^\dagger S_i^\top A\right) \overset{\text{Lemma 13}}{=} \mathbf{Null}\left(S_i^\top A\right). \tag{63}$$

Consequently

$$\mathbf{Null}\left(Z_i \otimes Z_i\right) = \{\vec{R} \,:\, Z_i R Z_i = 0\} \overset{(63)}{=} \{\vec{R} \,:\, S_i^\top ARZ_i = 0\} \overset{(63)}{=} \{\vec{R} \,:\, S_i^\top ARAS_i = 0\}. \tag{64}$$

Finally

$$\mathbf{Null}\left(\mathbf{E}\left[Z \otimes Z\right]\right) \overset{(62)+(64)}{=} \{\vec{R} \,:\, S_i^\top ARAS_i = 0, \quad i = 1, \ldots, r\} = \bigcap_{i=1}^{r} \{\vec{R} \,:\, S_i^\top ARAS_i = 0\},$$

which proves that (51) and (59) are equivalent. Using vectorization, the condition (59) can be rewritten as $\{v \,:\, (S_i^\top \otimes S_i^\top)(A \otimes A)v = 0, \text{ for } i = 1, \ldots, r\} \subset \{v \,:\, (A \otimes A)v = 0\}$, which is clearly equivalent to (60). ∎

Lemma 12 gives us a practical rule for designing a distribution for S such that convergence is guaranteed. Given that $\mathbf{Null}\left(A \otimes A\right)$ is not known to us, the easiest way to ensure that (60) holds is if we choose a distribution for S such that $\mathbb{S}$ has a full column rank. Clearly (59) holds when S is a fixed invertible matrix with probability one, but this does not result in a practical method. In the next section we show how to construct S so that $\mathbb{S}$ has a full column rank and results in a practical method.

3.4 Discrete examples

Based on the two sketching strategies presented in Section 2.3, we define two variants of the `SAXAS` method (31). Let the `SAXAS_uni` and the `SAXAS_ada` methods be the result of using a uniform τ–batch sketching and an adaptive sketching with the `SAXAS` method, respectively. We found that these two variants work well in practice, as we show later on in Section 5. Though we observe in empirical experiments that the two variants of `SAXAS` converge in practice, it is hard to verify Lemma 12 and thus prove convergence. So instead we introduce a new sketching very similar to the uniform τ–batch sketching, but that allows us to easily prove convergence of the resulting method.

τ–batch sketching with replacement. Let $S = I_{:v}$ where $v \in \{1, \dots, n\}^\tau$ is an array and $I_{:v} \in \mathbb{R}^{n \times \tau}$ is the column concatenation of the columns in the identity matrix I indexed by v. Furthermore, let $\mathbf{P}(S = I_{:v}) = p_v > 0$ for each $v \in \{1, \dots, n\}^\tau$.

We refer to the `SAXAS` method with a τ–batch sketching with replacement as the `SAXAS_rep` method. As we will now show, under the condition that $\tau \geq 2$, the `SAXAS_rep` method satisfies Lemma 12 and thus convergence of the `SAXAS_rep` method is guaranteed.

Convergence. We will prove that `SAXAS_rep` method converges by showing that the matrix $\mathbb{S}$ defined in (58) has full column rank, and thus according to Lemma 12 the iterates converge. First note that since the sampling is done over all $v \in \{1, \dots, n\}^\tau$, there are n^τ different sketching matrices. Thus $\mathbb{S} \in \mathbb{R}^{\tau n^\tau \times n^2}$. To prove that $\mathbb{S}$ has full column rank, we will show that for $\tau \geq 2$ that the row rank of $\mathbb{S}$ is n^2. Note that for $\tau = 1$ the matrix $\mathbb{S}$ has n rows, thus it is not possible for $\mathbb{S}$ to have full column rank. For simplicity, consider the case $\tau = 2$. Fix $i \in \{1, \dots, n^2\}$. We will now show that for the ith coordinate vector $e_i \in \mathbb{R}^{n^2}$, there exists $v \in \{1, \dots, n\}^\tau$ such that e_i is a row of $I_{:v}^\top \otimes I_{:v}^\top$, and consequently, e_i is a row of $\mathbb{S}$. First, for $v = (s, j)$ we have from the definition of Kronecker product (50) that

$$(I_{:v}^\top \otimes I_{:v}^\top)_{2,n(s-1)+j} = [I_{:v}]_{1s}[I_{:v}]_{2j} = 1. \tag{65}$$

Moreover, every other element on row 2 of $I_{:v}^\top \otimes I_{:v}^\top$ is zero apart from the element in column $n(s-1)+j$. Now note that the integer i can be written as

$$i = n\left\lfloor \frac{i}{n} \right\rfloor + \mod(i,n) = n\underbrace{\left(\left\lfloor \frac{i}{n} \right\rfloor - 1\right)}_{=s-1} + \underbrace{\mod(i,n) + n}_{=j}.$$

By setting $s = \lfloor \frac{i}{n} \rfloor$ and $j = \mod(i,n) + n$, we have from the above that $n(s-1)+j = i$. Though there is problem when $\lfloor \frac{i}{n} \rfloor = 0$, since s cannot be zero. To remedy this, consider the indices

$$s = \begin{cases} 1 & \text{if } i < n \\ \lfloor \frac{i}{n} \rfloor & \text{if } i \geq n, \end{cases} \qquad \text{and} \qquad j = \begin{cases} i & \text{if } i < n \\ \mathrm{mod}(i,n) + n & \text{if } i \geq n. \end{cases}$$

With $v = (s, j)$ we now have that the 2nd row of the matrix in (65) is the ith unit coordinate vector in $\mathbb{R}^{n^2}$. Consequently $\mathbb{S}$ has row rank n^2 and the `SAXAS_rep` method converges.

4 Projections and Full Rank Matrices

In this section we comment on calculating approximate projections onto the range space of a given matrix, and on certain specifics related to calculating the pseudoinverse of a full rank matrix.

