THE UNIVERSITY of EDINBURGH

Edinburgh Research Explorer

Stochastic Dual Ascent for Solving Linear Systems

Citation for published version:
Gower, RM & Richtarik, P 2015 'Stochastic Dual Ascent for Solving Linear Systems' ArXiv.

Link:
Link to publication record in Edinburgh Research Explorer

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

OPEN ACCESS

Download date: 14. Jun. 2018

arXiv:1512.06890v2 [math.NA] 28 Jan 2016

Stochastic Dual Ascent for Solving Linear Systems

Robert M. Gower and Peter Richtárik*

School of Mathematics
University of Edinburgh
United Kingdom

December 21, 2015

Abstract

We develop a new randomized iterative algorithm—*stochastic dual ascent (SDA)*—for finding the projection of a given vector onto the solution space of a linear system. The method is dual in nature: with the dual being a non-strongly concave quadratic maximization problem without constraints. In each iteration of SDA, a dual variable is updated by a carefully chosen point in a subspace spanned by the columns of a random matrix drawn independently from a fixed distribution. The distribution plays the role of a parameter of the method. Our complexity results hold for a wide family of distributions of random matrices, which opens the possibility to fine-tune the stochasticity of the method to particular applications. We prove that primal iterates associated with the dual process converge to the projection exponentially fast in expectation, and give a formula and an insightful lower bound for the convergence rate. We also prove that the same rate applies to dual function values, primal function values and the duality gap. Unlike traditional iterative methods, SDA converges under no additional assumptions on the system (e.g., rank, diagonal dominance) beyond consistency. In fact, our lower bound improves as the rank of the system matrix drops. Many existing randomized methods for linear systems arise as special cases of SDA, including randomized Kaczmarz, randomized Newton, randomized coordinate descent, Gaussian descent, and their variants. In special cases where our method specializes to a known algorithm, we either recover the best known rates, or improve upon them. Finally, we show that the framework can be applied to the distributed average consensus problem to obtain an array of new algorithms. The randomized gossip algorithm arises as a special case.

1 Introduction

Probabilistic ideas and tools have recently begun to permeate into several fields where they had traditionally not played a major role, including fields such as numerical linear algebra and optimization. One of the key ways in which these ideas influence these fields is via the development and analysis of *randomized algorithms* for solving standard and new problems of these fields. Such methods are typically easier to analyze, and often lead to faster and/or more scalable and versatile methods in practice.

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

1.1 The problem

In this paper we consider a key problem in linear algebra, that of finding a solution of a system of linear equations

$$Ax = b, \tag{1}$$

where $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$. We shall assume throughout that the system is *consistent*, that is, that there exists x^* for which $Ax^* = b$. While we assume the existence of a solution, we do not assume uniqueness. In situations with multiple solutions, one is often interested in finding a solution with specific properties. For instance, in compressed sensing and sparse optimization, one is interested in finding the least ℓ_1-norm, or the least ℓ_0-norm (sparsest) solution.

In this work we shall focus on the canonical problem of finding the solution of (1) closest, with respect to a Euclidean distance, to a given vector $c \in \mathbb{R}^n$:

$$\begin{array}{rl} \text{minimize} & P(x) \stackrel{\text{def}}{=} \frac{1}{2}\|x-c\|_B^2 \\ \text{subject to} & Ax = b \\ & x \in \mathbb{R}^n. \end{array} \tag{2}$$

where B is an $n \times n$ symmetric positive definite matrix and $\|x\|_B \stackrel{\text{def}}{=} \sqrt{x^\top Bx}$. By x^* we denote the (necessarily) unique solution of (2). Of key importance in this paper is the *dual problem*[1] to (2), namely

$$\begin{array}{rl} \text{maximize} & D(y) \stackrel{\text{def}}{=} (b-Ac)^\top y - \frac{1}{2}\|A^\top y\|_{B^{-1}}^2 \\ \text{subject to} & y \in \mathbb{R}^m. \end{array} \tag{3}$$

Due to the consistency assumption, strong duality holds and we have $P(x^*) = D(y^*)$, where y^* is any dual optimal solution.

1.2 A new family of stochastic optimization algorithms

We propose to solve (2) via a new method operating in the dual (3), which we call *stochastic dual ascent* (SDA). The iterates of SDA are of the form

$$y^{k+1} = y^k + S\lambda^k, \tag{4}$$

where S is a random matrix with m rows drawn in each iteration independently from a pre-specified distribution $\mathcal{D}$, which should be seen as a parameter of the method. In fact, by varying $\mathcal{D}$, SDA should be seen as a family of algorithms indexed by $\mathcal{D}$, the choice of which leads to specific algorithms in this family. By performing steps of the form (4), we are moving in the range space of the random matrix S. A key feature of SDA enabling us to prove strong convergence results despite the fact that the dual objective is in general not strongly concave is the way in which the "stepsize" parameter λ^k is chosen: we chose λ^k to be the *least-norm* vector for which $D(y^k + S\lambda)$ is maximized in λ. Plugging this λ^k into (4), we obtain the SDA method:

$$\boxed{y^{k+1} = y^k + S\left(S^\top AB^{-1}A^\top S\right)^\dagger S^\top\left(b - A\left(c + B^{-1}A^\top y^k\right)\right)} \tag{5}$$

[1] Technically, this is both the Lagrangian and Fenchel dual of (2).

The symbol † denotes the Moore-Penrose pseudoinverse[2].

To the best of our knowledge, a randomized optimization algorithm with iterates of the *general* form (4) was not considered nor analyzed before. In the special case when S is chosen to be a random unit coordinate vector, SDA specializes to the *randomized coordinate descent method*, first analyzed by Leventhal and Lewis [21]. In the special case when S is chosen as a random column submatrix of the $m \times m$ identity matrix, SDA specializes to the *randomized Newton method* of Qu, Fercoq, Richtárik and Takáč [44].

With the dual iterates $\{y^k\}$ we associate a sequence of primal iterates $\{x^k\}$ as follows:

$$x^k \stackrel{\text{def}}{=} c + B^{-1}A^\top y^k. \tag{6}$$

In combination with (5), this yields the primal iterative process

$$\boxed{x^{k+1} = x^k - B^{-1}A^\top S\left(S^\top AB^{-1}A^\top S\right)^\dagger S^\top\left(Ax^k - b\right)} \tag{7}$$

Optimality conditions (see Section 2.1) imply that if y^* is any dual optimal point, then $c + B^{-1}A^\top y^*$ is necessarily primal optimal and hence equal to x^*, the optimal solution of (2). Moreover, we have the following useful and insightful correspondence between the quality of the primal and dual iterates (see Proposition 2.2):

$$D(y^*) - D(y^k) = \tfrac{1}{2}\|x^k - x^*\|_B^2. \tag{8}$$

Hence, *dual convergence in function values is equivalent to primal convergence in iterates.*

Our work belongs to a growing literature on randomized methods for various problems appearing in linear algebra, optimization and computer science. In particular, relevant methods include sketching algorithms, randomized Kaczmarz, stochastic gradient descent and their variants [55, 31, 9, 38, 66, 33, 34, 45, 50, 56, 17, 47, 19, 62, 8, 18, 65, 35, 7, 32, 26, 14, 40, 23] and randomized coordinate and subspace type methods and their variants [21, 16, 51, 36, 60, 3, 48, 37, 49, 57, 28, 58, 30, 46, 11, 53, 10, 12, 20, 41, 13, 42, 43, 64, 44, 63, 52, 22, 6, 14].

1.3 The main results

We now describe two complexity theorems which form the core theoretical contribution of this work. The results hold for a wide family of distributions $\mathcal{D}$, which we describe next.

Weak assumption on $\mathcal{D}$. In our analysis, we only impose a very weak assumption on $\mathcal{D}$. In particular, we only assume that the $m \times m$ matrix

$$H \stackrel{\text{def}}{=} \mathbf{E}_{S\sim\mathcal{D}}\left[S\left(S^\top AB^{-1}A^\top S\right)^\dagger S^\top\right] \tag{9}$$

is well defined and nonsingular[3]. Hence, we do not assume that S be picked from any particular random matrix ensamble: the options are, quite literally, limitless. This makes it possible for practitioners to choose the best distribution specific to a particular application.

[2]It is known that the vector $M^\dagger d$ is the least-norm solution of the least-squares problem $\min_\lambda \|M\lambda - d\|^2$. Hence, if the system $M\lambda = d$ has a solution, then $M^\dagger d = \arg\min_\lambda\{\|\lambda\| \;:\; M\lambda = d\}$.

[3]It is known that the pseudoinverse of a symmetric positive semidefinite matrix is again symmetric and positive semidefinite. As a result, if the expectation defining H is finite, H is also symmetric and positive semidefinite. Hence, we could equivalently assume that H be positive definite.

We cast the first complexity result in terms of the primal iterates since solving (2) is our main focus in this work. Let $\mathbf{Range}(M), \mathbf{Rank}(M)$ and $\lambda_{\min}^{+}(M)$ denote the range space, rank and the smallest nonzero eigenvalue of M, respectively.

Theorem 1.1 (Convergence of primal iterates and of the residual). *Assume that the matrix H, defined in (9), is nonsingular. Fix arbitrary $x^0 \in \mathbb{R}^n$. The primal iterates $\{x^k\}$ produced by (7) converge exponentially fast in expectation to $x^* + t$, where x^* is the optimal solution of the primal problem (2), and t is the projection of $x^0 - c$ onto $\mathbf{Null}(A)$:*

$$t \overset{def}{=} \arg\min_{t'} \left\{ \|x^0 - c - t'\|_B \ : \ t' \in \mathbf{Null}(A) \right\}. \tag{10}$$

In particular, for all $k \geq 0$ we have

$$\textit{Primal iterates:} \qquad \mathbf{E}\left[\|x^k - x^* - t\|_B^2\right] \leq \rho^k \cdot \|x^0 - x^* - t\|_B^2, \tag{11}$$

$$\textit{Residual:} \qquad \mathbf{E}\left[\|Ax^k - b\|_B\right] \leq \rho^{k/2}\|A\|_B\|x^0 - x^* - t\|_B + \|At\|_B, \tag{12}$$

where $\|A\|_B \overset{def}{=} \max\{\|Ax\|_B \ : \ \|x\|_B \leq 1\}$ and

$$\rho \overset{def}{=} 1 - \lambda_{\min}^{+}\left(B^{-1/2}A^\top HAB^{-1/2}\right). \tag{13}$$

Furthermore, the convergence rate is bounded by

$$1 - \frac{\mathbf{E}\left[\mathbf{Rank}\left(S^\top A\right)\right]}{\mathbf{Rank}(A)} \leq \rho < 1. \tag{14}$$

If we let S be a unit coordinate vector chosen at random, B be the identity matrix and set $c = 0$, then (7) reduces to the *randomized Kaczmarz (RK)* method proposed and analyzed in a seminal work of Strohmer and Vershynin [55]. Theorem 1.1 implies that RK converges with an exponential rate so long as the system matrix has no zero rows (see Section 3). To the best of our knowledge, such a result was not previously established: current convergence results for RK assume that the system matrix is full rank [26, 45]. Not only do we show that the RK method converges to the least-norm solution for any consistent system, but we do so through a single all encompassing theorem covering a wide family of algorithms. Likewise, convergence of block variants of RK has only been established for full column rank [33, 35]. Block versions of RK can be obtained from our generic method by choosing $B = I$ and $c = 0$, as before, but letting S to be a random column submatrix of the identity matrix (see [14]). Again, our general complexity bound holds under no assumptions on A, as long as one can find S such that H becomes nonsingular.

