Stochastic Reformulations of Linear Systems

Algorithms and Convergence Theory

Stochastic Reformulations of Linear Systems: Algorithms and Convergence Theory*

Peter Richtárik† Martin Takáč‡

June 7, 2017

Abstract

We develop a family of reformulations of an arbitrary consistent linear system into a *stochastic problem*. The reformulations are governed by two user-defined parameters: a positive definite matrix defining a norm, and an arbitrary discrete or continuous distribution over random matrices. Our reformulation has several equivalent interpretations, allowing for researchers from various communities to leverage their domain specific insights. In particular, our reformulation can be equivalently seen as a stochastic optimization problem, stochastic linear system, stochastic fixed point problem and a probabilistic intersection problem. We prove sufficient, and necessary and sufficient conditions for the reformulation to be exact.

Further, we propose and analyze three stochastic algorithms for solving the reformulated problem—basic, parallel and accelerated methods—with global linear convergence rates. The rates can be interpreted as condition numbers of a matrix which depends on the system matrix and on the reformulation parameters. This gives rise to a new phenomenon which we call *stochastic preconditioning*, and which refers to the problem of finding parameters (matrix and distribution) leading to a sufficiently small condition number. Our basic method can be equivalently interpreted as stochastic gradient descent, stochastic Newton method, stochastic proximal point method, stochastic fixed point method, and stochastic projection method, with fixed stepsize (relaxation parameter), applied to the reformulations.

*All theoretical results in this paper were obtained by August 2016 and a first draft was circulated to a few selected colleagues in September 2016. The first author gave several talks on these results before this draft was made publicly available on arXiv: *Linear Algebra and Parallel Computing at the Heart of Scientific Computing*, Edinburgh, UK (Sept 21, 2016), *Seminar on Combinatorics, Games and Optimisation*, London School of Economics, London, UK, (Nov 16, 2016), *Workshop on Distributed Machine Learning*, Télécom ParisTech, Paris, France (Nov 25, 2016), Skoltech Seminar, Moscow, Russia (Dec 1, 2016), *BASP Frontiers Workshop*, Villars-sur-Ollon, Switzerland (Feb 1, 2017), and *SIAM Conference on Optimization*, Vancouver, Canada (May 22, 2017). In addition, the first author has included the results of this paper in the MSc/PhD course *Modern optimization methods for big data problems*, delivered in Spring 2017 at the University of Edinburgh, as an introduction into the role of randomized decomposition in linear algebra, optimization and machine learning. All main results of this paper were distributed to the students in the form of slides.

†KAUST and University of Edinburgh

‡Lehigh University

1 Introduction

Linear systems form the backbone of most numerical codes used in academia and industry. With the advent of the age of big data, practitioners are looking for ways to solve linear systems of unprecedented sizes. The present work is motivated by the need to design such algorithms. As an algorithmic tool enabling faster computation, *randomization* is well developed, understood and appreciated in several fields, typically traditionally of a "discrete" nature, most notably theoretical computer science [28]. However, probabilistic ideas are also increasingly and successfully penetrating "continuous" fields, such as numerical linear algebra [57, 10, 11, 60, 27, 50, 1], optimization [24, 33, 35, 46, 59, 40], control theory [4, 5, 62], machine learning [52, 47, 21, 14, 44], and signal processing [7, 22].

In this work we are concerned with the problem of solving a consistent linear system. In particular, consider the problem

$$\text{solve} \quad \mathbf{A}x = b, \tag{1}$$

where $0 \neq \mathbf{A} \in \mathbb{R}^{m\times n}$. We shall assume throughout the paper that the system is consistent, i.e., $\mathcal{L} \overset{\text{def}}{=} \{x \,:\, \mathbf{A}x = b\} \neq \emptyset$. Problem (1) is arguably one of the most important problems in linear algebra. As such, a tremendous amount of research has been done to design efficient iterative algorithms [51]. However, surprisingly little is know about randomized iterative algorithms for solving linear systems. In this work we aim to contribute to closing this gap.

1.1 Stochastic reformulations of linear systems

We propose a fundamental and flexible way of reformulating each consistent linear system into a *stochastic problem*. To the best of our knowledge, this is the first systematic study of such reformulations. Stochasticity is introduced in a controlled way, into an otherwise deterministic problem, as a decomposition tool which can be leveraged to design efficient, granular and scalable randomized algorithms.

Parameters defining the reformulation. Stochasticity enters our reformulations through a user-defined distribution $\mathcal{D}$ describing an ensemble of random matrices $\mathbf{S} \in \mathbb{R}^{m\times q}$. We make use of one more parameter: a user-defined $n\times n$ symmetric positive definite matrix $\mathbf{B}$. Our approach and underlying theory support virtually all thinkable distributions. The choice of the distribution should ideally depend on the problem itself, as it will affect the conditioning of the reformulation. However, for now we leave such considerations aside.

One stochastic reformulation in four disguises. Our reformulation of (1) as a stochastic problem has several seemingly different, yet equivalent interpretations, and hence we describe them here side by side.

1. *Stochastic optimization problem.* Consider the problem

$$\text{minimize} \quad f(x) \overset{\text{def}}{=} \mathrm{E}_{\mathbf{S}\sim\mathcal{D}}\left[f_{\mathbf{S}}(x)\right], \tag{2}$$

where $f_{\mathbf{S}}(x) = \frac{1}{2}(\mathbf{A}x-b)^\top \mathbf{H}(\mathbf{A}x-b)$, $\mathbf{H} = \mathbf{S}(\mathbf{S}^\top\mathbf{A}\mathbf{B}^{-1}\mathbf{A}^\top\mathbf{S})^\dagger\mathbf{S}^\top$, and $\dagger$ denotes the Moore-Penrose pseudoinverse. When solving the problem, we do not have (or do not wish to exercise, as it may be prohibitively expensive) explicit access to f, its gradient or Hessian. Rather, we can repeatedly sample $\mathbf{S}\sim\mathcal{D}$ and receive unbiased samples of these quantities at points of interest. That is, we may obtain local information about the *stochastic function* $f_{\mathbf{S}}$, such as the *stochastic gradient* $\nabla f_{\mathbf{S}}(x)$, and use this to drive an iterative process for solving (2).