4.1 Calculating approximate range space projections

With very similar methods, we can calculate an approximate projection operator onto the range space of A. Note that $AA^\dagger$ projects onto $\mathbf{Range}(A)$ as can be seen by $(P1)$. But rather than calculate $A^\dagger$ and then left multiply by A, it is more efficient to calculate $AA^\dagger$ directly. For this, let $P \overset{\text{def}}{=} AA^\dagger$ and note that from the identities $AA^\dagger A = A$ and $A^\top AA^\dagger = A^\top$ we have that P satisfies

1. $PA = A$
2. $A^\top P = A^\top$.

We can design a sketch and project method based on either property. For instance, based on item 1 we have the method

$$X_{k+1} \overset{\text{def}}{=} \arg\min \frac{1}{2}\|X - X_k\|^2, \quad \text{subject to} \quad PAS = AS. \tag{66}$$

The advantage of this approach, over calculating $A^\dagger$ separately, is a resulting faster method. Indeed, if we were to carry out the analysis of this method, following analogous steps to the convergence in Section 2.2, and together with a conveniently chosen probability distribution based on Lemma 17, the iterates (66) would converge according to

$$\mathbf{E}\left[\|X_{k+1} - P\|^2\right] = \left(1 - \frac{\lambda^+_{\min}(\mathbb{S}^\top A^\top A\mathbb{S})}{\mathbf{Tr}\left(\mathbb{S}^\top A^\top A\mathbb{S}\right)}\right)\mathbf{E}\left[\|X_k - P\|^2\right]. \tag{67}$$

Since the rate is proportional to a scaled condition number with fewer powers of A as compared to our previous convergence results (18), the method (66) is less sensitive to ill conditioning in the matrix A.

Such a method would be useful in a solving linearly constrained optimization problems [14, 6] which often require projecting the gradient onto the range space of system matrix. In particular, in a iteration of a Newton-CG framework [10, 13], one needs only inexact solutions to a quadratic optimization problem with linear constraints. A method based on (66) can be used to calculate a projection operator to within the precision required by the Newton-CG framework, and thus save on the computational effort of calculating the exact projection matrix.

4.2 Pseudoinverse of full rank matrices

In the special case when A has full rank, there are two alternative sketch-and-project methods that are more effective than our generic method. In particular, when A has full row rank ($m \leq n$) then there exists X such that $AX = I$, furthermore, $AA^\dagger = I$. In this case, we have that

$$A^\dagger = \arg\min \|X\|_F^2, \quad \text{subject to} \quad AX = I. \tag{68}$$

Applying a sketching and projecting strategy to the above gives

$$X^{k+1} = \arg\min \|X - X^k\|_F^2, \quad \text{subject to} \quad S^\top AX = S^\top. \tag{69}$$

This method (69) was presented in [20] as a method for inverting matrices. The analysis in [20] still holds in this situation by using the techniques we presented in Section 3.2. Again, the resulting rate of convergence of the method defined by (69) is less sensitive to ill conditioning in the matrix A, as can be seen in Theorem 6.2 in [20].

Alternatively, when A has full column rank, then $A^\dagger A = I$, and one should apply a sketching and projecting method using the equation $XA = I$.

Consequently the methods `SATAX` (11) and `SAXAS` (31) are better suited for calculating the pseudoinverse of rank deficient matrices, which is the focus of our experiments in the next section.

5 Numerical Experiments

We now perform several numerical experiments comparing two variants of the `SATAX` and the `SAXAS` methods to the Newton-Schulz method

$$X^{k+1} = 2X^k + X^k A X^k, \tag{70}$$

as introduced by Ben-Israel and Cohen [3, 2] for calculating the pseudoinverse matrix. The Newton-Schulz method is guaranteed to converge as long as $\|I - X_0 A\|_2 < 1$. Consequently, we set $X_0 = \frac{1}{2}\frac{A^\top}{\|A\|_F^2}$ for the Newton-Schulz method to guarantee its convergence. Furthermore, the Newton-Schulz method enjoys quadratic local convergence [3, 2], in contrast to the randomized methods which are globally linearly convergent. Thus in theory the Newton-Schulz should be more effective at calculating a highly accurate approximation to the pseudoinverse as compared to the randomized methods, as we confirm in the next experiments.

All the code for the experiments is written in the julia programming language and can be downloaded from `http://www.di.ens.fr/~rgower/` or `https://github.com/gowerrobert/`.

In each figure presented below we plot the evolution of the residual $\|AXA - A\|_F$ against time and flops of each method.

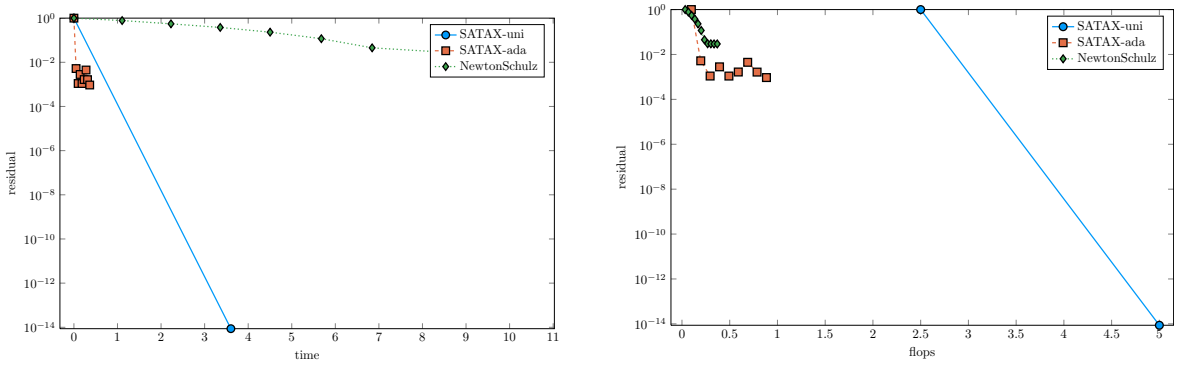

Figure 1: LPnetlib/lp_fit2d $(m; n) = (10, 524; 25)$.