The lower bound (14) says that for a singular system matrix, the number of steps required by SDA to reach an expected accuracy is at best inversely proportional to the rank of A. If A has row rank equal to one, for instance, then RK converges in one step (this is no surprise, given that RK projects onto the solution space of a single row, which in this case, is the solution space of the whole system). Our lower bound in this case becomes 0, and hence is tight.

While Theorem 1.1 is cast in terms of the primal iterates, if we assume that $x^0 = c + B^{-1}A^\top y^0$ for some $y^0 \in \mathbb{R}^m$, then an equivalent dual characterization follows by combining (6) and (8). In fact, in that case we can also establish the convergence of the primal function values and of the duality gap. *No such results were previously known.*

Theorem 1.2 (Convergence of function values). *Assume that the matrix H, defined in (9), is nonsingular. Fix arbitrary $y^0 \in \mathbb{R}^m$ and let $\{y^k\}$ be the SDA iterates produced by (5). Further, let $\{x^k\}$ be the associated primal iterates, defined by (6), $OPT \overset{def}{=} P(x^*) = D(y^*)$,*

$$U_0 \overset{def}{=} \tfrac{1}{2}\|x^0 - x^*\|_B^2 \overset{(8)}{=} OPT - D(y^0),$$

and let ρ be as in Theorem 1.1. Then for all $k \geq 0$ we have the following complexity bounds:

$$\textit{Dual suboptimality:} \qquad \mathbf{E}\left[OPT - D(y^k)\right] \leq \rho^k U_0 \tag{15}$$

$$\textit{Primal suboptimality:} \qquad \mathbf{E}\left[P(x^k) - OPT\right] \leq \rho^k U_0 + 2\rho^{k/2}\sqrt{OPT \times U_0} \tag{16}$$

$$\textit{Duality gap:} \qquad \mathbf{E}\left[P(x^k) - D(y^k)\right] \leq 2\rho^k U_0 + 2\rho^{k/2}\sqrt{OPT \times U_0} \tag{17}$$

Note that the dual objective function is *not* strongly concave in general, and yet we prove linear convergence (see (15)). It is known that for *some* structured optimization problems, linear convergence results can be obtained without the need to assume strong concavity (or strong convexity, for minimization problems). Typical approaches to such results would be via the employment of error bounds [25, 59, 15, 27, 29]. *In our analysis, no error bounds are necessary.*

1.4 Outline

The paper is structured as follows. Section 2 describes the algorithm in detail, both in its dual and primal form, and establishes several useful identities. In Section 3 we characterize discrete distributions for which our main assumption on H is satisfied. We then specialize our method to several simple discrete distributions to better illustrate the results. We then show in Section 4 how SDA can be applied to design new randomized gossip algorithms. We also show that our framework can recover some standard methods. Theorem 1.1 is proved in Section 5 and Theorem 1.2 is proved in Section 6. In Section **??** we perform a simple experiment illustrating the convergence of the randomized Kaczmarz method on rank deficient linear systems. We conclude in Section 8. To the appendix we relegate two elementary but useful technical results which are needed multiple times in the text.

2 Stochastic Dual Ascent

By *stochastic dual ascent* (SDA) we refer to a randomized optimization method for solving the dual problem (3) performing iterations of the form

$$y^{k+1} = y^k + S\lambda^k, \tag{18}$$

where S is a random matrix with m rows drawn in each iteration independently from a prespecified distribution. We shall not fix the number of columns of S; in fact, we even allow for the number of columns to be random. By performing steps of the form (18), we are moving in the range space of the random matrix S, with λ^k describing the precise linear combination of the columns used in computing the step. In particular, we shall choose λ^k from the set

$$Q^k \overset{\text{def}}{=} \arg\max_\lambda D(y^k + S\lambda) \overset{(3)}{=} \arg\max_\lambda \left\{ (b - Ac)^\top (y^k + S\lambda) - \tfrac{1}{2}\left\| A^\top (y^k + S\lambda) \right\|_{B^{-1}}^2 \right\}.$$

Since D is bounded above (a consequence of weak duality), this set is nonempty. Since D is a concave quadratic, Q^k consists of all those vectors λ for which the gradient of the mapping $\phi_k(\lambda) : \lambda \mapsto D(y^k + S\lambda)$ vanishes. This leads to the observation that Q^k is the set of solutions of a random linear system:

$$Q^k = \left\{ \lambda \in \mathbb{R}^m \ : \ \left(S^\top A B^{-1} A^\top S \right) \lambda = S^\top \left(b - Ac - AB^{-1} A^\top y^k \right) \right\}.$$

If S has a small number of columns, this is a small easy-to-solve system.

A key feature of our method enabling us to prove exponential error decay despite the lack of strong concavity is the way in which we choose λ^k from Q^k. In SDA, λ^k is chosen to be the least-norm element of Q^k,

$$\lambda^k \stackrel{\text{def}}{=} \arg\min_{\lambda \in Q^k} \|\lambda\|,$$

where $\|\lambda\| = (\sum_i \lambda_i^2)^{1/2}$ denotes standard Euclidean norm. The least-norm solution of a linear system can be written down in a compact way using the (Moore-Penrose) pseudoinverse. In our case, we obtain the formula

$$\lambda^k = \left(S^\top A B^{-1} A^\top S \right)^\dagger S^\top \left(b - Ac - AB^{-1} A^\top y^k \right), \tag{19}$$

where $\dagger$ denotes the pseudoinverse operator. Note that if S has only a few columns, then (19) requires projecting the origin onto a small linear system. The SDA algorithm is obtained by combining (18) with (19).

Algorithm 1 Stochastic Dual Ascent (SDA)

parameter: $\mathcal{D}$ = distribution over random matrices
Choose $y^0 \in \mathbb{R}^m$ ▷ Initialization
for $k = 0, 1, 2, \ldots$ **do**
    Sample an independent copy $S \sim \mathcal{D}$
    $\lambda^k = \left(S^\top A B^{-1} A^\top S \right)^\dagger S^\top \left(b - Ac - AB^{-1} A^\top y^k \right)$
    $y^{k+1} = y^k + S\lambda^k$ ▷ Update the dual variable

The method has one parameter: the distribution $\mathcal{D}$ from which the random matrices S are drawn. Sometimes, one is interested in finding any solution of the system $Ax = b$, rather than the particular solution described by the primal problem (2). In such situations, B and c could also be seen as parameters.

2.1 Optimality conditions

For any x for which $Ax = b$ and for any y we have

$$P(x) - D(y) \stackrel{(2)+(3)}{=} \tfrac{1}{2}\|x - c\|_B^2 + \tfrac{1}{2}\|A^\top y\|_{B^{-1}}^2 + (c - x)^\top A^\top y \geq 0,$$

where the inequality (weak duality) follows from the Fenchel-Young inequality[4]. As a result, we obtain the following necessary and sufficient optimality conditions, characterizing primal and dual optimal points.

Proposition 2.1 (Optimality conditions)**.** *Vectors $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^m$ are optimal for the primal* (2) *and dual* (3) *problems respectively, if and only if they satisfy the following relation*

$$Ax = b, \qquad x = c + B^{-1}A^\top y. \tag{20}$$

In view of this, it will be useful to define a linear mapping from $\mathbb{R}^m$ to $\mathbb{R}^n$ as follows:

$$x(y) = c + B^{-1}A^\top y. \tag{21}$$

As an immediate corollary of Proposition 2.1 we observe that for any dual optimal y^*, the vector $x(y^*)$ must be primal optimal. Since the primal problem has a unique optimal solution, x^*, we must necessarily have

$$x^* = x(y^*) = c + B^{-1}A^\top y^*. \tag{22}$$

Another immediate corollary of Proposition 2.1 is the following characterization of dual optimality: y is dual optimal if and only if

$$b - Ac = AB^{-1}A^\top y. \tag{23}$$

Hence, the set of dual optimal solutions is $\mathcal{Y}^* = (AB^{-1}A^\top)^\dagger(b - Ac) + \mathbf{Null}\left(AB^{-1}A^\top\right)$. Since, $\mathbf{Null}\left(AB^{-1}A^\top\right) = \mathbf{Null}\left(A^\top\right)$ (see Lemma 10.1), we have

$$\mathcal{Y}^* = \left(AB^{-1}A^\top\right)^\dagger (b - Ac) + \mathbf{Null}\left(A^\top\right).$$

Combining this with (22), we get

$$x^* = c + B^{-1}A^\top \left(AB^{-1}A^\top\right)^\dagger (b - Ac).$$

Remark 2.1 (The dual is also a least-norm problem.)**.** *Observe that:*

1. *The particular dual optimal point $y^* = (AB^{-1}A^\top)^\dagger(b - Ac)$ is the solution of the following optimization problem:*

$$\min\left\{\tfrac{1}{2}\|y\|^2 \;:\; AB^{-1}A^\top y = b - Ac\right\}. \tag{24}$$

   Hence, this particular formulation of the dual problem has the same form as the primal problem: projection onto a linear system.

[4]Let U be a vector space equipped with an inner product $\langle\cdot,\cdot\rangle : U \times U \to \mathbb{R}$. Given a function $f : U \to \mathbb{R}$, its convex (or Fenchel) conjugate $f^* : U \to \mathbb{R} \cup \{+\infty\}$ is defined by $f^*(v) = \sup_{u\in U}\langle u, v\rangle - f(u)$. A direct consequence of this is the Fenchel-Young inequality, which asserts that $f(u) + f^*(v) \geq \langle u, v\rangle$ for all u and v. The inequality in the main text follows by choosing $f(u) = \frac{1}{2}\|u\|_B^2$ (and hence $f^*(v) = \frac{1}{2}\|v\|_{B^{-1}}^2$), $u = x - c$ and $v = A^\top y$. If f is differentiable, then equality holds if and only if $v = \nabla f(u)$. In our case, this condition is $x = c + B^{-1}A^\top y$. This, together with primal feasibility, gives the optimality conditions (20). For more details on Fenchel duality, see [1].

2. *If $A^\top A$ is positive definite (which can only happen if A is of full column rank, which means that $Ax = b$ has a unique solution and hence the primal objective function does not matter), and we choose $B = A^\top A$, then the dual constraint* (24) *becomes*

$$A(A^\top A)^{-1}A^\top y = b - Ac.$$

This constraint has a geometric interpretation: we are seeking vector y whose orthogonal projection onto the column space of A is equal to $b - Ac$. Hence the reformulated dual problem (24) *is asking us to find the vector y with this property having the least norm.*

2.2 Primal iterates associated with the dual iterates

With the sequence of dual iterates $\{y^k\}$ produced by SDA we can associate a sequence of primal iterates $\{x^k\}$ using the mapping (21):

$$x^k \overset{\text{def}}{=} x(y^k) = c + B^{-1}A^\top y^k. \tag{25}$$

This leads to the following *primal version of the SDA method.*

Algorithm 2 Primal Version of Stochastic Dual Ascent (SDA-Primal)

parameter: $\mathcal{D}$ = distribution over random matrices
Choose $x^0 \in \mathbb{R}^n$ ▷ Initialization
for $k = 0, 1, 2, \ldots$ **do**
    Sample an independent copy $S \sim \mathcal{D}$
    $x^{k+1} = x^k - B^{-1}A^\top S\left(S^\top AB^{-1}A^\top S\right)^\dagger S^\top(Ax^k - b)$ ▷ Update the primal variable

Remark 2.2. *A couple of observations:*

1. Self-duality. *If A is positive definite, $c = 0$, and if we choose $B = A$, then in view of* (25) *we have $x^k = y^k$ for all k, and hence Algorithms 1 and 2 coincide. In this case, Algorithm 2 can be described as* self-dual.