2. *Stochastic linear system.* Consider now a preconditioned version of the linear system (1) given by

$$\text{solve} \quad \mathbf{B}^{-1}\mathbf{A}^\top \mathrm{E}_{\mathbf{S}\sim\mathcal{D}}\left[\mathbf{H}\right]\mathbf{A}x = \mathbf{B}^{-1}\mathbf{A}^\top \mathrm{E}_{\mathbf{S}\sim\mathcal{D}}\left[\mathbf{H}\right]b, \tag{3}$$

where $\mathbf{P} = \mathbf{B}^{-1}\mathbf{A}^\top \mathrm{E}_{\mathbf{S}\sim\mathcal{D}}\left[\mathbf{H}\right]$ is the preconditioner. The preconditioner is not assumed to be known explicitly. Instead, when solving the problem, we are able to repeatedly sample $\mathbf{S}\sim\mathcal{D}$, obtaining an unbiased estimate of the preconditioner (not necessarily explicitly), $\mathbf{B}^{-1}\mathbf{A}^\top\mathbf{H}$, for which we coin the name *stochastic preconditioner*. This gives us access to an unbiased sample of the preconditioned system (3): $\mathbf{B}^{-1}\mathbf{A}^\top\mathbf{H}\mathbf{A}x = \mathbf{B}^{-1}\mathbf{A}^\top\mathbf{H}b$. As we shall see—in an analogy with stochastic optimization—the information contained in such systems can be utilized by an iterative algorithm to solve (3).

3. *Stochastic fixed point problem.* Let $\Pi^{\mathbf{B}}_{\mathcal{L}_\mathbf{S}}(x)$ denote the projection of x onto $\mathcal{L}_\mathbf{S} \overset{\text{def}}{=} \{x \;:\; \mathbf{S}^\top\mathbf{A}x = \mathbf{S}^\top b\}$, in the norm $\|x\|_\mathbf{B} \overset{\text{def}}{=} \sqrt{x^\top\mathbf{B}x}$. Consider the *stochastic fixed point problem*

$$\text{solve} \quad x = \mathrm{E}_{\mathbf{S}\sim\mathcal{D}}\left[\Pi^{\mathbf{B}}_{\mathcal{L}_\mathbf{S}}(x)\right]. \tag{4}$$

That is, we seek to find a *fixed point* of the mapping $x \to \mathrm{E}_{\mathbf{S}\sim\mathcal{D}}\left[\Pi^{\mathbf{B}}_{\mathcal{L}_\mathbf{S}}(x)\right]$. When solving the problem, we do not have an explicit access to the average projection map. Instead, we are able to repeatedly sample $\mathbf{S}\sim\mathcal{D}$, and use the stochastic projection map $x \to \Pi^{\mathbf{B}}_{\mathcal{L}_\mathbf{S}}(x)$.

4. *Probabilistic intersection problem.* Note that $\mathcal{L} \subseteq \mathcal{L}_\mathbf{S}$ for all $\mathbf{S}$. We would wish to design $\mathcal{D}$ in such a way that a suitably chosen notion of an intersection of the sets $\mathcal{L}_\mathbf{S}$ is equal to $\mathcal{L}$. The correct notion is what we call *probabilistic intersection*, denoted $\cap_{\mathbf{S}\sim\mathcal{D}}\mathcal{L}_\mathbf{S}$, and defined as the set of points x which belong to $\mathcal{L}_\mathbf{S}$ with probability one. This leads to the problem:

$$\text{find} \quad x \in \cap_{\mathbf{S}\sim\mathcal{D}}\mathcal{L}_\mathbf{S} \overset{\text{def}}{=} \{x \;:\; \mathrm{Prob}(x\in\mathcal{L}_\mathbf{S}) = 1\}. \tag{5}$$

As before, we typically do not have an explicit access to the probabilistic intersection when designing an algorithm. Instead, we can repeatedly sample $\mathbf{S}\sim\mathcal{D}$, and utilize the knowledge of $\mathcal{L}_\mathbf{S}$ to drive the iterative process. If $\mathcal{D}$ is a discrete distribution, probabilistic intersection reduces to standard intersection.

All of the above formulations have a common feature: they all involve an expectation over $\mathbf{S}\sim\mathcal{D}$, and we either do not assume this expectation is known explicitly, or even if it is, we prefer, due to efficiency or other considerations, to sample from unbiased estimates of the objects (e.g., stochastic gradient $\nabla f_\mathbf{S}$, stochastic preconditioner $\mathbf{B}^{-1}\mathbf{A}^\top\mathbf{H}$, stochastic projection map $x \to \Pi^{\mathbf{B}}_{\mathcal{L}_\mathbf{S}}(x)$, random set $\mathcal{L}_\mathbf{S}$) appearing in the formulation.

Equivalence and exactness. We show that all these stochastic reformulations are equivalent (see Theorem 3.4). In particular, the following sets are identical: the set of minimizers of the stochastic optimization problem (2), the solution set of the stochastic linear system (3), the set of fixed points of the stochastic fixed point problem (4), and the probabilistic intersection (5). Further, we give necessary and sufficient conditions for this set to be equal to $\mathcal{L}$. If this is the case, we say the the reformulation is *exact* (see Section 3.4). Distributions $\mathcal{D}$ satisfying these conditions always exist, independently of any assumptions on the system beyond consistency. The simplest, but also the least useful choice of a distribution is to pick $\mathbf{S} = \mathbf{I}$ (the $m\times m$ identity matrix), with probability one. In this case, all of our reformulations become trivial.