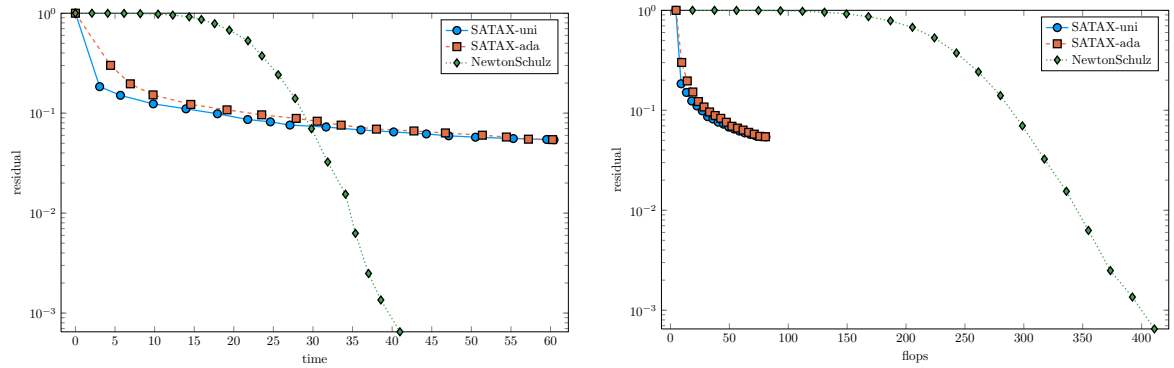

Figure 2: LPnetlib/lp_ken_07 $(m; n) = (2, 426; 3, 602)$.

5.1 Nonsymmetric matrices

In this section we compare the `SATAX_uni`, `SATAX_ada` and Newton-Schulz methods presented earlier in Section 2.3. In setting the initial iterate X_0 for the `SATAX` methods, we know from Lemma 4 and Theorem 3 that we need $X_0 = \alpha A^\top$ for some $\alpha \in \mathbb{R}$ to guarantee that the method converges. We choose α as

$$\alpha = \frac{\min\{n, m\}}{\|A\|_F^2},$$

which is an approximation to the solution of

$$\alpha^* = \arg\min \|A^\dagger - \alpha A^\top\|_F^2,$$

to which the exact solution is $\alpha^* = \mathbf{Rank}\,(A)/\,\|A\|_F^2$.

To verify the performance of the methods, we test several rank deficient matrices from the UF sparse matrix collection[9]. In Figures 1, 2, 3 and 4 we tested the three methods on the LPnetlib/lp_fit2d, the LPnetlib/lp_ken_07, NYPA/Maragal_6 and the Meszaros/primagaz problems, respectively.

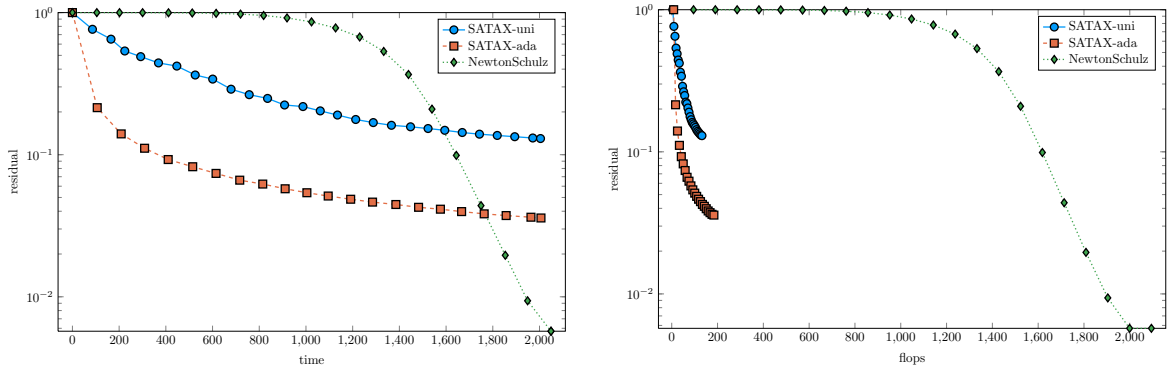

Figure 3: NYPA/Maragal_6 $(m; n) = (21,255; 10,152)$.