2. Space of iterates. *A direct consequence of the correspondence between the dual and primal iterates* (25) *is the following simple observation (a generalized version of this, which we prove later as Lemma 5.1, will be used in the proof of Theorem 1.1): Choose $y^0 \in \mathbb{R}^m$ and let $x^0 = c + B^{-1}A^\top y^0$. Then the iterates $\{x^k\}$ of Algorithm 2 are of the form $x^k = c + B^{-1}A^\top y^k$ for some $y^k \in \mathbb{R}^m$.*

3. Starting point. *While we have defined the primal iterates of Algorithm 2 via a linear transformation of the dual iterates—see* (25)*—we* can, *in principle, choose x^0 arbitrarily, thus breaking the primal-dual connection which helped us to define the method. In particular, we can choose x^0 in such a way that there does not exist y^0 for which $x^0 = c + B^{-1}A^\top y^0$. As is clear from Theorem 1.1, in this case the iterates $\{x^k\}$ will not converge to x^*, but to $x^* + t$, where t is the projection of $x^0 - c$ onto the nullspace of A.*

It turns out that Algorithm 2 is equivalent to the *sketch-and-project* method (26) of Gower and Richtárik [14]:

$$x^{k+1} = \arg\min_x \left\{ \|x - x^k\|_B \ : \ S^\top A x = S^\top b \right\}, \tag{26}$$

where S is a random matrix drawn in an i.i.d. fashion from a fixed distribution, just as in this work. In this method, the "complicated" system $Ax = b$ is first replaced by its sketched version $S^\top Ax = S^\top b$, the solution space of which contains all solutions of the original system. If S has a few columns only, this system will be small and easy to solve. Then, progress is made by projecting the last iterate onto the sketched system.

We now briefly comment on the relationship between [14] and our work.

- **Dual nature of sketch-and-project.** It was shown in [14] that Algorithm 2 is equivalent to the sketch-and-project method. In fact, the authors of [14] provide five additional equivalent formulations of sketch-and-project, with Algorithm 2 being one of them. Here we show that their method can be seen as a primal process associated with SDA, which is a new method operating in the dual. By observing this, we uncover a hidden dual nature of the sketch-and-project method. For instance, this allows us to formulate and prove Theorem 1.2. No such results appear in [14].

- **No assumptions on the system matrix.** In [14] the authors only studied the convergence of the primal iterates $\{x^k\}$, establishing a (much) weaker variant of Theorem 1.1. Indeed, convergence was only established in the case when A has full column rank. In this work, we lift this assumption completely and hence establish complexity results in the general case.

- **Convergence to a shifted point.** As we show in Theorem 1.1, Algorithm 2 converges to $x^* + t$, where t is the projection of $x^0 - c$ onto $\mathbf{Null}\,(A)$. Hence, in general, the method does not converge to the optimal solution x^*. This is not an issue if A is of full column rank—an assumption used in the analysis in [14]—since then $\mathbf{Null}\,(A)$ is trivial and hence $t = 0$. As long as $x^0 - c$ lies in $\mathbf{Range}\left(B^{-1}A^\top\right)$, however, we have $x^k \to x^*$. This can be easily enforced (for instance, we can choose $x^0 = c$).

2.3 Relating the quality of the dual and primal iterates

The following simple but insightful result (mentioned in the introduction) relates the "quality" of a dual vector y with that of its primal counterpart, $x(y)$. It says that the dual suboptimality of y in terms of function values is equal to the primal suboptimality of $x(y)$ in terms of distance.

Proposition 2.2. *Let y^* be any dual optimal point and $y \in \mathbb{R}^m$. Then*

$$D(y^*) - D(y) = \tfrac{1}{2}\|x(y^*) - x(y)\|_B^2.$$

Proof: Straightforward calculation shows that

$$\begin{aligned}
D(y^*) - D(y) &\overset{(3)}{=} (b - Ac)^\top (y^* - y) - \tfrac{1}{2}(y^*)^\top AB^{-1}A^\top y^* + \tfrac{1}{2}y^\top AB^{-1}A^\top y \\
&\overset{(23)}{=} (y^*)^\top AB^{-1}A^\top (y^* - y) - \tfrac{1}{2}(y^*)^\top AB^{-1}A^\top y^* + \tfrac{1}{2}y^\top AB^{-1}A^\top y \\
&= \tfrac{1}{2}(y - y^*)^\top AB^{-1}A^\top (y - y^*) \\
&\overset{(21)}{=} \tfrac{1}{2}\|x(y) - x(y^*)\|_B^2.
\end{aligned}$$

□

Applying this result to sequence $\{(x^k, y^k)\}$ of dual iterates produced by SDA and their corresponding primal images, as defined in (25), we get the identity:

$$D(y^*) - D(y^k) = \tfrac{1}{2}\|x^k - x^*\|_B^2.$$

Therefore, *dual convergence in function values $D(y^k)$ is equivalent to primal convergence in iterates x^k*. Furthermore, a direct computation leads to the following formula for the *duality gap*:

$$P(x^k) - D(y^k) \overset{(25)}{=} (AB^{-1}A^\top y^k + Ac - b)^\top y^k = -(\nabla D(y^k))^\top y^k. \tag{27}$$

Note that computing the gap is significantly more expensive than the cost of a single iteration (in the interesting regime when the number of columns of S is small). Hence, evaluation of the duality gap should generally be avoided. If it is necessary to be certain about the quality of a solution however, the above formula will be useful. The gap should then be computed from time to time only, so that this extra work does not significantly slow down the iterative process.

3 Discrete Distributions

Both the SDA algorithm and its primal counterpart are generic in the sense that the distribution $\mathcal{D}$ is not specified beyond assuming that the matrix H defined in (9) is well defined and nonsingular. In this section we shall first characterize finite discrete distributions for which H is nonsingular. We then give a few examples of algorithms based on such distributions, and comment on our complexity results in more detail.

3.1 Nonsingularity of H for finite discrete distributions

For simplicity, we shall focus on *finite discrete* distributions $\mathcal{D}$. That is, we set $S = S_i$ with probability $p_i > 0$, where $S_1, \ldots, S_r$ are fixed matrices (each with m rows). The next theorem gives a necessary and sufficient condition for the matrix H defined in (9) to be nonsingular.

Theorem 3.1. *Let $\mathcal{D}$ be a finite discrete distribution, as described above. Then H is nonsingular if and only if*

$$\mathbf{Range}\left([S_1 S_1^\top A, \cdots, S_r S_r^\top A]\right) = \mathbb{R}^m.$$

Proof: Let $K_i = S_i^\top A B^{-1/2}$. In view of the identity $\left(K_i K_i^\top\right)^\dagger = (K_i^\dagger)^\top K_i^\dagger$, we can write

$$H \overset{(9)}{=} \sum_{i=1}^r H_i,$$

where $H_i = p_i S_i (K_i^\dagger)^\top K_i^\dagger S_i^\top$. Since H_i are symmetric positive semidefinite, so is H. Now, it is easy to check that $y^\top H_i y = 0$ if and only if $y \in \mathbf{Null}\,(H_i)$ (this holds for any symmetric positive semidefinite H_i). Hence, $y^\top H y = 0$ if and only if $y \in \cap_i \mathbf{Null}\,(H_i)$ and thus H is positive definite if and only if

$$\bigcap_i \mathbf{Null}\,(H_i) = \{0\}. \tag{28}$$

In view of Lemma 10.1, $\mathbf{Null}\left(H_i\right) = \mathbf{Null}\left(\sqrt{p_i}K_i^\dagger S_i^\top\right) = \mathbf{Null}\left(K_i^\dagger S_i^\top\right)$. Now, $y \in \mathbf{Null}\left(K_i^\dagger S_i^\top\right)$ if and only of $S_i^\top y \in \mathbf{Null}\left(K_i^\dagger\right) = \mathbf{Null}\left(K_i^\top\right) = \mathbf{Null}\left(A^\top S_i\right)$. Hence, $\mathbf{Null}\left(H_i\right) = \mathbf{Null}\left(A^\top S_i S_i^\top\right)$, which means that (28) is equivalent to $\mathbf{Null}\left([S_1S_1^\top A, \cdots, S_rS_r^\top A]^\top\right) = \{0\}$. □

We have the following corollary.[5]

Corollary 3.1. *Assume that $S_i^\top A$ has full row rank for all i and that $\bar{S} \overset{def}{=} [S_1, \ldots, S_r]$ is of full row rank. Then H is nonsingular.*

We now give a few illustrative examples:

1. *Coordinate vectors.* Let $S_i = e_i$ (i^{th} unit coordinate vector) for $i = 1, 2, \ldots, r = m$. In this case, $\bar{S} = [S_1, \ldots, S_m]$ is the identity matrix in $\mathbb{R}^m$, and $S_i^\top A$ has full row rank for all i as long as the rows of A are all nonzero. By Corollary 3.1, H is positive definite.

2. *Submatrices of the identity matrix.* We can let S be a random column submatrix of the $m \times m$ identity matrix I. There are $2^m - 1$ such potential submatrices, and we choose $1 \leq r \leq 2^m - 1$. As long as we choose $S_1, \ldots, S_r$ in such a way that each column of I is represented in some matrix S_i, the matrix $\bar{S}$ will have full row rank. Furthermore, if $S_i^\top A$ has full row rank for all i, then by the above corollary, H is nonsingular. Note that if the row rank of A is r, then the matrices S_i selected by the above process will necessarily have at most r columns.

3. *Count sketch and Count-min sketch.* Many other "sketching" matrices S can be employed within SDA, including the count sketch [4] and the count-min sketch [5]. In our context (recall that we sketch with the transpose of S), S is a count-sketch matrix (resp. count-min sketch) if it is assembled from random columns of $[I, -I]$ (resp I), chosen uniformly with replacement, where I is the $m \times m$ identity matrix.

3.2 Randomized Kaczmarz as the primal process associated with randomized coordinate ascent

Let $B = I$ (the identity matrix). The primal problem then becomes

$$\begin{array}{rl} \text{minimize} & P(x) \overset{\text{def}}{=} \frac{1}{2}\|x - c\|^2 \\ \text{subject to} & Ax = b \\ & x \in \mathbb{R}^n. \end{array}$$

and the dual problem is

$$\begin{array}{rl} \text{maximize} & D(y) \overset{\text{def}}{=} (b - Ac)^\top y - \frac{1}{2} y^\top AA^\top y \\ \text{subject to} & y \in \mathbb{R}^m. \end{array}$$

[5] We can also prove the corollary directly as follows: The first assumption implies that $S_i^\top AB^{-1}A^\top S_i$ is invertible for all i and that $V \overset{\text{def}}{=} \text{Diag}\left(p_i^{1/2}(S_i^\top AB^{-1}A^\top S_i)^{-1/2}\right)$ is nonsingular. It remains to note that

$$H \overset{(9)}{=} \mathbf{E}\left[S\left(S^\top AB^{-1}A^\top S\right)^{-1}S^\top\right] = \sum_i p_i S_i \left(S_i^\top AB^{-1}A^\top S_i\right)^{-1} S_i^\top = \bar{S}V^2\bar{S}^\top.$$

Dual iterates. Let us choose $S = e^i$ (unit coordinate vector in $\mathbb{R}^m$) with probability $p_i > 0$ (to be specified later). The SDA method (Algorithm 1) then takes the form

$$\boxed{y^{k+1} = y^k + \frac{b_i - A_i c - A_{i:} A^\top y^k}{\|A_{i:}\|^2} e_i} \tag{29}$$

This is the randomized coordinate ascent method applied to the dual problem. In the form popularized by Nesterov [36], it takes the form

$$y^{k+1} = y^k + \frac{e_i^\top \nabla D(y^k)}{L_i} e_i,$$

where $e_i^\top \nabla D(y^k)$ is the ith partial derivative of D at y^k and $L_i > 0$ is the Lipschitz constant of the ith partial derivative, i.e., constant for which the following inequality holds for all $\lambda \in \mathbb{R}$:

$$|e_i^\top \nabla D(y + \lambda e_i) - e_i^\top \nabla D(y)| \leq L_i |\lambda|. \tag{30}$$

It can be easily verified that (30) holds with $L_i = \|A_{i:}\|^2$ and that $e_i^\top \nabla D(y^k) = b_i - A_{i:} c - A_{i:} A^\top y^k$.