1.2 Stochastic algorithms

Besides proposing a family of stochastic reformulations of the linear system (1), we also propose three stochastic algorithms for solving them: Algorithm 1 (basic method), Algorithm 2 (parallel/minibatch method), and Algorithm 3 (accelerated method). Each method can be interpreted naturally from the viewpoint of each of the reformulations.

Basic method. Below we list some of the interpretations of Algorithm 1 (basic method), which performs updates of the form

$$x_{k+1} = \phi_\omega(x_k, \mathbf{S}_k) \overset{\text{def}}{=} x_k - \omega \mathbf{B}^{-1}\mathbf{A}^\top \mathbf{S}_k (\mathbf{S}_k^\top \mathbf{A}\mathbf{B}^{-1}\mathbf{A}^\top \mathbf{S}_k)^\dagger \mathbf{S}_k^\top (\mathbf{A}x_k - b), \tag{6}$$

where $\mathbf{S}_k \sim \mathcal{D}$ is sampled afresh in each iteration. The method is formally presented and analyzed in Section 4.

1. **Stochastic gradient descent.** Algorithm 1 can be seen as *stochastic gradient descent* [48], with fixed stepsize, applied to (2). At iteration k of the method, we sample $\mathbf{S}_k \sim \mathcal{D}$, and compute $\nabla f_{\mathbf{S}_k}(x_k)$, which is an unbiased stochastic approximation of $\nabla f(x_k)$. We then perform the step

$$x_{k+1} = x_k - \omega \nabla f_{\mathbf{S}_k}(x_k), \tag{7}$$

   where $\omega > 0$ is a stepsize.

   Let us note that in order to achieve linear convergence it is not necessary to use any explicit variance reduction strategy [52, 21, 23, 9], nor do we need to use decreasing stepsizes. This is because the stochastic gradients vanish at the optimum, which is a consequence of the consistency assumption. Surprisingly, we get linear convergence in spite of the fact that we deal with a non-finite-sum problem (2), and without the need to assume boundedness of the stochastic gradients, and without f being strongly convex. To the best of our knowledge, this is the first linearly convergent accelerated method for stochastic optimization without requiring strong convexity. This beats the minimax bounds given by Srebro [58]. This is because (2) is not a black-box stochastic optimization objective; indeed, we have constructed it in a particular way from the underlying linear system (1).

2. **Stochastic Newton method.** However, Algorithm 1 can also be seen as a *stochastic Newton method*. At iteration k we sample $\mathbf{S}_k \sim \mathcal{D}$, and instead of applying the inverted Hessian of $f_{\mathbf{S}_k}$ to the stochastic gradient (this is not possible as the Hessian is not necessarily invertible), we apply the $\mathbf{B}$-pseudoinverse. That is, we perform the step

$$x_{k+1} = x_k - \omega (\nabla^2 f_{\mathbf{S}_k}(x_k))^{\dagger_{\mathbf{B}}} \nabla f_{\mathbf{S}_k}(x_k), \tag{8}$$

   where $\omega > 0$ is a stepsize, and the $\mathbf{B}$-pseudoinverse of a matrix $\mathbf{M}$ is defined as $\mathbf{M}^{\dagger_{\mathbf{B}}} \overset{\text{def}}{=} \mathbf{B}^{-1}\mathbf{M}^\top(\mathbf{M}\mathbf{B}^{-1}\mathbf{M}^\top)^\dagger$. One may wonder why methods (7) and (8) are equivalent; after all, the (stochastic) gradient descent and (stochastic) Newton methods are not equivalent in general. However, in our setting it turns out that the stochastic gradient $\nabla f_{\mathbf{S}_k}(x)$ is always an eigenvector of $(\nabla^2 f_{\mathbf{S}_k}(x))^{\dagger_{\mathbf{B}}}$, with eigenvalue 1 (see Lemma 3.1).

   Stochastic Newton-type methods were recently developed and analyzed in the optimization and machine learning literature [39, 38, 41, 29]. However, they are design to solve different problems, and operate in a different manner.

3. **Stochastic proximal point method.** If we restrict our attention to stepsizes satisfying $0 < \omega \leq 1$, then Algorithm 1 can be equivalently (see Theorem A.3 in the Appendix) written down as

$$x_{k+1} = \arg\min_{x \in \mathbb{R}^n} \left\{ f_{\mathbf{S}_k}(x) + \frac{1-\omega}{2\omega} \|x - x_k\|^2_{\mathbf{B}} \right\}. \tag{9}$$

That is, (9) is a *stochastic* variant of the *proximal point method* for solving (2), with a fixed regularization parameter [49]. The proximal point method is obtained from (9) by replacing $f_{\mathbf{S}_k}$ with f. If we define the *prox operator* of a function $\psi : \mathbb{R}^n \to \mathbb{R}$ with respect to the $\mathbf{B}$-norm as $\operatorname{prox}^{\mathbf{B}}_{\psi}(y) \overset{\text{def}}{=} \arg\min_{x \in \mathbb{R}^n} \left\{ \psi(x) + \frac{1}{2}\|x - y\|^2_{\mathbf{B}} \right\}$, then iteration (9) can be written compactly as $x_{k+1} = \operatorname{prox}^{\mathbf{B}}_{\frac{\omega}{1-\omega} f_{\mathbf{S}_k}}(x_k)$.