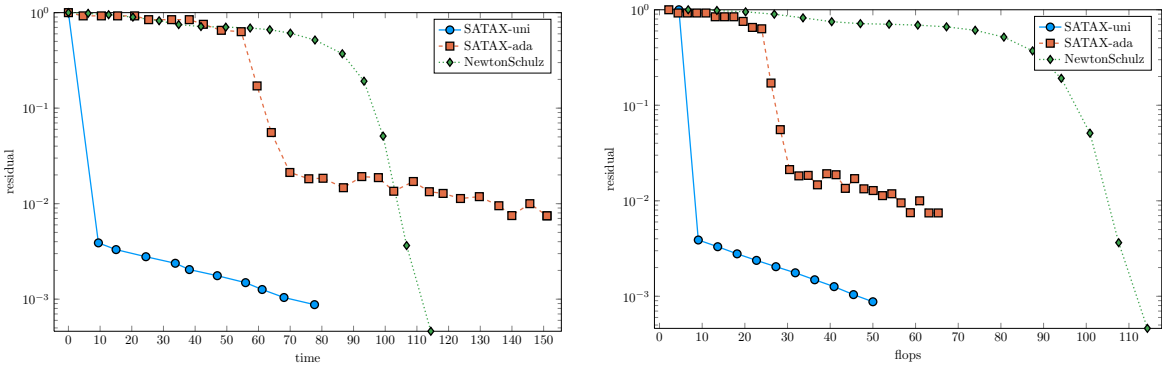

Figure 4: Meszaros/primagaz $(m; n) = (1,554; 10,836)$

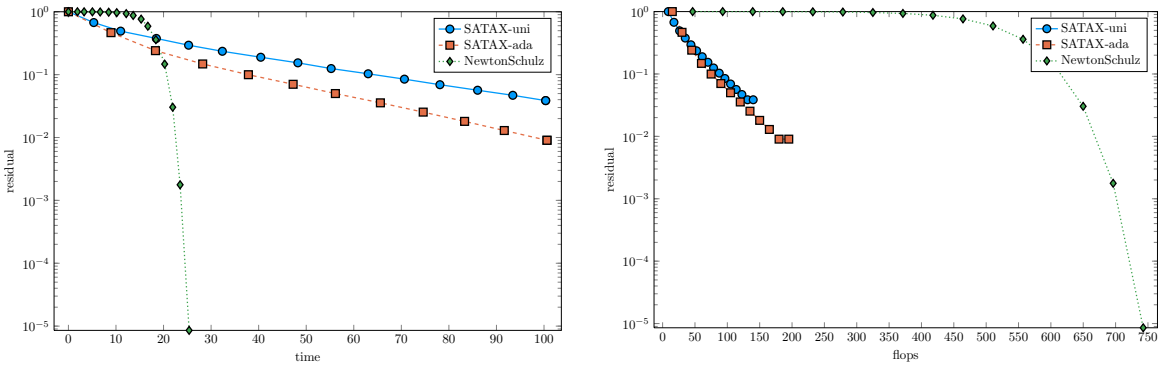

Figure 5: The matrix A is the best rank 1000 approximation to a randomly generated 5000×2500 normal Gaussian matrix.

From Figure 1 we see that the `SATAX` methods are considerably faster at calculating the pseudoinverse on highly rectangular matrices ($n \ll m$ or $n \gg m$) as compared to the Newton-Schultz method. Indeed, by the time the Newton-Schultz method completes three iterations, the stochastic methods have already encountered a pseudoinverse within the desired accuracy. On the remaining problems in Figures 2, 3 and 4 the results are mixed, in that, the `SATAX` methods are very fast at encountering a rough approximation of the pseudoinverse with a residual between 10^{-1} and 10^{-3}, but for reaching a lower residual the Newton-Schultz method proved to be the most efficient.

In calculating the approximate pseudoinverse of the the best rank $r = 1000$ approximation to a random 5000×2500 Gaussian matrix the Newton-Schultz method outperforms the randomized methods in terms of time taken but is less efficient is terms of flops, see in Figure 5. We observed this same result holds for Gaussian matrices with a range of different dimensions and different ranks.

The faster initial convergence of `SATAX` methods and the local quadratic convergence of the Newton-Schulz method can be combined to create an efficient method. To illustrate, we create a combined method named `NS-SATAX` where we use the `SATAX` method for the first few iterations before switching to the Newton-Schulz method, see Figure 6. Through experiments we have identified that we should switch to the Newton-Schulz method after the `SATAX` method has performed one *effective pass over the data*. In other words, we should switch methods after t iterations such that t times the cost of computing the sketched matrix AS is equal to the cost of performing one full matrix-matrix product AX where $X \in \mathbb{R}^{n \times m}$. Though this requires care, in particular, if X_t is the last iteration of the `SATAX` method, then we need to ensure that X_t satisfies the starting condition $\|I - X_t A\|_2 < 1$ of the Newton-Schulz method. For this we normalize the iterate X_t according to $X_t \leftarrow X_t / \|X_t A\|_F$. This normalization is a heuristic and is not guaranteed to satisfy the Newton-Schulz starting condition. Despite this, it does work in practice as we can see in Figure 6 where the combined method `NS-SATAX` outperforms the Newton-Schulz method during the entire execution.

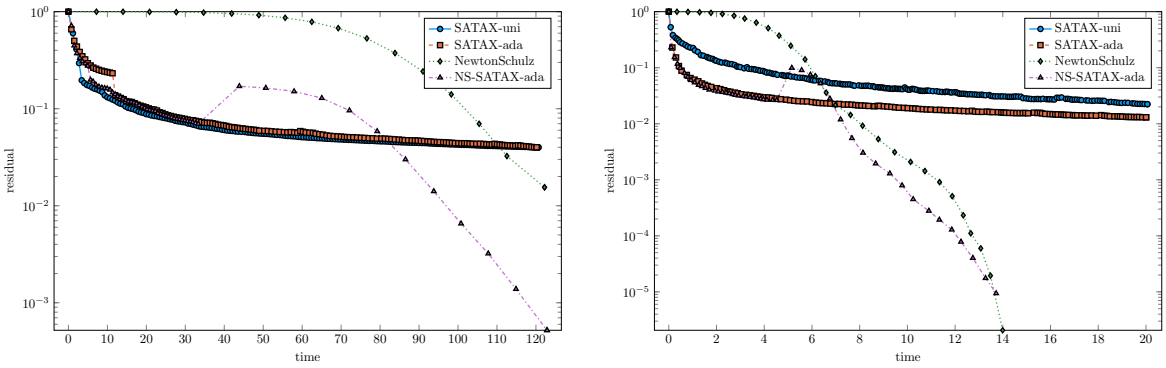

Figure 6: Comparing the new combined method `NS-SATAX` to the `SATAX_uni`, `SATAX_ada` and the Newton-Schulz method on the LPnetlib/lp_ken_07 matrix (LEFT figure) and the NYPA/Maragal_3 (RIGHT figure).

5.2 Symmetric matrices

In this section we compare the `SAXAS_uni`, `SAXAS_ada` and Newton-Schulz methods. In setting the initial iterate X_0 for the `SAXAS` methods, we know from Lemma 7 and Theorem 6 that we need $X_0 = \alpha A^2$ for some $\alpha \in \mathbb{R}$ to guarantee that the method converges. We choose α so that $\|X_0\|_F^2 = 1$, that is $\alpha = 1/\|A\|_F^2$.

To test the symmetric methods we used the Hessian matrix $A^\top A$ of the linear regression problem

$$\min_{x \in \mathbb{R}^n} \frac{1}{2}\|Ax - b\|_2^2 \tag{71}$$

using data from LIBSVM [7], see Figure 7, 8, 9 and 10. These experiments show that the two variants of the `SAXAS` method are much more efficient at calculating an approximate pseudoinverse as compared to the Newton-Schulz method, even for reaching a relative residual with a high precision of around 10^{-6}. The only exception being the `rcv1_train.binary` problem in Figure 10, where the `SAXAS_uni` and `SAXAS_ada` methods make very good progress in the first few iterations, but then struggle to bring the residual much below 10^{-2}. Again looking at Figure 10, the trend appears that the Newton-Schulz method will reach a lower precision than the the `SAXAS_uni` and `SAXAS_ada` after approximately 4000 seconds, though we were not prepared to wait so long. We leave it as an observation that we could again get the best of both worlds by combining an initial execution of the `SAXAS` methods and later switching to the Newton-Schulz method as was done with the `SATAX` and Newton-Schulz method in the previous section.