Primal iterates. The associated primal iterative process (Algorithm 2) takes the form

$$\boxed{x^{k+1} = x^k - \frac{A_{i:} x^k - b_i}{\|A_{i:}\|^2} A_{i:}^\top} \tag{31}$$

This is the randomized Kaczmarz method of Strohmer and Vershynin [55].

The rate. Let us now compute the rate ρ as defined in (13). It will be convenient, but *not* optimal, to choose the probabilities via

$$p_i = \frac{\|A_{i:}\|_2^2}{\|A\|_F^2}, \tag{32}$$

where $\|\cdot\|_F$ denotes the Frobenius norm (we assume that A does not contain any zero rows). Since

$$H \overset{(9)}{=} \mathbf{E}\left[S\left(S^\top A A^\top S\right)^\dagger S^\top\right] = \sum_{i=1}^m p_i \frac{e_i e_i^\top}{\|A_{i:}\|^2} \overset{(32)}{=} \frac{1}{\|A\|_F^2} I,$$

we have

$$\rho = 1 - \lambda_{\min}^+\left(A^\top H A\right) = 1 - \frac{\lambda_{\min}^+\left(A^\top A\right)}{\|A\|_F^2}. \tag{33}$$

In general, the rate ρ is a function of the probabilities p_i. The inverse problem: "How to set the probabilities so that the rate is optimized?" is difficult. If A is of full column rank, however, it leads to a semidefinite program [14].

Furthermore, if $r = \mathbf{Rank}(A)$, then in view of (14), the rate is bounded as

$$1 - \frac{1}{r} \leq \rho < 1.$$

Assume that A is of rank $r = 1$ and let $A = uv^\top$. Then $A^\top A = (u^\top u) v v^\top$, and hence this matrix is also of rank 1. Therefore, $A^\top A$ has a single nonzero eigenvalue, which is equal its trace. Therefore, $\lambda_{\min}^+(A^\top A) = \mathbf{Tr}\left(A^\top A\right) = \|A\|_F^2$ and hence $\rho = 0$. Note that the rate ρ reaches its lower bound and the method converges in one step.

Remarks. For randomized coordinate ascent applied to (non-strongly) concave quadratics, rate (33) has been established by Leventhal and Lewis [21]. However, to the best of our knowledge, this is the first time this rate has also been established for the randomized Kaczmarz method. We do not only prove this, but show that this is because the iterates of the two methods are linked via a linear relationship. In the $c = 0, B = I$ case, and for row-normalized matrix A, this linear relationship between the two methods was recently independently observed by Wright [61]. While all linear complexity results for RK we are aware of require full rank assumptions, there exist nonstandard variants of RK which do not require such assumptions, one example being the asynchronous parallel version of RK studied by Liu, Wright and Sridhar [24]. Finally, no results of the type (16) (primal suboptimality) and (17) (duality gap) previously existed for these methods in the literature.

3.3 Randomized block Kaczmarz is the primal process associated with randomized Newton

Let $B = I$, so that we have the same pair of primal dual problems as in Section 3.2.

Dual iterates. Let us now choose S to be a random column submatrix of the $m \times m$ identity matrix I. That is, we choose a random subset $C \subset \{1, 2, \ldots, m\}$ and then let S be the concatenation of columns $j \in C$ of I. We shall write $S = I_C$. Let p_C be the probability that $S = I_C$. Assume that for each $j \in \{1, \ldots, m\}$ there exists C with $j \in C$ such that $p_C > 0$. Such a random set is called *proper* [44].

The SDA method (Algorithm 1) then takes the form

$$\boxed{y^{k+1} = y^k + I_C \lambda^k} \tag{34}$$

where λ^k is chosen so that the dual objective is maximized (see (19)). This is a variant of the *randomized Newton method* studied in [44]. By examining (19), we see that this method works by "inverting" randomized submatrices of the "Hessian" $AA^\top$. Indeed, λ^k is in each iteration computed by solving a system with the matrix $I_C^\top AA^\top I_C$. This is the random submatrix of $AA^\top$ corresponding to rows and columns in C.

Primal iterates. In view of the equivalence between Algorithm 2 and the sketch-and-project method (26), the primal iterative process associated with the randomized Newton method has the form

$$\boxed{x^{k+1} = \arg\min_x \left\{ \|x - x^k\| \ : \ I_C^\top A x = I_C^\top b \right\}} \tag{35}$$

This method is a variant of the *randomized block Kaczmarz* method of Needell [33]. The method proceeds by projecting the last iterate x^k onto a subsystem of $Ax = b$ formed by equations indexed by the set C.

The rate. Provided that H is nonsingular, the shared rate of the randomized Newton and randomized block Kaczmarz methods is

$$\rho = 1 - \lambda_{\min}^+ \left(A^\top \mathbf{E} \left[I_C \left(I_C^\top AA^\top I_C \right)^\dagger I_C^\top \right] A \right).$$

Qu et al [44] study the randomized Newton method for the problem of minimizing a smooth strongly convex function and prove linear convergence. In particular, they study the above rate in the case when $AA^\top$ is positive definite. Here we show that linear converges also holds for *weakly* convex quadratics (as long as H is nonsingular).

An interesting feature of the randomized Newton method, established in [44], is that when viewed as a family of methods indexed by the size $\tau = |C|$, it enjoys superlinear speedup in τ. That is, as τ increases by some factor, the iteration complexity drops by a factor that is at least as large. It is possible to conduct a similar study in our setting with a possibly singular matrix $AA^\top$, but such a study is not trivial and we therefore leave it for future research.

3.4 Self-duality for positive definite A

If A is positive definite, then we can choose $B = A$. As mentioned before, in this setting SDA is self-dual: $x^k = y^k$ for all k. The primal problem then becomes

$$\begin{array}{rl} \text{minimize} & P(x) \overset{\text{def}}{=} \frac{1}{2} x^\top A x \\ \text{subject to} & Ax = b \\ & x \in \mathbb{R}^n. \end{array}$$

and the dual problem becomes

$$\begin{array}{rl} \text{maximize} & D(y) \overset{\text{def}}{=} b^\top y - \frac{1}{2} y^\top A y \\ \text{subject to} & y \in \mathbb{R}^m. \end{array}$$

Note that the primal objective function does not play any role in determining the solution; indeed, the feasible set contains a single point only: $A^{-1}b$. However, it does affect the iterative process.

Primal and dual iterates. As before, let us choose $S = e^i$ (unit coordinate vector in $\mathbb{R}^m$) with probability $p_i > 0$, where the probabilities p_i are arbitrary. Then both the primal and the dual iterates take the form

$$\boxed{y^{k+1} = y^k - \frac{A_{i:}y^k - b_i}{A_{ii}} e_i}$$

This is the randomized coordinate ascent method applied to the dual problem.

The rate. If we choose $p_i = A_{ii}/\mathbf{Tr}\,(A)$, then

$$H = \mathbf{E}\left[S\left(S^\top A S\right)^\dagger S^\top\right] = \frac{I}{\mathbf{Tr}\,(A)},$$

whence

$$\rho \overset{(13)}{=} 1 - \lambda_{\min}^+\left(A^{1/2} H A^{1/2}\right) = 1 - \frac{\lambda_{\min}(A)}{\mathbf{Tr}\,(A)}.$$

It is known that for this problem, randomized coordinate descent applied to the dual problem, with this choice of probabilities, converges with this rate [21].

4 Application: Randomized Gossip Algorithms

In this section we apply our method and results to the distributed consensus (averaging) problem.

Let (V, E) be a connected network with $|V| = n$ nodes and $|E| = m$ edges, where each edge is an unordered pair $\{i, j\} \in E$ of distinct nodes. Node $i \in V$ stores a private value $c_i \in \mathbb{R}$. The goal of a "distributed consensus problem" is for the network to compute the average of these private values in a distributed fashion [2, 39]. This means that the exchange of information can only occur along the edges of the network.

The nodes may represent people in a social network, with edges representing friendship and private value representing certain private information, such as salary. The goal would be to compute the average salary via an iterative process where only friends are allowed to exchange information. The nodes may represent sensors in a wireless sensor network, with an edge between two sensors if they are close to each other so that they can communicate. Private values represent measurements of some quantity performed by the sensors, such as the temperature. The goal is for the network to compute the average temperature.

4.1 Consensus as a projection problem

We now show how one can model the consensus (averaging) problem in the form (2). Consider the projection problem

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}\|x - c\|^2 \\ \text{subject to} & x_1 = x_2 = \cdots = x_n, \end{array} \tag{36}$$

and note that the optimal solution x^* must necessarily satisfy

$$x_i^* = \bar{c} \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} c_i,$$

for all i. There are many ways in which the constraint forcing all coordinates of x to be equal can be represented in the form of a linear system $Ax = b$. Here are some examples:

1. *Each node is equal to all its neighbours.* Let the equations of the system $Ax = b$ correspond to constraints

   $$x_i = x_j,$$

   for $\{i, j\} \in E$. That is, we are enforcing all pairs of vertices joined by an edge to have the same value. Each edge $e \in E$ can be written in two ways: $e = \{i, j\}$ and $e = \{j, i\}$, where i, j are the incident vertices. In order to avoid duplicating constraints, for each edge $e \in E$ we use $e = (i, j)$ to denote an arbitrary but fixed order of its incident vertices i, j. We then let $A \in \mathbb{R}^{m \times n}$ and $b = 0 \in \mathbb{R}^m$, where

   $$(A_{e:})^\top = f_i - f_j, \tag{37}$$

   and where $e = (i, j) \in E$, f_i (resp. f_j) is the i^{th} (resp. j^{th}) unit coordinate vector in $\mathbb{R}^n$. Note that the constraint $x_i = x_j$ is represented only once in the linear system. Further, note that the matrix

   $$L = A^\top A \tag{38}$$

is the *Laplacian* matrix of the graph (V, E):

$$L_{ij} = \begin{cases} d_i & i = j \\ -1 & i \neq j,\ (i,j) \in E \\ 0 & \text{otherwise,} \end{cases}$$

where d_i is the degree of node i.