4. **Stochastic fixed point method.** From the perspective of the stochastic fixed point problem (4), Algorithm 1 can be interpreted as a *stochastic fixed point method*, with relaxation. We first reformulate the problem into an equivalent form using relaxation, which is done to improve the contraction properties of the map. We pick a parameter $\omega > 0$, and instead consider the equivalent fixed point problem $x = \mathrm{E}_{\mathbf{S} \sim \mathcal{D}} \left[\omega \Pi^{\mathbf{B}}_{\mathcal{L}_{\mathbf{S}}}(x) + (1-\omega)x \right]$. Now, at iteration k, we sample $\mathbf{S}_k \sim \mathcal{D}$, which enables us to obtain an unbiased estimate of the new fixed point mapping, and then simply perform one step of a fixed point method on this mapping:

$$x_{k+1} = \omega \Pi^{\mathbf{B}}_{\mathcal{L}_{\mathbf{S}_k}}(x_k) + (1-\omega)x_k. \tag{10}$$

5. **Stochastic projection method.** Algorithm 1 can also be seen as a *stochastic projection method* applied to the probabilistic intersection problem (5). By sampling $\mathbf{S}_k \sim \mathcal{D}$, we are one of the sets defining the intersection, namely $\mathcal{L}_{\mathbf{S}_k}$. We then project the last iterate onto this set, in the $\mathbf{B}$-norm, followed by a relaxation step with relaxation parameter $\omega > 0$. That is, we perform the update

$$x_{k+1} = x_k + \omega(\Pi^{\mathbf{B}}_{\mathcal{L}_{\mathbf{S}_k}}(x_k) - x_k). \tag{11}$$

This is a randomized variant of an alternating projection method. Note that the representation of $\mathcal{L}$ as a probabilistic intersection of sets is not given to us. Rather, we construct it with the hope to obtain faster convergence.

An optimization algorithm utilizing stochastic projection steps was developed in [30]. For a comprehensive survey of projection methods for convex feasibility problems, see [2].

Parallel method. A natural parallelization strategy is to perform one step of the basic method independently τ times, starting from the same point x_k, and average the results:

$$x_{k+1} = \frac{1}{\tau} \sum_{i=1}^{\tau} \phi_{\omega}(x_k, \mathbf{S}^i_k), \tag{12}$$

where $\mathbf{S}^1_k, \ldots, \mathbf{S}^\tau_k$ are independent samples from $\mathcal{D}$ (recall that ϕ_ω is defined in (6)). This method is formalized as Algorithm 2, and studied in Section 5.1. Betrayed by our choice of the name, this method is useful in scenarios where τ parallel workers are available, allowing for the τ basic steps to be computed in parallel, followed by an averaging operation.

From the stochastic optimization viewpoint, this is a *minibatch* method. Considering the SGD interpretation (7), we can equivalently write (12) in the form $x_{k+1} = x_k - \frac{1}{\tau}\sum_{i=1}^{\tau} \nabla f_{\mathbf{S}_k^i}(x_k)$. This is *minibatch SGD*. Iteration complexity of minibatch SGD was first understood in the context of training support vector machines with the hinge loss [61]. Complexity under a lock-free paradigm, in a different setting from ours, was first studied in [36]. Notice that in the limit $\tau \to \infty$, we obtain gradient descent. It is therefore interesting to study the complexity of the parallel method as a function τ. Of course, this method can also be interpreted as a minibatch stochastic Newton method, minibatch proximal point method and so on.

From the probabilistic intersection point of view, method (12) can be interpreted as a stochastic variant of the parallel projection method. In particular, we obtain the iterative process

$$x_{k+1} = x_k + \omega \left[\left(\frac{1}{\tau} \sum_{i=1}^{\tau} \Pi^{\mathbf{B}}_{\mathcal{L}_{\mathbf{S}_k^i}}(x_k) \right) - x_k \right].$$

That is, we move from the current iterate, x_k, towards the average of the τ projection points, with undershooting (if $\omega < 1$), precisely landing on (if $\omega = 1$), or overshooting (if $\omega > 1$) the average. Projection methods have a long history and are well studied [12, 3]. However, much less is known about stochastic projection methods.

Accelerated method. In order to obtain acceleration without parallelization—that is, acceleration in the sense of Nesterov [34]—we suggest to perform an update step in which x_{k+1} depends on both x_k and x_{k-1}. In particular, we take two *dependent* steps of Algorithm 1, one from x_k and one from x_{k-1}, and then take an affine combination of the results. That is, the process is started with $x_0, x_1 \in \mathbb{R}^n$, and for $k \geq 1$ involves an iteration of the form

$$x_{k+1} = \gamma \phi_\omega(x_k, \mathbf{S}_k) + (1-\gamma)\phi_\omega(x_{k-1}, \mathbf{S}_{k-1}), \tag{13}$$

where the matrices $\{\mathbf{S}_k\}$ are independent samples from $\mathcal{D}$, and $\gamma \in \mathbb{R}$ is an *acceleration parameter*. Note that by choosing $\gamma = 1$ (no acceleration), we recover Algorithm 1. This method is formalized as Algorithm 3 and analyzed in Section 5.2. Our theory suggests that γ should be always between 1 and 2. In particular, for well conditioned problems (small ζ), one should choose $\gamma \approx 1$, and for ill conditioned problems (large ζ), one should choose $\gamma \approx 2$.

1.3 Complexity

The complexity of our methods is completely described by the spectrum of the (symmetric positive semidefinite) matrix

$$\mathbf{W} \overset{\text{def}}{=} \mathbf{B}^{-1/2}\mathbf{A}^\top \mathrm{E}_{\mathbf{S}\sim\mathcal{D}}\left[\mathbf{H}\right]\mathbf{A}\mathbf{B}^{-1/2}.$$

Let $\mathbf{W} = \mathbf{U}\Lambda\mathbf{U}^\top$ be the eigenvalue decomposition of $\mathbf{W}$, where $\mathbf{U} = [u_1, \ldots, u_n]$ are the eigenvectors, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0$ are the eigenvalues, and $\Lambda = \mathbf{Diag}(\lambda_1, \ldots, \lambda_n)$. It can be shown that the largest eigenvalue, $\lambda_{\max} \overset{\text{def}}{=} \lambda_1$ is bounded above by 1 (see Lemma 4.1). Let $\lambda_{\min}^+$ be the smallest nonzero eigenvalue.