Again we found that the Newton-Schultz method was more efficient in calculating pseudoinverse of randomly generated Gaussian matrices A, where A is the best rank 1000 approximation to a matrix $G + G^\top$, where G is a 5000×5000 random Gaussian matrix; see Figure 11.

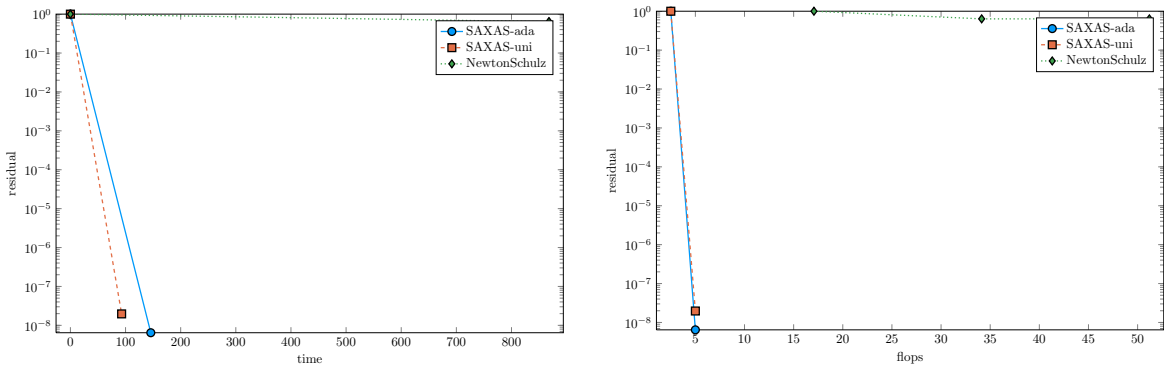

Figure 7: a9a: $(m; n) = (123; 32, 561)$.