2. *Each node is the average of its neighbours.* Let the equations of the system $Ax = b$ correspond to constraints

$$x_i = \frac{1}{d_i} \sum_{j \in N(i)} x_j,$$

for $i \in V$, where $N(i) \stackrel{\text{def}}{=} \{j \in V \ : \ \{i,j\} \in E\}$ is the set of neighbours of node i and $d_i \stackrel{\text{def}}{=} |N(i)|$ is the degree of node i. That is, we require that the values stored at each node are equal to the average of the values of its neighbours. This corresponds to the choice $b = 0$ and

$$(A_{i:})^\top = f_i - \frac{1}{d_i} \sum_{j \in N(i)} f_j. \tag{39}$$

Note that $A \in \mathbb{R}^{n \times n}$.

3. *Spanning subgraph.* Let (V, E') be any connected subgraph of (V, E). For instance, we can choose a spanning tree. We can now apply any of the 2 models above to this new graph and either require $x_i = x_j$ for all $\{i,j\} \in E'$, or require the value x_i to be equal to the average of the values x_j for all neighbours j of i in (V, E').

Clearly, the above list does not exhaust the ways in which the constraint $x_1 = \cdots = x_n$ can be modeled as a linear system. For instance, we could build the system from constraints such as $x_1 = x_2 + x_4 - x_3$, $x_1 = 5x_2 - 4x_7$ and so on.

Different representations of the constraint $x_1 = \cdots = x_n$, in combination with a choice of $\mathcal{D}$, will lead to a wide range of specific algorithms for the consensus problem (36). Some (but not all) of these algorithms will have the property that communication only happens along the edges of the network, and these are the ones we are interested in. The number of combinations is very vast. We will therefore only highlight two options, with the understanding that based on this, the interested reader can assemble other specific methods as needed.

4.2 Model 1: Each node is equal to its neighbours

Let $b = 0$ and A be as in (37). Let the distribution $\mathcal{D}$ be defined by setting $S = e_i$ with probability $p_i > 0$, where e_i is the i^{th} unit coordinate vector in $\mathbb{R}^m$. We have $B = I$, which means that Algorithm 2 is the randomized Kaczmarz (RK) method (31) and Algorithm 1 is the randomized coordinate ascent method (29).

Let us take $y^0 = 0$ (which means that $x^0 = c$), so that in Theorem 1.1 we have $t = 0$, and hence $x^k \to x^*$. The particular choice of the starting point $x^0 = c$ in the primal process has a very tangible meaning: for all i, node i initially knows value c_i. The primal iterative process will dictate how the local values are modified in an iterative fashion so that eventually all nodes contain the optimal value $x_i^* = \bar{c}$.

Primal method. In view of (37), for each edge $e = (i, j) \in E$, we have $\|A_{e:}\|^2 = 2$ and $A_{e:}x^k = x_i^k - x_j^k$. Hence, if the edge e is selected by the RK method, (31) takes the specific form

$$\boxed{x^{k+1} = x^k - \frac{x_i^k - x_j^k}{2}(f_i - f_j)} \tag{40}$$

From (40) we see that only the i^{th} and j^{th} coordinates of x^k are updated, via

$$x_i^{k+1} = x_i^k - \frac{x_i^k - x_j^k}{2} = \frac{x_i^k + x_j^k}{2}$$

and

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

Note that in each iteration of RK, a random edge is selected, and the nodes on this edge replace their local values by their average. This is a basic variant of the *randomized gossip* algorithm [2, 67].

Invariance. Let f be the vector of all ones in $\mathbb{R}^n$ and notice that from (40) we obtain $f^\top x^{k+1} = f^\top x^k$ for all k. This means that for all $k \geq 0$ we have the invariance property:

$$\sum_{i=1}^n x_i^k = \sum_{i=1}^n c_i. \tag{41}$$

Insights from the dual perspective. We can now bring new insight into the randomized gossip algorithm by considering the dual iterative process. The dual method (29) maintains weights y^k associated with the edges of E via the process:

$$y^{k+1} = y^k - \frac{A_{e:}(c - A^\top y^k)}{2} e_e,$$

where e is a randomly selected edge. Hence, only the weight of a single edge is updated in each iteration. At optimality, we have $x^* = c + A^\top y^*$. That is, for each i

$$\delta_i \overset{\text{def}}{=} \bar{c} - c_i = x_i^* - c_i = (A^\top y^*)_i = \sum_{e \in E} A_{ei} y_e^*,$$

where δ_i is the correction term which needs to be added to c_i in order for node i to contain the value $\bar{c}$. From the above we observe that these correction terms are maintained by the dual method as an inner product of the i^{th} column of A and y^k, with the optimal correction being $\delta_i = A_{:i}^\top y^*$.

Rate. Both Theorem 1.1 and Theorem 1.2 hold, and hence we automatically get several types of convergence for the randomized gossip method. In particular, to the best of our knowledge, no primal-dual type of convergence exist in the literature. Equation (27) gives a stopping criterion certifying convergence via the duality gap, which is also new.

In view of (33) and (38), and since $\|A\|_F^2 = 2m$, the convergence rate appearing in all these complexity results is given by

$$\rho = 1 - \frac{\lambda_{\min}^+(L)}{2m},$$

where L is the Laplacian of (V, E). While it is know that the Laplacian is singular, the rate depends on the smallest nonzero eigenvalue. This means that the number of iterations needed to output an ϵ-solution in expectation scales as $O\big(\big(2m/\lambda^+_{\min}(L)\big)\log(1/\epsilon)\big)$, i.e., linearly with the number of edges.

4.3 Model 2: Each node is equal to the average of its neighbours

Let A be as in (39) and $b = 0$. Let the distribution $\mathcal{D}$ be defined by setting $S = f_i$ with probability $p_i > 0$, where f_i is the i^{th} unit coordinate vector in $\mathbb{R}^n$. Again, we have $B = I$, which means that Algorithm 2 is the randomized Kaczmarz (RK) method (31) and Algorithm 1 is the randomized coordinate ascent method (29). As before, we choose $y^0 = 0$, whence $x^0 = c$.

Primal method. Observe that $\|A_{i:}\|^2 = 1 + 1/d_i$. The RK method (31) applied to this formulation of the problem takes the form

$$\boxed{x^{k+1} = x^k - \frac{x^k_i - \frac{1}{d_i}\sum_{j\in N(i)} x^k_j}{1 + 1/d_i}\left(f_i - \frac{1}{d_i}\sum_{j\in N(i)} f_j\right)} \tag{42}$$

where i is chosen at random. This means that only coordinates in $i \cup N(i)$ get updated in such an iteration, the others remain unchanged. For node i (coordinate i), this update is

$$x^{k+1}_i = \frac{1}{d_i + 1}\left(x^k_i + \sum_{j\in N(i)} x^k_j\right). \tag{43}$$

That is, the updated value at node i is the average of the values of its neighbours and the previous value at i. From (42) we see that the values at nodes $j \in N(i)$ get updated as follows:

$$x^{k+1}_j = x^k_j + \frac{1}{d_i + 1}\left(x^k_i - \frac{1}{d_i}\sum_{j'\in N(i)} x^k_{j'}\right). \tag{44}$$

Invariance. Let f be the vector of all ones in $\mathbb{R}^n$ and notice that from (42) we obtain

$$f^\top x^{k+1} = f^\top x^k - \frac{x^k_i - \frac{1}{d_i}\sum_{j\in N(i)} x^k_j}{1 + 1/d_i}\left(1 - \frac{d_i}{d_i}\right) = f^\top x^k,$$

for all k.

It follows that the method satisfies the invariance property (41).

Rate. The method converges with the rate ρ given by (33), where A is given by (39). If (V, E) is a complete graph (i.e., $m = \frac{n(n-1)}{2}$), then $L = \frac{(n-1)^2}{n}A^\top A$ is the Laplacian. In that case, $\|A\|_F^2 = \mathbf{Tr}\left(A^\top A\right) = \frac{n}{(n-1)^2}\mathbf{Tr}\left(L\right) = \frac{n}{(n-1)^2}\sum_i d_i = \frac{n^2}{n-1}$ and hence

$$\rho \overset{(39)}{=} 1 - \frac{\lambda^+_{\min}(A^\top A)}{\|A\|_F^2} = 1 - \frac{\frac{n}{(n-1)^2}\lambda^+_{\min}(L)}{\frac{n^2}{n-1}} = 1 - \frac{\lambda^+_{\min}(L)}{2m}.$$

5 Proof of Theorem 1.1

In this section we prove Theorem 1.1. We proceed as follows: in Section 5.1 we characterize the space in which the iterates move, in Section 5.2 we establish a certain key technical inequality, in Section 5.3 we establish convergence of iterates, in Section 5.4 we derive a rate for the residual and finally, in Section 5.5 we establish the lower bound on the convergence rate.

5.1 An error lemma

The following result describes the space in which the iterates move. It is an extension of the observation in Remark 2.2 to the case when x^0 is chosen arbitrarily.

Lemma 5.1. *Let the assumptions of Theorem 1.1 hold. For all $k \geq 0$ there exists $w^k \in \mathbb{R}^m$ such that $e^k \overset{def}{=} x^k - x^* - t = B^{-1}A^\top w^k$.*

Proof: We proceed by induction. Since by definition, t is the projection of $x^0 - c$ onto $\mathbf{Null}\,(A)$ (see (56)), applying Lemma 10.2 we know that $x^0 - c = s + t$, where $s = B^{-1}A^\top \hat{y}^0$ for some $\hat{y}^0 \in \mathbb{R}^m$. Moreover, in view of (22), we know that $x^* = c + B^{-1}A^\top y^*$, where y^* is any dual optimal solution. Hence,

$$e^0 = x^0 - x^* - t = B^{-1}A^\top(\hat{y}^0 - y^*).$$

Assuming the relationship holds for k, we have

$$\begin{aligned} e^{k+1} &= x^{k+1} - x^* - t \\ &\overset{\text{(Alg 2)}}{=} \left[x^k - B^{-1}A^\top S(S^\top AB^{-1}A^\top S)^\dagger S^\top(Ax^k - b)\right] - x^* - t \\ &= \left[x^* + t + B^{-1}A^\top w^k - B^{-1}A^\top S(S^\top AB^{-1}A^\top S)^\dagger S^\top(Ax^k - b)\right] - x^* - t \\ &= B^{-1}A^\top w^{k+1}, \end{aligned}$$

where $w^{k+1} = w^k - S(S^\top AB^{-1}A^\top S)^\dagger S^\top(Ax^k - b)$. □

5.2 A key inequality

The following inequality is of key importance in the proof of the main theorem.

Lemma 5.2. *Let $0 \neq W \in \mathbb{R}^{m\times n}$ and $G \in \mathbb{R}^{m\times m}$ be symmetric positive definite. Then the matrix $W^\top GW$ has a positive eigenvalue, and the following inequality holds for all $y \in \mathbb{R}^m$:*

$$y^\top WW^\top GWW^\top y \geq \lambda^+_{\min}(W^\top GW)\|W^\top y\|^2. \tag{45}$$

Furthermore, this bound is tight.