With all of the above reformulations we associate the same *condition number*

$$\zeta = \zeta(\mathbf{A}, \mathbf{B}, \mathcal{D}) \overset{\text{def}}{=} \|\mathbf{W}\|\|\mathbf{W}^\dagger\| = \frac{\lambda_{\max}}{\lambda_{\min}^+}, \tag{14}$$

Alg.	ω	τ	γ	Quantity	Rate	Complexity	Theorem
1	1	-	-	$\|\mathrm{E}\left[x_k - x_*\right]\|^2_{\mathbf{B}}$	$(1-\lambda^+_{\min})^{2k}$	$1/\lambda^+_{\min}$	4.3, 4.4, 4.6
1	$1/\lambda_{\max}$	-	-	$\|\mathrm{E}\left[x_k - x_*\right]\|^2_{\mathbf{B}}$	$(1-1/\zeta)^{2k}$	ζ	4.3, 4.4, 4.6
1	$\frac{2}{\lambda^+_{\min}+\lambda_{\max}}$	-	-	$\|\mathrm{E}\left[x_k - x_*\right]\|^2_{\mathbf{B}}$	$(1-2/(\zeta+1))^{2k}$	ζ	4.3, 4.4, 4.6
1	1	-	-	$\mathrm{E}\left[\|x_k - x_*\|^2_{\mathbf{B}}\right]$	$(1-\lambda^+_{\min})^{k}$	$1/\lambda^+_{\min}$	4.8
1	1	-	-	$\mathrm{E}\left[f(x_k)\right]$	$(1-\lambda^+_{\min})^{k}$	$1/\lambda^+_{\min}$	4.10
2	1	τ	-	$\mathrm{E}\left[\|x_k - x_*\|^2_{\mathbf{B}}\right]$	$\left(1-\lambda^+_{\min}\left(2-\xi(\tau)\right)\right)^{k}$		5.1
2	$1/\xi(\tau)$	τ	-	$\mathrm{E}\left[\|x_k - x_*\|^2_{\mathbf{B}}\right]$	$\left(1-\frac{\lambda^+_{\min}}{\xi(\tau)}\right)^{k}$	$\xi(\tau)/\lambda^+_{\min}$	5.1
2	$1/\lambda_{\max}$	∞	-	$\mathrm{E}\left[\|x_k - x_*\|^2_{\mathbf{B}}\right]$	$(1-1/\zeta)^{k}$	ζ	5.1
3	1	-	$\frac{2}{1+\sqrt{0.99\lambda^+_{\min}}}$	$\|\mathrm{E}\left[x_k - x_*\right]\|^2_{\mathbf{B}}$	$\left(1-\sqrt{0.99\lambda^+_{\min}}\right)^{2k}$	$\sqrt{1/\lambda^+_{\min}}$	5.3
3	$1/\lambda_{\max}$	-	$\frac{2}{1+\sqrt{0.99/\zeta}}$	$\|\mathrm{E}\left[x_k - x_*\right]\|^2_{\mathbf{B}}$	$\left(1-\sqrt{0.99/\zeta}\right)^{2k}$	$\sqrt{\zeta}$	5.3

Table 1: Summary of the main complexity results. In all cases, $x_* = \Pi^{\mathbf{B}}_{\mathcal{L}}(x_0)$ (the projection of the starting point onto the solution space of the linear system). "Complexity" refers to the number of iterations needed to drive "Quantity" below some error tolerance $\epsilon > 0$ (we suppress a $\log(1/\epsilon)$ factor in all expressions in the "Complexity" column). In the table we use the following expressions: $\xi(\tau) = \frac{1}{\tau} + (1-\frac{1}{\tau})\lambda_{\max}$ and $\zeta = \lambda_{\max}/\lambda^+_{\min}$.

where $\|\cdot\|$ is the spectral norm, $\lambda_{\max}$ is the largest eigenvalue of $\mathbf{W}$ and $\lambda^+_{\min}$ is the smallest nonzero eigenvalue of $\mathbf{W}$. Note that, for example, ζ is the condition number of the Hessian of f, and also the condition number of the stochastic linear system (3). Natural interpretations from the viewpoint of the stochastic fixed point and probabilistic intersection problems are also possible. As one varies the parameters defining the reformulation ($\mathcal{D}$ and $\mathbf{B}$), the condition number changes. For instance, choosing $\mathbf{S} = \mathbf{I}$ with probability one gives $\zeta = 1$.

Exact formula for the evolution of expected iterates. We first show (Theorem 4.3) that after the canonical linear transformation $x \mapsto \mathbf{U}^\top \mathbf{B}^{1/2} x$, the expected iterates of the basic method satisfy the identity

$$\mathrm{E}\left[\mathbf{U}^\top \mathbf{B}^{1/2}(x_k - x_*)\right] = (\mathbf{I} - \omega\Lambda)^k \mathbf{U}^\top \mathbf{B}^{1/2}(x_0 - x_*), \tag{15}$$

where x_* is an arbitrary solution of the linear system (i.e., $x_* \in \mathcal{L}$). This identity seems to suggest that zero eigenvalues cause an issue, preventing convergence of the corresponding elements of the error to zero. Indeed, if $\lambda_i = 0$, then (15) implies that $u_i^\top \mathbf{B}^{1/2}(x_k - x_*) = u_i^\top \mathbf{B}^{1/2}(x_0 - x_*)$, which does not change with k. However, it turns out that under exactness we have $u_i^\top \mathbf{B}^{1/2}(x_0 - x_*) = 0$ whenever $\lambda_i = 0$ if we let x_* to be the projection, in the $\mathbf{B}$-norm, of x_0 onto $\mathcal{L}$ (Theorem 4.3). This is then used to argue (Corollary 4.4) that $\|\mathrm{E}\left[x_k - x_*\right]\|_{\mathbf{B}}$ converges to zero if and only if $0 < \omega < 2/\lambda_{\max}$. The choice of stepsize issue is discussed in detail in Section 4.4.

The main complexity results obtained in this paper are summarized in Table 1. The full statements including the dependence of the rate on these parameters, as well as other alternative results (such as lower bounds, ergodic convergence) can be found in the theorems referenced in the table.