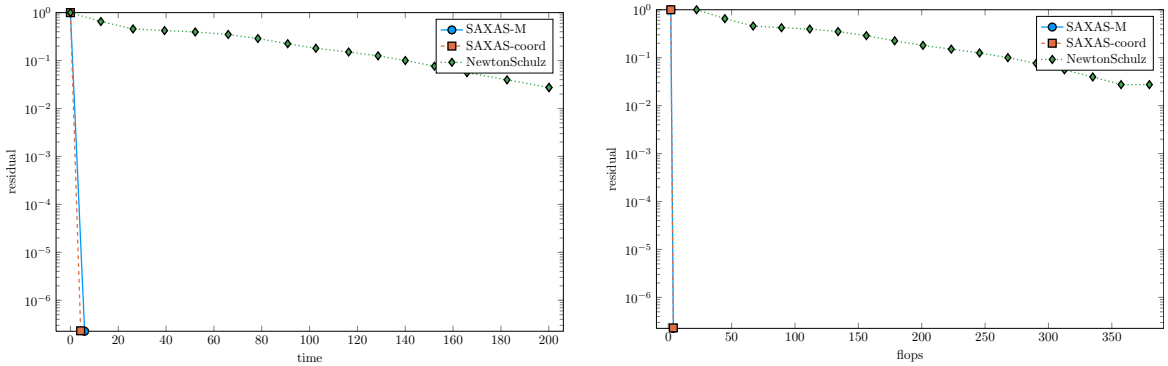

Figure 8: mushrooms: $(m; n) = (8, 124; 112)$

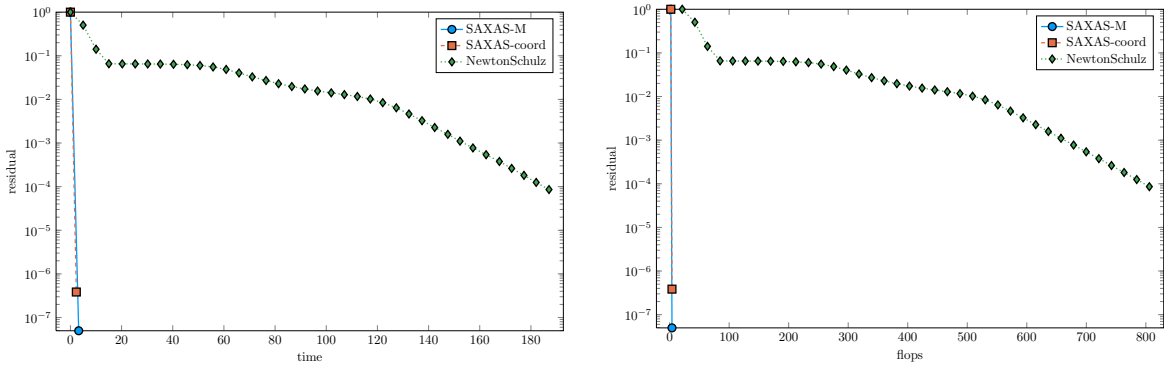

Figure 9: gisette_scale: $(m; n) = (6000; 5000)$

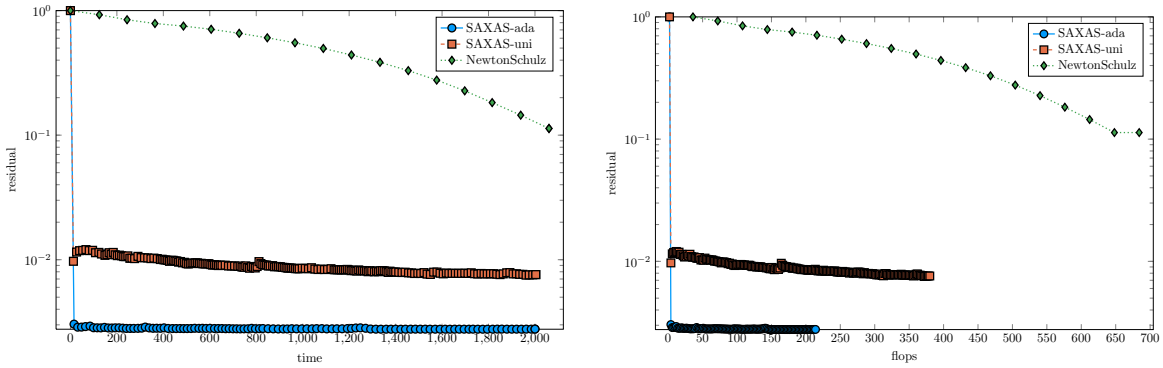

Figure 10: rcv1_train.binary: $(m; n) = (20,242; 47,236)$

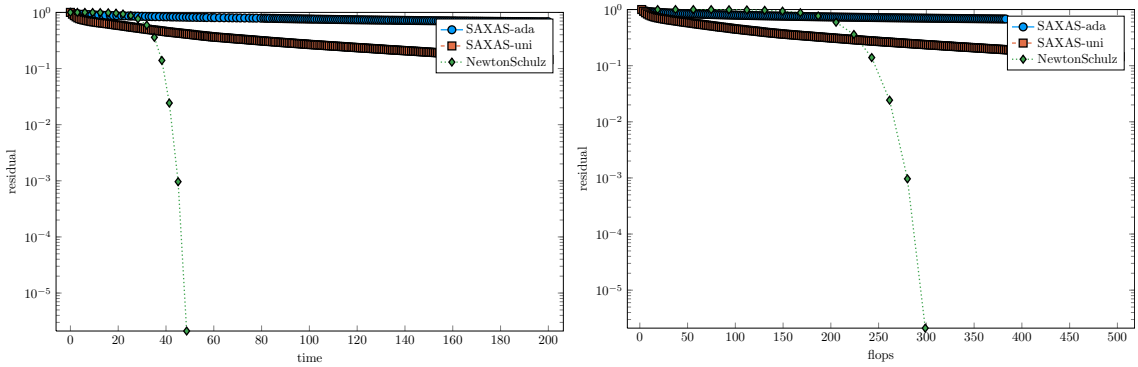

Figure 11: The matrix A is the best rank 10^3 approximation to the matrix $G + G^\top$ where G is a 5000×5000 random Gaussian matrix.

6 Conclusions and Future Work

We presented a new family of randomized methods for iteratively computing the pseudoinverse which are proven to converge linearly to the pseudoinverse matrix and, moreover, numeric experiments show that the new randomized methods are vastly superior at quickly obtaining an approximate pseudoinverse matrix. In such cases where an approximation of the pseudoinverse of a nonsymmetric matrix with a relative residual below 10^{-3} is required then our experiments show that the Newton Schultz method is more effective as compared to our randomized methods. These observations inspired a combined method which we illustrated in Figure 6 which has better overall performance than the Newton-Schulz method. Furthermore, we present new symmetric sketches used to design the `SAXAS` method. For future work, we have indicated how to design randomized methods for calculating approximate range space projections and pseudoinverse of full rank matrices.

References

[1] Y. Hel-Or A. Adler and M. Elad. "A Weighted Discriminative Approach For Image Denoising With Overcomplete Representations". In: *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Dallas, TX, 14-19 March* (2010).

[2] Adi Ben-Israel. "An Iterative Method for Computing the Generalized Inverse of a Matrix". In: *Mathematics of Computation* 19.91 (1965), pp. 452–455.

[3] Adi Ben-Israel and Dan Cohen. "On Iterative Computation of Generalized Inverses and Associated Projections". In: *SIAM Journal on Numerical Analysis* 3.3 (1966), pp. 410–419.

[4] Michele Benzi and Miroslav Tůma. "Comparative Study of Sparse Approximate Inverse Preconditioners". In: *Applied Numerical Mathematics* 30.2–3 (1999), pp. 305–340.

[5] C. G. Broyden. "Quasi-Newton Methods and their Application to Function Minimisation". In: *Mathematics of Computation* 21.99 (1967), pp. 368–381.

[6] Paul H. Calamai and Jorge J. Moré. "Projected gradient methods for linearly constrained problems". In: *Mathematical Programming* 39.1 (1987), pp. 93–116.

[7] Chih Chung Chang and Chih Jen Lin. "LIBSVM : A library for support vector machines". In: *ACM Transactions on Intelligent Systems and Technology* 2.3 (Apr. 2011), pp. 1–27.

[8] Edmond Chow and Yousef Saad. "Approximate Inverse Preconditioners Via Sparse-Sparse Iterations". In: *SIAM Journal of Scientific Computing* 19.3 (1998), pp. 995–1023.

[9] Timothy A. Davis and Yifan Hu. "The University of Florida Sparse Matrix Collection". In: *ACM Transactions on Mathematical Software* 38.1 (2011), 1:1–1:25.

[10] Ron S. Dembo, Stanley C. Eisenstat, and Trond Steihaug. "Inexact Newton Methods". In: *SIAM Journal on Numerical Analysis* 19.2 (1982), pp. 400–408.

[11] C. A. Desoer and B. H. Whalen. "A Note on Pseudoinverses". In: *Journal of the Society of Industrial and Applied Mathematics* 11.2 (1963), pp. 442–447.

[12] Hans G. Feichtinger. "Pseudoinverse matrix methods for signal reconstruction from partial data". In: *Proc. SPIE, Visual Communications and Image Processing '91* 1606 (1991), pp. 766–772.

[13] Jacek Gondzio. "Convergence Analysis of an Inexact Feasible Interior Point Method for Convex Quadratic Programming". In: *SIAM Journal on Optimization* 23.3 (2013), pp. 1510–1527.

[14] N. I. M. Gould, M. E. Hribar, and J. Nocedal. "On the solution of equality constrained quadratic problems arising in optimization". In: *SIAM Journal on Scientific Computing* 23.4 (2001), pp. 1375–1394.

[15] Nicholas I. M. Gould and Jennifer A. Scott. "Sparse Approximate-Inverse Preconditioners Using Norm-Minimization Techniques". In: *SIAM Journal on Scientific Computing* 19.2 (1998), pp. 605–625.