Proof: Fix arbitrary $y \in \mathbb{R}^m$. By Lemma 10.1, $W^\top y \in \mathbf{Range}\left(W^\top GW\right)$. Since, W is nonzero, the positive semidefinite matrix $W^\top GW$ is also nonzero, and hence it has a positive eigenvalue. Hence, $\lambda^+_{\min}(W^\top GW)$ is well defined. Let $\lambda^+_{\min}(W^\top GW) = \lambda_1 \leq \cdots \leq \lambda_\tau$ be the positive eigenvalues of $W^\top GW$, with associated orthonormal eigenvectors $q_1, \ldots, q_\tau$. We thus have

$$W^\top GW = \sum_{i=1}^{\tau} \lambda_i q_i q_i^\top.$$

It is easy to see that these eigenvectors span $\mathbf{Range}\left(W^\top GW\right)$. Hence, we can write $W^\top y = \sum_{i=1}^{\tau} \alpha_i q_i$ and therefore

$$y^\top WW^\top GWW^\top y = \sum_{i=1}^{\tau} \lambda_i \alpha_i^2 \geq \lambda_1 \sum_{i=1}^{\tau} \alpha_i^2 = \lambda_1 \|W^\top y\|^2.$$

Furthermore this bound is tight, as can be seen by selecting y so that $W^\top y = q_1$. □

5.3 Convergence of the iterates

Subtracting $x^* + t$ from both sides of the update step of Algorithm 2, and letting

$$Z = Z_{S^\top A} \overset{(57)}{=} A^\top S(S^\top AB^{-1}A^\top S)^\dagger S^\top A,$$

we obtain the identity

$$x^{k+1} - (x^* + t) = (I - B^{-1}Z)(x^k - (x^* + t)), \tag{46}$$

where we used that $t \in \mathbf{Null}\,(A)$. Let

$$e^k = x^k - (x^* + t) \tag{47}$$

and note that in view of (9), $\mathbf{E}\,[Z] = A^\top HA$. Taking norms and expectations (in S) on both sides of (46) gives

$$\begin{aligned}
\mathbf{E}\left[\|e^{k+1}\|_B^2 \,|\, e^k\right] &= \mathbf{E}\left[\|(I - B^{-1}Z)e^k\|_B^2\right] \\
&\overset{\text{(Lemma 10.2, Equation (58))}}{=} \mathbf{E}\left[(e^k)^\top (B - Z)e^k\right] \\
&= \|e^k\|_B^2 - (e^k)^\top \mathbf{E}\,[Z]\, e^k \\
&= \|e^k\|_B^2 - (e^k)^\top A^\top HAe^k, \qquad (48)
\end{aligned}$$

where in the second step we have used (58) from Lemma 10.2 with A replaced by $S^\top A$. In view of Lemma 5.1, let $w^k \in \mathbb{R}^m$ be such that $e^k = B^{-1}A^\top w^k$. Thus

$$\begin{aligned}
(e^k)^\top A^\top HAe^k &= (w^k)^\top AB^{-1}A^\top HAB^{-1}A^\top w^k \\
&\overset{\text{(Lemma 5.2)}}{\geq} \lambda_{\min}^+(B^{-1/2}A^\top HAB^{-1/2}) \cdot \|B^{-1/2}A^\top w^k\|^2 \\
&= (1 - \rho) \cdot \|B^{-1}A^\top w^k\|_B^2 \qquad (49) \\
&= (1 - \rho) \cdot \|e^k\|_B^2, \qquad (50)
\end{aligned}$$

where we applied Lemma 5.2 with $W = AB^{-1/2}$ and $G = H$, so that $W^\top GW = B^{-1/2}A^\top HAB^{-1/2}$. Substituting (50) into (48) gives $\mathbf{E}\left[\|e^{k+1}\|_B^2 \,|\, e^k\right] \leq \rho \cdot \|e^k\|_B^2$. Using the tower property of expectations, we obtain the recurrence

$$\mathbf{E}\left[\|e^{k+1}\|_B^2\right] \leq \rho \cdot \mathbf{E}\left[\|e^k\|_B^2\right].$$

To prove (11) it remains to unroll the recurrence.

5.4 Convergence of the residual

We now prove (12). Letting $V_k = \|x^k - x^* - t\|_B^2$, we have

$$\begin{aligned}
\mathbf{E}\left[\|Ax^k - b\|_B\right] &= \mathbf{E}\left[\|A(x^k - x^* - t) + At\|_B\right] \\
&\leq \mathbf{E}\left[\|A(x^k - x^* - t)\|_B\right] + \|At\|_B \\
&\leq \|A\|_B \mathbf{E}\left[\sqrt{V_k}\right] + \|At\|_B \\
&\leq \|A\|_B \sqrt{\mathbf{E}\left[V_k\right]} + \|At\|_B \\
&\overset{(11)}{\leq} \|A\|_B \sqrt{\rho^k V_0} + \|At\|_B,
\end{aligned}$$

where in the step preceding the last one we have used Jensen's inequality.

5.5 Proof of the lower bound

Since $\mathbf{E}\left[Z\right] = A^\top H A$, using Lemma 10.1 with $G = H$ and $W = AB^{-1/2}$ gives

$$\mathbf{Range}\left(B^{-1/2}\mathbf{E}\left[Z\right]B^{-1/2}\right) = \mathbf{Range}\left(B^{-1/2}A^\top\right),$$

from which we deduce that

$$\begin{aligned}
\mathbf{Rank}\left(A\right) &= \dim\left(\mathbf{Range}\left(A^\top\right)\right) \\
&= \dim\left(\mathbf{Range}\left(B^{-1/2}A^\top\right)\right) \\
&= \dim\left(\mathbf{Range}\left(B^{-1/2}\mathbf{E}\left[Z\right]B^{-1/2}\right)\right) \\
&= \mathbf{Rank}\left(B^{-1/2}\mathbf{E}\left[Z\right]B^{-1/2}\right).
\end{aligned}$$

Hence, $\mathbf{Rank}\left(A\right)$ is equal to the number of nonzero eigenvalues of $B^{-1/2}\mathbf{E}\left[Z\right]B^{-1/2}$, from which we immediately obtain the bound

$$\mathbf{Tr}\left(B^{-1/2}\mathbf{E}\left[Z\right]B^{-1/2}\right) \;\geq\; \mathbf{Rank}\left(A\right)\,\lambda_{\min}^{+}(B^{-1/2}\mathbf{E}\left[Z\right]B^{-1/2}).$$

In order to obtain (14), it only remains to combine the above inequality with

$$\mathbf{E}\left[\mathbf{Rank}\left(S^\top A\right)\right] \overset{(60)}{=} \mathbf{E}\left[\mathbf{Tr}\left(B^{-1}Z\right)\right] = \mathbf{E}\left[\mathbf{Tr}\left(B^{-1/2}ZB^{-1/2}\right)\right] = \mathbf{Tr}\left(B^{-1/2}\mathbf{E}\left[Z\right]B^{-1/2}\right).$$

6 Proof of Theorem 1.2

In this section we prove Theorem 1.2. We dedicate a subsection to each of the three complexity bounds.

6.1 Dual suboptimality

Since $x^0 \in c + \mathbf{Range}\left(B^{-1}A^\top\right)$, we have $t = 0$ in Theorem 1.1, and hence (11) says that

$$\mathbf{E}\left[U_k\right] \leq \rho^k U_0. \tag{51}$$

It remains to apply Proposition 2.2, which says that $U_k = D(y^*) - D(y^k)$.

6.2 Primal suboptimality

Letting $U_k = \frac{1}{2}\|x^k - x^*\|_B^2$, we can write

$$\begin{aligned} P(x^k) - OPT &= \tfrac{1}{2}\|x^k - c\|_B^2 - \tfrac{1}{2}\|x^* - c\|_B^2 \\ &= \tfrac{1}{2}\|x^k - x^* + x^* - c\|_B^2 - \tfrac{1}{2}\|x^* - c\|_B^2 \\ &= \tfrac{1}{2}\|x^k - x^*\|_B^2 + (x^k - x^*)^\top B(x^* - c) \\ &\leq U_k + \|x^k - x^*\|_B \|B(x^* - c)\|_{B^{-1}} \\ &= U_k + \|x^k - x^*\|_B \|x^* - c\|_B \\ &= U_k + 2\sqrt{U_k}\sqrt{OPT}. \end{aligned} \tag{52}$$

By taking expectations on both sides of (52), and using Jensen's inequality, we therefore obtain

$$\mathbf{E}\left[P(x^k) - OPT\right] \leq \mathbf{E}\left[U_k\right] + 2\sqrt{OPT}\sqrt{\mathbf{E}\left[U_k\right]} \overset{(51)}{\leq} \rho^k U_0 + 2\rho^{k/2}\sqrt{OPT \times U_0},$$

which establishes the bound on primal suboptimality (16).

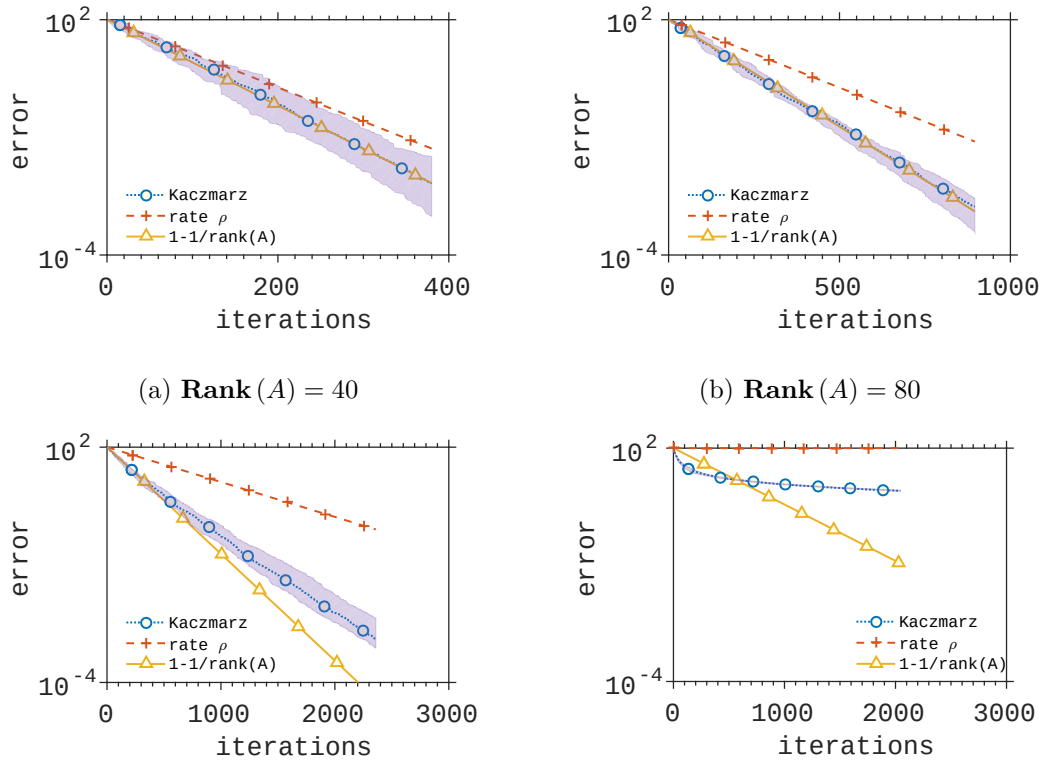
6.3 Duality gap

Having established rates for primal and dual suboptimality, the rate for the duality gap follows easily:

$$\begin{aligned} \mathbf{E}\left[P(x^k) - D(y^k)\right] &= \mathbf{E}\left[P(x^k) - OPT + OPT - D(y^k)\right] \\ &= \mathbf{E}\left[P(x^k) - OPT\right] + \mathbf{E}\left[OPT - D(y^k)\right] \\ &\overset{(15)+(16)}{=} 2\rho^k U_0 + 2\rho^{k/2}\sqrt{OPT \times U_0}. \end{aligned}$$

7 Numerical Experiments: Randomized Kaczmarz Method with Rank-Deficient System

To illustrate some of the novel aspects of our theory, we perform numerical experiments with the Randomized Kaczmarz method (31) (or equivalently the randomized coordinate ascent method applied to the dual problem (3)) and compare the empirical convergence to the convergence predicted by our theory. We test several randomly generated rank-deficient systems and compare the evolution of the empirical primal error $\|x^k - x^*\|_2^2/\|x^0 - x^*\|_2^2$ to the convergence dictated by the rate $\rho = 1 - \lambda_{\min}^+\left(A^\top A\right)/\|A\|_F^2$ given in (33) and the lower bound $1 - 1/\mathbf{Rank}\left(A\right) \leq \rho$. From Figure 1 we can see that the RK method converges despite the fact that the linear systems are rank deficient. While previous results do not guarantee that RK converges for rank-deficient matrices, our theory does as long as the system matrix has no zero rows. Furthermore, we observe that the lower the rank of the system matrix, the faster the convergence of the RK method, and moreover, the closer the empirical convergence is to the convergence dictated by the rate ρ and lower bound on ρ. In particular, on the low rank system in Figure 1a, the empirical convergence is very close to both the convergence dictated by ρ and the lower bound. While on the full rank system in Figure 1d, the convergence dictated by ρ and the lower bound on ρ are no longer an accurate estimate of the empirical convergence.