L2 convergence. The rate of decay of the quantity $\|\mathrm{E}\left[x_k - x_*\right]\|^2_{\mathbf{B}}$ for three different stepsize choices is summarized in the first three rows of Table 1. In particular, the default stepsize $\omega = 1$ leads to the complexity $1/\lambda^+_{\min}$, the long stepsize $\omega = 1/\lambda_{\max}$ gives the improved complexity $\lambda_{\max}/\lambda^+_{\min}$, and the

optimal stepsize $\omega = 2/(\lambda_{\max} + \lambda'_{\min})$ gives the best complexity $0.5 + 0.5\lambda_{\max}/\lambda'_{\min}$. However, if we are interested in the convergence of the larger quantity $\mathrm{E}\left[\|x_k - x_*\|^2_{\mathbf{B}}\right]$ (L2 convergence), it turns out that $\omega = 1$ is the optimal choice, leading to the complexity $1/\lambda'_{\min}$.

Parallel and accelerated methods. The parallel method improves upon the basic method in that it is capable of faster L2 convergence. We give a complexity formula as a function of τ, recovering the complexity the $1/\lambda'_{\min}$ rate of the basic method in the $\tau = 1$ case, and achieving the improved asymptotic complexity $\lambda_{\max}/\lambda'_{\min}$ as $\tau \to \infty$ (recall that $\lambda_{\max} \leq 1$, whence the improvement). Because of this, $\lambda_{\max}$ is the quantity driving *parallelizability*. If $\lambda_{\max}$ is close to 1, then there is little or no reason to parallelize. If $\lambda_{\max}$ is very small, parallelization helps. The smaller $\lambda_{\max}$ is, the more gain is achieved by utilizing more processors.

With the correct stepsize (ω) and acceleration (γ) parameters, the accelerated method improves the complexity $\lambda_{\max}/\lambda'_{\min}$ achieved by the basic method to $\sqrt{\lambda_{\max}/\lambda'_{\min}}$. However, this is established for the quantity $\mathrm{E}\left[\|x_k - x_*\|^2_{\mathbf{B}}\right]$. We conjecture that the L2 convergence rate of the accelerated method (for a suitable choice of the parameters ω and γ) is $\sqrt{1/\lambda'_{\min}}$.

1.4 Stochastic preconditioning

We coin the phrase *stochastic preconditioning* to refer to the general problem of *designing* matrix $\mathbf{B}$ and distribution $\mathcal{D}$ such that the appropriate condition number of $\mathbf{W}$ is well behaved. For instance, one might be interested in minimizing (or reducing) the condition number $1/\lambda'_{\min}$ if the basic method with unit stepsize is used, and the quantity we wish to converge to zero is either $\mathrm{E}\left[\|x_k - x_*\|^2_{\mathbf{B}}\right]$, $\|\mathrm{E}\left[x_k - x_*\right]\|^2_{\mathbf{B}}$, or $\mathrm{E}\left[f(x_k)\right]$ (see Lines 1, 4 and 5 of Table 1). On the other hand, if we can estimate $\lambda_{\max}$, then one may use the basic method with the larger stepsize $1/\lambda_{\max}$, in which case we may wish to minimize (or reduce) the condition number $\lambda_{\max}/\lambda'_{\min}$ (see Line 2 of Table 1).

One possible approach to stochastic preconditioning is to choose some $\mathbf{B}$ and then focus on a reasonably simple parametric family of distributions $\mathcal{D}$, trying to find the parameters which minimize (or reduce) the condition number of interest. The distributions in this family should be "comparable" in the sense that the cost of one iteration of the method of interest should be comparable for all distributions; as otherwise comparing bounds on the number of iterations does not make sense.

To illustrate this through a simple example, consider the family of discrete uniform distributions over m vectors in $\mathbb{R}^m$ (that is, $\mathbf{S}_1, \ldots, \mathbf{S}_m \in \mathbb{R}^{m \times 1}$), where the vectors themselves are the parameters defining the family. The cost of one iteration of the basic method will be proportional to the cost of performing a matrix-vector product of the form $\mathbf{S}^\top \mathbf{A}$, which is comparable across all distributions in the family (assuming the vectors are dense, say). To illustrate this approach, consider this family, and further assume that $\mathbf{A}$ is $n \times n$ symmetric and positive definite. Choose $\mathbf{B} = \mathbf{A}$. It can be shown that $1/\lambda'_{\min}$ is maximized precisely when $\{\mathbf{S}_i\}$ correspond to the eigenvectors of $\mathbf{A}$. In this case, $1/\lambda'_{\min} = n$, and hence our stochastic preconditioning strategy results in a condition number which is *independent* of the condition number of $\mathbf{A}$. If we now apply the basic method to the stochastic optimization reformulation, we can interpret it as a *spectral* variant of stochastic gradient descent (spectral SGD). Ignoring logarithmic terms, spectral SGD only needs to perform n matrix vector multiplications to solve the problem. While this is not a practical preconditioning strategy—computing the eigenvectors is hard, and if we actually had access to them, we could construct the solution directly, without the need to resort to an iterative scheme—it sheds light on the opportunities and challenges associated with stochastic preconditioning.

All standard sketching matrices $\mathbf{S}$ can be employed within our framework, including the count sketch [6] and the count-min sketch [8]. In the context to this paper (since we sketch with the transpose of $\mathbf{S}$), $\mathbf{S}$ is a count-sketch matrix (resp. count-min sketch) if it is assembled from random columns of $[\mathbf{I}, -\mathbf{I}]$ (resp $\mathbf{I}$), chosen uniformly with replacement, where $\mathbf{I}$ is the $m \times m$ identity matrix.

The notion of *importance sampling* developed in the last 5 years in the randomized optimization and machine learning literature [45, 64, 42, 40] can be seen a type of stochastic preconditioning, somewhat reverse to what we have outlined above. In these methods, the atoms forming the distribution $\mathcal{D}$ are fixed, and one is seeking to associate them with appropriate probabilities. Thus, the probability simplex is the parameter space defining the class of distributions one is considering.