[16] Robert M. Gower. "Sketch and Project: Randomized Iterative Methods for Linear Systems and Inverting Matrices". PhD thesis. University of Edinburgh, 2016.

[17] Robert M. Gower, Donald Goldfarb, and Peter Richtárik. "Stochastic Block BFGS: Squeezing More Curvature out of Data". In: *Proceedings of the 33rd International Conference on Machine Learning* (2016).

[18] Robert M. Gower and Peter Richtárik. "Stochastic Dual Ascent for Solving Linear Systems". In: *arXiv:1512.06890* (2015).

[19] Robert Mansel Gower and Peter Richtárik. "Randomized Iterative Methods for Linear Systems". In: *SIAM Journal on Matrix Analysis and Applications* 36.4 (2015), pp. 1660–1690.

[20] Robert Mansel Gower and Peter Richtárik. "Randomized Quasi-Newton Updates are Linearly Convergent Matrix Inversion Algorithms". In: *arXiv:1602.01768* (2016).

[21] Marcus J. Grote and Thomas Huckle. "Parallel Preconditioning with Sparse Approximate Inverses". In: *SIAM J. Sci. Comput* 18 (1996), pp. 838–853.

[22] Nicolas Loizou and Peter Richtárik. "A new perspective on randomized gossip algorithms". In: *4th IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. 2016.

[23] E. H. Moore. "Abstract for "On the reciprocal of the general algebraic matrix"". In: *Bulletin of the American Mathematical Society* 26 (1920), pp. 394–395.

[24] Victor Pan and Robert Schreiber. "An Improved Newton Iteration for the Generalized Inverse of a Matrix, with Applications". In: *SIAM J. Scientific Computing* 12.5 (1991), pp. 1109–1130.

[25] R. Penrose. "A generalized inverse for matrices". In: *Mathematical Proceedings of the Cambridge Philosophical Society* 51.3 (1955), pp. 406–413.

[26] Gregory Robertson, R. Lynn Kirlin, and W. S. Lu. "A Pseudoinverse Update Algorithm for Rank-Reduced Covariance Matrices from 2-D Data". In: *IEEE Signal Processing Letters* 3 (8 1997).

[27] Thomas Strohmer and Roman Vershynin. "A Randomized Kaczmarz Algorithm with Exponential Convergence". In: *Journal of Fourier Analysis and Applications* 15.2 (2009), pp. 262–278.

[28] J. Tapson and A. Van Schaik. "Learning the Pseudoinverse Solution to Network Weights". In: *Neural Networks* 45 (Sept. 2013), pp. 94–100.

7 Appendix

Here we present and prove several fundamental linear algebra lemmas that are required to develop the main theorems in the paper.

7.1 Key linear algebra lemmas

Lemma 13 *For any matrix W and symmetric positive semidefinite matrix G such that*

$$\mathbf{Null}\,(G) \subset \mathbf{Null}\left(W^\top\right), \tag{72}$$

we have that

$$\mathbf{Null}\,(W) = \mathbf{Null}\left(W^\top G W\right) \tag{73}$$

and

$$\mathbf{Range}\left(W^\top\right) = \mathbf{Range}\left(W^\top G W\right). \tag{74}$$

Proof: In order to establish (73), it suffices to show the inclusion $\mathbf{Null}\,(W) \supseteq \mathbf{Null}\left(W^\top G W\right)$ since the reverse inclusion trivially holds. Letting $s \in \mathbf{Null}\left(W^\top G W\right)$, we see that $\|G^{1/2} W s\|^2 = 0$, which implies $G^{1/2} W s = 0$. Consequently

$$W s \in \mathbf{Null}\left(G^{1/2}\right) = \mathbf{Null}\,(G) \overset{(72)}{\subset} \mathbf{Null}\left(W^\top\right).$$

Thus $W s \in \mathbf{Null}\left(W^\top\right) \cap \mathbf{Range}\,(W)$ which are orthogonal complements which shows that $W s = 0$.

Finally, (74) follows from (73) by taking orthogonal complements. Indeed, $\mathbf{Range}\left(W^\top\right)$ is the orthogonal complement of $\mathbf{Null}\left(W\right)$ and $\mathbf{Range}\left(W^\top GW\right)$ is the orthogonal complement of $\mathbf{Null}\left(W^\top GW\right)$. ∎

The following two lemmas are of key importance throughout the paper.

Lemma 14 *For any matrix $M \in \mathbb{R}^{m\times n}$ and any matrix $R \in \mathbb{R}^{n\times d}$ such that $\mathbf{Range}\left(R\right) \subset \mathbf{Range}\left(M^\top\right)$ we have that*

$$\left\langle M^\top MR, R\right\rangle \geq \lambda^+_{\min}(M^\top M)\left\langle R, R\right\rangle, \tag{75}$$

Proof: Since

$$\left\langle M^\top MR, R\right\rangle = \mathbf{Tr}\left(R^\top M^\top MR\right) = \sum_{i=1}^{d}\left\langle M^\top MR_{:i}, R_{:i}\right\rangle,$$

the inequality (75) follows from the known inequality

$$\left\langle M^\top Mvv\right\rangle \geq \lambda^+_{\min}(M^\top M)\left\langle v, v\right\rangle,$$

where $v \in \mathbf{Range}\left(M^\top\right)$, which can be proved be diagonalizing $M^\top M$.

Lemma 15 *Let $0 \neq W \in \mathbb{R}^{m\times n}$ and $G \in \mathbb{R}^{m\times m}$ be symmetric positive semi-definite with $\mathbf{Null}\left(G\right) \subset \mathbf{Null}\left(W^\top\right)$. Then the matrix $W^\top GW$ has a positive eigenvalue, and the following inequality holds:*

$$\left\langle W^\top GWR, R\right\rangle \geq \lambda^+_{\min}(W^\top GW)\left\langle R, R\right\rangle, \tag{76}$$

where R is a matrix with n rows and $\mathbf{Range}\left(R\right) \subset \mathbf{Range}\left(W^\top\right)$.