(a) $\mathbf{Rank}\,(A) = 40$ (b) $\mathbf{Rank}\,(A) = 80$

(c) $\mathbf{Rank}\,(A) = 160$ (d) $\mathbf{Rank}\,(A) = 300$

Figure 1: Synthetic MATLAB generated problems. Rank deficient matrix $A = \sum_{i=1}^{\mathbf{Rank}(A)} \sigma_i u_i v_i^T$ where $\sum_{i=1}^{300} \sigma_i u_i v_i^T = \texttt{rand}(300, 300)$ is an svd decomposition of a 300×300 uniform random matrix. We repeat each experiment ten times. The blue shaded region is the 90% percentile of relative error achieved in each iteration.

8 Conclusion

We have developed a versatile and powerful algorithmic framework for solving linear systems: *stochastic dual ascent (SDA)*. In particular, SDA finds the projection of a given point, in a fixed but arbitrary Euclidean norm, onto the solution space of the system. Our method is dual in nature, but can also be described in terms of primal iterates via a simple affine transformation of the dual variables. Viewed as a dual method, SDA belongs to a novel class of randomized optimization algorithms: it updates the current iterate by adding the product of a random matrix, drawn independently from a fixed distribution, and a vector. The update is chosen as the best point lying in the random subspace spanned by the columns of this random matrix.

While SDA is the first method of this type, particular choices for the distribution of the random matrix lead to several known algorithms: randomized coordinate descent [21] and randomized Kaczmarz [55] correspond to a discrete distribution over the columns of the identity matrix, randomized Newton method [44] corresponds to a discrete distribution over column submatrices of the identity matrix, and Gaussian descent [54] corresponds to the case when the random matrix is a Gaussian vector.

We equip the method with several complexity results with the same rate of exponential decay in expectation (aka linear convergence) and establish a tight lower bound on the rate. In particular, we prove convergence of primal iterates, dual function values, primal function values, duality gap and of the residual. The method converges under very weak conditions beyond consistency of the linear system. In particular, no rank assumptions on the system matrix are needed. For instance, randomized Kaczmarz method converges linearly as long as the system matrix contains no zero rows.

Further, we show that SDA can be applied to the distributed (average) consensus problem. We recover a standard randomized gossip algorithm as a special case, and show that its complexity is proportional to the number of edges in the graph and inversely proportional to the smallest nonzero eigenvalue of the graph Laplacian. Moreover, we illustrate how our framework can be used to obtain new randomized algorithms for the distributed consensus problem.

Our framework extends to several other problems in optimization and numerical linear algebra. For instance, one can apply it to develop new stochastic algorithms for computing the inverse of a matrix and obtain state-of-the art performance for inverting matrices of huge sizes.

9 Acknowledgements

The second author would like to thank Julien Hendrickx from Université catholique de Louvain for a discussion regarding randomized gossip algorithms.

References

[1] Jonathan M. Borwein and Adrian S. Lewis. *Convex analysis and nonlinear optimization*. Springer-Verlag New York, 2006.

[2] Stephen Boyd, Arpita Ghosh, and Devavrat Prabhakar Balajiand Shah. "Randomized Gossip Algorithms". In: *IEEE Transactions on Information Theory* 52.6 (2006), pp. 2508–2530.

[3] Joseph K. Bradley, Aapo Kyrola, Danny Bickson, and Carlos Guestrin. "Parallel Coordinate Descent for L1-Regularized Loss Minimization". In: *28th Int. Conf. on Machine Learning*. 2011.

[4] Moses Charikar, Kevin Chen, and Martin Farach-Colton. "Finding frequent items in data streams". In: *Proceedings of the 29th International Colloquium on Automata, Languages and Programming (ICALP)*. Springer-Verlag London, 2002, pp. 693–703.

[5] Graham Cormode and S. Muthukrishnan. "An improved data stream summary: the count-min sketch and its applications". In: *Journal of Algorithms* 55 (2005), pp. 29–38.

[6] Dominik Csiba, Zheng Qu, and Peter Richtárik. "Stochastic dual coordinate ascent with adaptive probabilities". In: *Proceedings of the 32nd International Conference on Machine Learning*. 2015.

[7] Liang Dai, Mojtaba Soltanalian, and Kristiaan Pelckmans. "On the Randomized Kaczmarz Algorithm". In: *IEEE Signal Processing Letters* 21.3 (2014), pp. 330–333. arXiv:`arXiv:1402.2863v1`.

[8] Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. "SAGA: A Fast Incremental Gradient Method With Support for Non-Strongly Convex Composite Objectives". In: *arXiv:1407.0202* (2014).

[9] Petros Drineas, Michael W. Mahoney, S. Muthukrishnan, and Tamás Sarlós. "Faster Least Squares Approximation". In: *Numerische Mathematik* 117.2 (2011), pp. 219–249. arXiv:0710.1435.

[10] Olivier Fercoq. "Parallel coordinate descent for the Adaboost problem". In: *Proc. of the International Conf. on Machine Learning and Applications*. 2013.

[11] Olivier Fercoq, Zheng Qu, Peter Richtárik, and Martin Takáč. "Fast distributed coordinate descent for minimizing non-strongly convex losses". In: *IEEE International Workshop on Machine Learning for Signal Processing* (2014).

[12] Olivier Fercoq and Peter Richtárik. "Accelerated, parallel and proximal coordinate descent". In: *SIAM Journal on Optimization* 25 (4 2015), pp. 1997–2023.

[13] Olivier Fercoq and Peter Richtárik. "Smooth minimization of nonsmooth functions by parallel coordinate descent". In: *arXiv:1309.5885* (2013).

[14] Robert Mansel Gower and Peter Richtárik. "Randomized Iterative Methods for Linear Systems". In: *SIAM Journal on Matrix Analysis and Applications* (2015).

[15] Mingyi Hong and Zhi-Quan Luo. "On the Linear Convergence of the Alternating Direction Method of Multipliers". In: *arXiv:1208.3922* (2012).

[16] Cho-Jui Hsieh, Kai-Wei Chang, Chih-Jen Lin, S Sathiya Keerthi, and S Sundararajan. "A dual coordinate descent method for large-scale linear SVM". In: *Proceedings of the 25th International Conference on Machine Learning*. 2008, pp. 408–415.

[17] Rie Johnson and Tong Zhang. "Accelerating Stochastic Gradient Descent using Predictive Variance Reduction". In: *Advances in Neural Information Processing Systems 26*. 2013, pp. 315–323.

[18] Jakub Konečný, Jie Lu, Peter Richtárik, and Martin Takáč. "Mini-Batch Semi-Stochastic Gradient Descent in the Proximal Setting". In: *IEEE Journal of Selected Topics in Signal Processing* (2016).

[19] Jakub Konečný and Peter Richtárik. "S2GD: Semi-stochastic gradient descent methods". In: *arXiv:1312.1666* (2014).

[20] Yin Tat Lee and Aaron Sidford. "Efficient Accelerated Coordinate Descent Methods and Faster Algorithms for Solving Linear Systems". In: *Proceedings - Annual IEEE Symposium on Foundations of Computer Science, FOCS* (2013), pp. 147–156. arXiv:1305.1922.

[21] Dennis Leventhal and Adrian S. Lewis. "Randomized Methods for Linear Constraints: Convergence Rates and Conditioning". In: *Mathematics of Operations Research* 35.3 (2010), pp. 641–654. eprint: http://mor.journal.informs.org/cgi/reprint/35/3/641.pdf.

[22] Qihang Lin, Zhaosong Lu, and Lin Xiao. "An accelerated proximal coordinate gradient method". In: *Advances in Neural Information Processing Systems 27*. 2014.

[23] Ji Liu and Stephen J Wright. "An accelerated randomized Kaczmarz algorithm". In: *Mathematics of Computation* 85.297 (2016), pp. 153–178.

[24] Ji Liu, Stephen J. Wright, and Sridhar Srikrishna. "An Asynchronous Parallel Randomized Kaczmarz Algorithm". In: *arXiv:1401.4780* (2014).

[25] Zhi-Quan Luo and Paul Tseng. "Error bounds and convergence analysis of feasible descent methods: a general approach". In: *Annals of Operations Research* 46 (1 1993), pp. 157–178.

[26] Anna Ma, Deanna Needell, Aaditya Ramdas, and N A Mar. "Convergence Properties of the Randomized Extended Gauss-Seidel and Kaczmarz methods". In: *arXiv:1503.08235* (2015), pp. 1–16.

[27] Chenxin Ma, Rachael Tappenden, and Martin Takáč. "Linear convergence of the randomized feasible descent method under the weak strong convexity assumption". In: *arXiv:1506.02530* (2015).

[28] Ion Necoara and Dragos Clipici. *Efficient parallel coordinate descent algorithm for convex optimization problems with separable constraints: application to distributed MPC*. Tech. rep. Politehnica University of Bucharest, 2012.

[29] Ion Necoara and Dragos Clipici. "Parallel coordinate descent methods for composite minimization: convergence analysis and error bounds". In: *SIAM Journal on Optimization* (2016).

[30] Ion Necoara and Andrei Patrascu. "A random coordinate descent algorithm for optimization problems with composite objective function and linear coupled constraints". In: *Computational Optimization and Applications* 57 (2 2014), pp. 307–337.

[31] Deana Needell. "Randomized Kaczmarz solver for noisy linear systems". In: *BIT* 50.2 (2010), pp. 395–403.

[32] Deanna Needell, Nathan Srebro, and Rachel Ward. "Stochastic gradient descent, weighted sampling, and the randomized Kaczmarz algorithm". In: *Mathematical Programming* (2015).

[33] Deanna Needell and Joel a. Tropp. "Paved with good intentions: Analysis of a randomized block Kaczmarz method". In: *Linear Algebra and Its Applications* 441.August (2012), pp. 199–221. arXiv:`arXiv:1208.3805v3`.

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

[35] Deanna Needell, Ran Zhao, and Anastasios Zouzias. "Randomized block Kaczmarz method with projection for solving least squares". In: *arXiv:1403.4192* (2014).

[36] Yurii Nesterov. "Efficiency of coordinate descent methods on huge-scale optimization problems". In: *SIAM Journal on Optimization* 22.2 (2012), pp. 341–362.

[37] Yurii Nesterov. *Random gradient-free minimization of convex functions*. Tech. rep. Université catholique de Louvain, CORE discussion paper, 2011.