Stochastic preconditioning is fundamentally different from the idea of *randomized preconditioning* [50, 1], which is based on a two-stage procedure. In the first step, the input matrix is randomly projected and an good preconditioning matrix is extracted. In the second step, an iterative least squares solver is applied to solve the preconditioned system.

Much like standard preconditioning, different stochastic preconditioning strategies will need to be developed for different classes of problems, with structure of $\mathbf{A}$ informing the choice of $\mathbf{B}$ and $\mathcal{D}$. Due to its inherent difficulty, stochastic preconditioning is beyond the scope of this paper.

1.5 Notation

For convenience, a table of the most frequently used notation is included in Appendix D. All matrices are written in bold capital letters. By $\mathrm{Range}\,(\mathbf{M})$ and $\mathrm{Null}\,(\mathbf{M})$ we mean the range space and null space of matrix $\mathbf{M}$, respectively. Given a symmetric positive definite matrix $\mathbf{B} \in \mathbb{R}^{n\times n}$, we equip $\mathbb{R}^n$ with the Euclidean inner product defined by $\langle x, h\rangle_{\mathbf{B}} \overset{\text{def}}{=} x^\top \mathbf{B} h$. We also define the induced norm: $\|x\|_{\mathbf{B}} \overset{\text{def}}{=} \sqrt{\langle x, x\rangle_{\mathbf{B}}}$. The short-hand notation $\|\cdot\|$ means $\|\cdot\|_{\mathbf{I}}$, where $\mathbf{I}$ is the identity matrix. We shall often write $\|x\|_{\mathbf{M}}$ for matrix $\mathbf{M} \in \mathbb{R}^{n\times n}$ being merely positive definite; this constitutes a pseudonorm.

2 Further Connections to Existing Work

In this section we outline several connections of our work with existing developments. We do not aim to be comprehensive.

2.1 Randomized Kaczmarz method, with relaxation and acceleration

Let $\mathbf{B} = \mathbf{I}$, and choose $\mathcal{D}$ as follows: $\mathbf{S} = e_i$ with probability $p_i = \|\mathbf{A}_{i:}\|_2^2/\|\mathbf{A}\|_F^2$. Since

$$\mathbf{W} = \mathbf{B}^{-1/2}\mathrm{E}\left[\mathbf{Z}\right]\mathbf{B}^{-1/2} = \mathrm{E}\left[\mathbf{Z}\right] = \sum_{i=1}^{m} p_i \frac{\mathbf{A}_{i:}^\top \mathbf{A}_{i:}}{\|\mathbf{A}_{i:}\|_2^2} = \frac{1}{\|\mathbf{A}\|_F^2}\sum_{i=1}^{m} \mathbf{A}_{i:}^\top \mathbf{A}_{i:} = \frac{\mathbf{A}^\top\mathbf{A}}{\|\mathbf{A}\|_F^2}.$$

the condition number is

$$\zeta = \|\mathbf{W}\|\|\mathbf{W}^\dagger\| = \|\mathrm{E}\left[\mathbf{Z}\right]\|\|\mathrm{E}\left[\mathbf{Z}\right]^\dagger\| = \frac{\lambda_{\max}(\mathbf{A}^\top\mathbf{A})}{\lambda_{\min}^{+}(\mathbf{A}^\top\mathbf{A})}. \tag{16}$$

Basic method. In this setup, Algorithm 1 simplifies to

$$x_{k+1} = x_k - \frac{\omega(\mathbf{A}_{i:}x_k - b_i)}{\|\mathbf{A}_{i:}\|_2^2}\mathbf{A}_{i:}^\top.$$

For $\omega = 1$, this reduces to the celebrated randomized Kaczmarz method (RK) of Strohmer and Vershyinin [60]. For $\omega > 1$, this is *RK with overrelaxation* – a new method not considered before. Based on Theorem 4.6, for $\omega \in [1/\lambda_{\max}, \omega_*]$ the iteration complexity of Algorithm 1 is

$$\tilde{\mathcal{O}}(\zeta) \overset{(16)}{=} \frac{\lambda_{\max}(\mathbf{A}^\top \mathbf{A})}{\lambda^+_{\min}(\mathbf{A}^\top \mathbf{A})}.$$

This is an improvement on standard RK method (with $\omega = 1$), whose complexity depends on $\mathrm{Trace}\left(\mathbf{A}^\top \mathbf{A}\right)$ instead of $\lambda_{\max}$. Thus, the improvement can be as large as by a factor n.

Accelerated method. In the same setup, Algorithm 3 simplifies to

$$x_{k+1} = \gamma\left(x_k - \frac{\omega(\mathbf{A}_{i_k:}x_k - b_{i_k})}{\|\mathbf{A}_{i_k:}\|_2^2}\mathbf{A}_{i_k:}^\top\right) + (1-\gamma)\left(x_{k-1} - \frac{\omega(\mathbf{A}_{i_{k-1}:}x_{k-1} - b_{i_{k-1}})}{\|\mathbf{A}_{i_{k-1}:}\|_2^2}\mathbf{A}_{i_{k-1}:}^\top\right)$$

This is accelerated RK method with overrelaxation – a new method not considered before. Based on Theorem 5.3, for the parameter choice $\omega = 1/\lambda_{\max}$ and $\gamma = 2/(1+\zeta^{-2})$, the iteration complexity of this method is

$$\tilde{\mathcal{O}}(\sqrt{\zeta}) \overset{(16)}{=} \sqrt{\frac{\lambda_{\max}(\mathbf{A}^\top \mathbf{A})}{\lambda^+_{\min}(\mathbf{A}^\top \mathbf{A})}}.$$

If we instead choose $\omega = 1$ and $\gamma = 2/(1+\zeta^{-2})$, the iteration complexity gets slightly worse: $1/\sqrt{\lambda^+_{\min}(\mathbf{A}^\top \mathbf{A})}$. To the best of our knowledge, this is the best known complexity for a variant of RK. Let us remark that an asynchronous accelerated RK method was developed in [25].