Proof: By Lemma 14 with $M = G^{1/2}W$ we have that (76) holds for $\mathbf{Range}\left(R\right) \subset \mathbf{Range}\left(W^\top G^{1/2}\right)$. The proof now follows by observing

$$\mathbf{Range}\left(W^\top G^{1/2}\right) \overset{\text{Lemma 13}}{=} \mathbf{Range}\left(W^\top GW\right) \overset{\text{Lemma 13}}{=} \mathbf{Range}\left(W^\top\right). \quad ∎$$

7.2 Smallest nonzero eigenvalue of the product of two matrices

Lemma 16 *Let $A, B \in \mathbb{R}^{n\times n}$ be symmetric positive semidefinite matrices. If*

$$\mathbf{Null}\left(A\right) \subset \mathbf{Null}\left(B\right) \tag{77}$$

then

$$\lambda^+_{\min}(AB) \geq \lambda^+_{\min}(A)\lambda^+_{\min}(B). \tag{78}$$

Proof: Using the variational formulation we have that

$$\begin{aligned}\lambda^+_{\min}(AB) &= \min_{v\in\mathbf{Null}(AB)^\perp}\frac{\|ABv\|}{\|v\|}\\ &= \min_{v\in\mathbf{Null}(AB)^\perp}\frac{\|ABv\|}{\|Bv\|}\frac{\|Bv\|}{\|v\|}\\ &\leq \min_{v\in\mathbf{Null}(AB)^\perp}\frac{\|ABv\|}{\|Bv\|}\min_{v\in\mathbf{Null}(AB)^\perp}\frac{\|Bv\|}{\|v\|}. \end{aligned}\tag{79}$$

Given that

$$\mathbf{Null}\,(B)\subset\mathbf{Null}\,(AB)$$

ergo

$$\mathbf{Null}\,(AB)^\perp\subset\mathbf{Null}\,(B)^\perp\overset{(77)}{\subset}\mathbf{Null}\,(A)^\perp.$$

The above shows that

$$\min_{v\in\mathbf{Null}(AB)^\perp}\frac{\|Bv\|}{\|v\|}\geq\min_{v\in\mathbf{Null}(B)^\perp}\frac{\|Bv\|}{\|v\|}=\lambda^+_{\min}(B).\tag{80}$$

But also, since $\mathbf{Null}\,(B)^\perp=\mathbf{Range}\,(B)=\mathbf{Range}\,(BB)$ which follows from B being symmetric and Lemma 13, we have that

$$\begin{aligned}\min_{v\in\mathbf{Null}(AB)^\perp}\frac{\|ABv\|}{\|Bv\|} &\geq \min_{v\in\mathbf{Range}(B)}\frac{\|ABv\|}{\|Bv\|}\\ &= \min_{w\in\mathbf{Range}(BB)}\frac{\|Aw\|}{\|w\|}\\ &= \min_{w\in\mathbf{Range}(B)}\frac{\|Aw\|}{\|w\|}=\lambda^+_{\min}(A).\end{aligned}\tag{81}$$

Inserting (80) and (81) in (79) gives the desired result.

7.3 Convenient probability lemma

Theorem 17 *Let G be a positive symmetric semidefinite matrix. Let S be a random matrix with a finite discrete distribution with r $S=S_i\in\mathbb{R}^{n\times q_i}$ with probability $p_i>0$ for $i=1,\ldots,r$. Let $\mathbb{S}\overset{def}{=}[S_1,\ldots,S_r]\in\mathbb{R}^{n\times n}$. If*

$$p_i = \frac{\mathbf{Tr}\left(S_i^\top G^2 S_i\right)}{\mathbf{Tr}\left(\mathbb{S}^\top G^2\mathbb{S}\right)},\quad \textit{for}\quad i=1,\ldots,r.\tag{82}$$

then

$$\lambda^+_{\min}\left(G\mathbf{E}\left[S(S^\top G^2 S)^\dagger S^\top\right]G\right)\quad\geq\quad\frac{\lambda^+_{\min}\left(\mathbb{S}^\top G^2\mathbb{S}\right)}{\mathbf{Tr}\left(\mathbb{S}^\top G^2\mathbb{S}\right)}.\tag{83}$$

Proof: Let $Z = GS(S^\top G^2 S)^\dagger S^\top G$. Note that

$$\mathbf{E}\left[Z\right] = G\mathbb{S}D\mathbb{S}^\top G, \tag{84}$$

with

$$D \quad \stackrel{\text{def}}{=} \quad \text{diag}(p_i(S^\top G^2 S)^\dagger). \tag{85}$$

Let $t_i = \mathbf{Tr}\left(S_i^\top G^2 S_i\right)$, and with (82) in (85) we have

$$D = \frac{1}{\mathbf{Tr}\left(\mathbb{S}^\top G^2 \mathbb{S}\right)} \text{diag}\left(t_1(S_1^\top G^2 S_1)^\dagger, \ldots, t_r(S_r^\top G^2 S_r)^\dagger\right),$$

thus

$$\lambda_{\min}^+(D) = \frac{1}{\mathbf{Tr}\left(\mathbb{S}^\top G^2 \mathbb{S}\right)} \min_{i\,:\,t_i \neq 0} \left\{\frac{t_i}{\lambda_{\max}(S_i^\top G^2 S_i)}\right\} \geq \frac{1}{\mathbf{Tr}\left(\mathbb{S}^\top G^2 \mathbb{S}\right)}. \tag{86}$$

Thus

$$\begin{aligned} \lambda_{\min}^+\left(\mathbf{E}\left[Z\right]\right) \quad &\stackrel{(84)}{=} \quad \lambda_{\min}^+\left(G\mathbb{S}D\mathbb{S}^\top G\right) \\ &= \quad \lambda_{\min}^+\left(D\mathbb{S}^\top G G\mathbb{S}\right) \\ &\stackrel{\text{Lemma 16}}{\geq} \quad \frac{\lambda_{\min}^+\left(\mathbb{S}^\top G G\mathbb{S}\right)}{\mathbf{Tr}\left(\mathbb{S}^\top G^2 \mathbb{S}\right)}, \end{aligned} \tag{87}$$

where in the second line we used that, for any matrices A, B, the matrices AB and BA share the same nonzero eigenvalues. ∎