[38] Feng Niu, Benjamin Recht, Christopher Ré, and Stephen Wright. "HOGWILD!: A Lock-Free Approach to Parallelizing Stochastic Gradient Descent". In: *Advances in Neural Information Processing Systems 24*. 2011.

[39] Alex Olshevsky and John N. Tsitsiklis. "Convergence Speed in Distributed Consensus and Averaging". In: *SIAM Journal on Control and Optimization* 48.1 (2009), 3355.

[40] Peter Oswald and Weiqi Zhou. "Convergence analysis for Kaczmarz-type methods in a Hilbert space framework". In: *Linear Algebra and its Applications* 478 (2015), pp. 131–161.

[41] Zheng Qu, Peter Richtárik, and Tong Zhang. "Quartz: Randomized dual coordinate ascent with arbitrary sampling". In: *Advances in Neural Information Processing Systems 28*. 2015.

[42] Zheng Qu and Richtárik. "Coordinate descent with arbitrary sampling I: algorithms and complexity". In: *arXiv:1412.8060* (2014).

[43] Zheng Qu and Richtárik. "Coordinate descent with arbitrary sampling II: expected separable overapproximation". In: *arXiv:1412.8063* (2014).

[44] Zheng Qu, Peter Richtárik, Martin Takáč, and Olivier Fercoq. "SDNA: Stochastic dual newton ascent for empirical risk minimization". In: *arXiv:1502.02268* (2015).

[45] Aaditya Ramdas. "Rows vs Columns for Linear Systems of Equations - Randomized Kaczmarz or Coordinate Descent ?" In: *arXiv:1406.5295* (2014). arXiv:arXiv:1406.5295v1.

[46] Peter Richtárik and Martin Takáč. "Distributed coordinate descent method for learning with big data". In: *Journal of Machine Learning Research* (2015).

[47] Peter Richtárik and Martin Takáč. "On optimal probabilities in stochastic coordinate descent methods". In: *Optimization Letters* (2015), pp. 1–11.

[48] Peter Richtárik and Martin Takáč. "Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function". In: *Mathematical Programming* 144 (2 2014), pp. 1–38.

[49] Peter Richtárik and Martin Takáč. "Parallel coordinate descent methods for big data optimization problems". In: *Mathematical Programming* (2015), pp. 1–52.

[50] Mark Schmidt, Nicolas Le Roux, and Francis Bach. "Minimizing Finite Sums with the Stochastic Average Gradient". In: *arXiv:1309.2388* (2013).

[51] Shai Shalev-Shwartz and Ambuj Tewari. "Stochastic methods for ℓ_1-regularized loss minimization". In: *Journal of Machine Learning Research* 12 (2011), pp. 1865–1892.

[52] Shai Shalev-Shwartz and Tong Zhang. "Accelerated Mini-Batch Stochastic Dual Coordinate Ascent". In: *Advances in Neural Information Processing Systems 26*. 2013, pp. 378–385.

[53] Shai Shalev-Shwartz and Tong Zhang. "Stochastic dual coordinate ascent methods for regularized loss". In: *Journal of Machine Learning Research* 14.1 (2013), pp. 567–599.

[54] S. U. Stich, C. L. Müller, and B. Gärtner. "Optimization of Convex Functions with Random Pursuit". In: *SIAM Journal on Optimization* 23.2 (2014), pp. 1284–1309.

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

[56] Martin Takáč, Avleen Bijral, Peter Richtárik, and Nathan Srebro. "Mini-batch primal and dual methods for SVMs". In: *Proceedings of the 30th International Conference on Machine Learning*. 2013.

[57] Qing Tao, Kang Kong, Dejun Chu, and Gaowei Wu. "Stochastic Coordinate Descent Methods for Regularized Smooth and Nonsmooth Losses". In: *Machine Learning and Knowledge Discovery in Databases* (2012), pp. 537–552.

[58] Rachael Tappenden, Peter Richtárik, and Jacek Gondzio. "Inexact block coordinate descent method: complexity and preconditioning". In: *arXiv:1304.5530* (2013).

[59] Paul Tseng. "On linear convergence of iterative methods for the variational inequality problem". In: *Journal of Computational and Applied Mathematics* 60 (1–2 1995), pp. 237–252.

[60] Stephen J. Wright. "Accelerated block-coordinate relaxation for regularized optimization". In: *SIAM Journal on Optimization* 22 (1 2012), pp. 159–186.

[61] Stephen J. Wright. "Coordinate descent methods". In: *Mathematical Programming* 151 (1 2015), pp. 3–34.

[62] Lin Xiao and Tong Zhang. "A Proximal Stochastic Gradient Method with Progressive Variance Reduction". In: *arXiv:1403.4699* (2014).

[63] Tianbao Yang. "Trading Computation for Communication: Distributed Stochastic Dual Coordinate Ascent". In: *Advances in Neural Information Processing Systems 26*. Curran Associates, Inc., 2013, pp. 629–637.

[64] Yuchen Zhang and Lin Xiao. "Stochastic Primal-Dual Coordinate Method for Regularized Empirical Risk Minimization". In: *Proceedings of the 32nd International Conference on Machine Learning*. 2015, pp. 353–361.

[65] Peilin Zhao and Tong Zhang. "Stochastic Optimization with Importance Sampling". In: *arXiv:1401.2753* (2014).

[66] Anastasios Zouzias and Nikolaos Freris. "Randomized extended Kaczmarz for solving least-squares". In: *arXiv:1205.5770* (2012), p. 19.

[67] Anastasios Zouzias and Nikolaos M. Freris. "Randomized gossip algorithms for solving Laplacian systems". In: *IEEE European Control Conference (ECC)*. 2015, pp. 1920–1925.

10 Appendix: Elementary Results Often Used in the Paper

We first state a lemma comparing the null spaces and range spaces of certain related matrices. While the result is of an elementary nature, we use it several times in this paper, which justifies its elevation to the status of a lemma. The proof is brief and hence we include it for completeness.

Lemma 10.1. *For any matrix W and symmetric positive definite matrix G,*

$$\mathbf{Null}\left(W\right) = \mathbf{Null}\left(W^\top G W\right) \tag{53}$$

and

$$\mathbf{Range}\left(W^\top\right) = \mathbf{Range}\left(W^\top G W\right). \tag{54}$$

Proof: In order to establish (53), it suffices to show the inclusion $\mathbf{Null}\left(W\right) \supseteq \mathbf{Null}\left(W^\top G W\right)$ 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 $G^{1/2} W s = 0$. Therefore, $s \in \mathbf{Null}\left(W\right)$. Finally, (54) follows from (53) by taking orthogonal complements. Indeed, $\mathbf{Range}\left(W^\top\right)$ is the orthogonal complement of $\mathbf{Null}\left(W\right)$ and $\mathbf{Range}\left(W^\top G W\right)$ is the orthogonal complement of $\mathbf{Null}\left(W^\top G W\right)$. □

The following technical lemma is a variant of a standard result of linear algebra (which is recovered in the $B = I$ case). While the results are folklore and easy to establish, in the proof of our main theorem we need certain details which are hard to find in textbooks on linear algebra, and hence hard to refer to. For the benefit of the reader, we include the detailed statement and proof.

Lemma 10.2 (Decomposition and Projection)**.** *Each $x \in \mathbb{R}^n$ can be decomposed in a unique way as $x = s(x) + t(x)$, where $s(x) \in \mathbf{Range}\left(B^{-1}A^\top\right)$ and $t(x) \in \mathbf{Null}\,(A)$. Moreover, the decomposition can be computed explicitly as*

$$s(x) = \arg\min_s \left\{ \|x - s\|_B \ : \ s \in \mathbf{Range}\left(B^{-1}A^\top\right) \right\} = B^{-1}Z_A x \tag{55}$$

and

$$t(x) = \arg\min_t \{\|x - t\|_B \ : \ t \in \mathbf{Null}\,(A)\} = (I - B^{-1}Z_A)x, \tag{56}$$

where

$$Z_A \overset{def}{=} A^\top (AB^{-1}A^\top)^\dagger A. \tag{57}$$

Hence, the matrix $B^{-1}Z_A$ is a projector in the B-norm onto $\mathbf{Range}\left(B^{-1}A^\top\right)$, and $I - B^{-1}Z_A$ is a projector in the B-norm onto $\mathbf{Null}\,(A)$. Moreover, for all $x \in \mathbb{R}^n$ we have $\|x\|_B^2 = \|s(x)\|_B^2 + \|t(x)\|_B^2$, with

$$\|t(x)\|_B^2 = \|(I - B^{-1}Z_A)x\|_B^2 = x^\top (B - Z_A)x \tag{58}$$

and

$$\|s(x)\|_B^2 = \|B^{-1}Z_A x\|_B^2 = x^\top Z_A x. \tag{59}$$

Finally,

$$\mathbf{Rank}\,(A) = \mathbf{Tr}\left(B^{-1}Z_A\right). \tag{60}$$

Proof: Fix arbitrary $x \in \mathbb{R}^n$. We first establish existence of the decomposition. By Lemma 10.1 applied to $W = A^\top$ and $G = B^{-1}$ we know that there exists u such that $Ax = AB^{-1}A^\top u$. Now let $s = B^{-1}A^\top u$ and $t = x - s$. Clearly, $s \in \mathbf{Range}\left(B^{-1}A^\top\right)$ and $t \in \mathbf{Null}\,(A)$. For uniqueness, consider two decompositions: $x = s_1 + t_1$ and $x = s_2 + t_2$. Let u_1, u_2 be vectors such that $s_i = B^{-1}A^\top u_i$, $i = 1, 2$. Then $AB^{-1}A^\top(u_1 - u_2) = 0$. Invoking Lemma 10.1 again, we see that $u_1 - u_2 \in \mathbf{Null}\left(A^\top\right)$, whence $s_1 = B^{-1}A^\top u_1 = B^{-1}A^\top u_2 = s_2$. Therefore, $t_1 = x - s_1 = x - s_2 = t_2$, establishing uniqueness.

Note that $s = B^{-1}A^\top y$, where y is any solution of the optimization problem

$$\min_y \tfrac{1}{2}\|x - B^{-1}A^\top y\|_B^2.$$

The first order necessary and sufficient optimality conditions are $Ax = AB^{-1}A^\top y$. In particular, we may choose y to be the least norm solution of this system, which gives $y = (AB^{-1}A^\top)^\dagger Ax$, from which (55) follows. The variational formulation (56) can be established in a similar way, again via first order optimality conditions (note that the closed form formula (56) also directly follows from (55) and the fact that $t = x - s$).

Next, since $x = s + t$ and $s^\top Bt = 0$,

$$\|t\|_B^2 = (t + s)^\top Bt = x^\top Bt \overset{(56)}{=} x^\top B(I - B^{-1}Z_A)x = x^\top (B - Z_A)x \tag{61}$$

and

$$\|s\|_B^2 = \|x\|_B^2 - \|t\|_B^2 \overset{(61)}{=} x^\top Z_A x.$$

It only remains to establish (60). Since $B^{-1}Z_A$ is a projector onto $\mathbf{Range}\left(B^{-1}A^\top\right)$ and since the trace of each projector is equal to the dimension of the space they project onto, we have $\mathbf{Tr}\left(B^{-1}Z_A\right) = \dim(\mathbf{Range}\left(B^{-1}A^\top\right)) = \dim(\mathbf{Range}\left(A^\top\right)) = \mathbf{Rank}\,(A)$. □