The randomized Kaczmarz method, its variants have received considerable attention recently [31, 65, 32, 43], and several connections to existing methods were made. Kaczmarz-type methods in a Hilbert setting were developed in [37].

2.2 Basic method with unit stepsize

The method $x_{k+1} \leftarrow \Pi^{\mathbf{B}}_{\mathcal{L}_{\mathbf{S}_k}}(x_k)$ was first proposed and analyzed (under a full rank assumption on $\mathbf{A}$) in [17]. Note that in view of (10), this is the basic method with unit stepsize. However, it was not interpreted as a method for solving any of the reformulations presented here, and as a result, all the interpretations we are giving here also remained undiscovered. Instead, it was developed and presented as a method for finding the unique solution of (1).

2.3 Duality

As we have seen, all three methods developed in this paper converge to a specific solution of the linear system (1), namely, to the projection of the starting point onto the solution space: $x_* = \Pi^{\mathbf{B}}_{\mathcal{L}}(x_0)$. Therefore, our methods solve the best approximation problem

$$\min_{x \in \mathbb{R}^n} \{\|x - x_0\|^2_{\mathbf{B}} \ : \ \mathbf{A}x = b\}. \tag{17}$$

The "dual" of the basic method with *unit stepsize* was studied in this context in [18]. The Fenchel dual of the best approximation problem is an unconstrained concave quadratic maximization problem of the form $\max_{y \in \mathbb{R}^m} D(y)$, where the dual objective D depends on $\mathbf{A}, b, \mathbf{B}$ and x_0. In [18] it was shown that

the basic method with unit stepsize closely related to a *dual method* (stochastic dual subspace ascent) performing iterations of the form

$$y_{k+1} = y_k + \mathbf{S}_k \lambda_k, \tag{18}$$

where $\mathbf{S}_k \sim \mathcal{D}$, and λ_k is chosen in such a way that the dual objective is as large as possible. Notice that the dual method in each iteration performs exact "line search" in a random subspace of $\mathbb{R}^m$ spanned by the columns of the random matrix $\mathbf{S}_k$, and passing through y_k. In particular, the iterates of the basic method with unit stepsize arise as affine images of the iterates of the dual method: $x_k = x_0 + \mathbf{B}^{-1}\mathbf{A}^\top y_k$.

In a similar fashion, it is possible to interpret the methods developed in this paper as images of appropriately designed dual methods. In [18], the authors focus on establishing convergence of various quantities, such as dual objective, primal-dual gap, primal objective and so on. They obtain the complexity $1/\lambda_{\min}^+$, which is identical to the rate we obtain here for the basic method with unit stepsize. However, their results require a stronger assumption on $\mathcal{D}$ (their assumption implies exactness, but not vice versa). We perform a much deeper analysis from the novel viewpoint of stochastic reformulations of linear systems, include a stepsize, and propose and analyze parallel and accelerated variants.

In the special case when $\mathbf{S}$ is chosen to be a random unit coordinate vector, (18) specializes to the *randomized coordinate descent method*, first analyzed by Leventhal and Lewis [24]. In the special case when $\mathbf{S}$ is chosen as a random column submatrix of the $m \times m$ identity matrix, (18) specializes to the *randomized Newton method* of Qu et al. [41]. Randomized coordinate descent methods are the state of the art methods for certain classes of convex optimization problems with a very large number of variables. The first complexity analysis beyond quadratics was performed in [54, 35, 46], a parallel method was developed in [47], duality was explored in [55] and acceleration in [14].

2.4 Randomized gossip algorithms

It was shown in [17, 26] that for a suitable matrix $\mathbf{A}$ encoding the structure of a graph, and for $b = 0$, the application of the randomized Kaczmarz and randomized block Kaczmarz methods to (17) lead to classical and new *randomized gossip algorithms* developed and studied in the signal processing literature, with new insights and novel proofs. Likewise, when applied in the same context, our new methods lead to new parallel and accelerated gossip algorithms.

2.5 Empirical risk minimization

Regularized empirical risk minimization (ERM) problems are optimization problems of the form

$$\min_{x \in \mathbb{R}^n} \frac{1}{m} \sum_{i=1}^{m} f_i(x) + g(x), \tag{19}$$

where f_i is a loss function and g a regularizer. Problems of this form are of key importance in machine learning and statistics [53]. Let $f_i(x) = 0$ if $\mathbf{A}_{i:}x = b_i$ and $f_i(x) = +\infty$ otherwise, further let $g(x) = \|x - x_0\|_{\mathbf{B}}^2$. In this setting, the ERM problem (19) becomes equivalent to (17). While quadratic regularizers similar to g are common in machine learning, zero/infinity loss functions are not used. For this reason, this specific instance of ERM was not studied in the machine learning literature. Since all our methods solve (17), they can be seen as stochastic algorithms for solving the ERM problem (19).

Since there is no reason to expect that any of our methods will satisfy $\mathbf{A}x_k = b$ for any finite k, the ERM objective value can remain to be equal to $+\infty$ throughout the entire iterative process. From this perspective, the value of the ERM objective is unsuitable as a measure of progress.

2.6 Matrix inversion and quasi-Newton methods

Given an invertible matrix $\mathbf{A}$, its inverse is the unique solution of the matrix equation $\mathbf{AX} = \mathbf{I}$. In [20] the authors have extended the "sketch and project" method [17] to this equation. In each iteration of the method, one projects the current iterate matrix $\mathbf{X}_k$, with respect to a weighted Frobenius norm, onto the sketched equation $\mathbf{S}_k^\top \mathbf{AX} = \mathbf{S}_k^\top \mathbf{I}$. This is a similar iterative process to the basic method with unit stepsize. The authors of [20] prove that the iterates of method converge to the inverse matrix at a linear rate, and detail connections of their method to quasi-Newton updates and approximate inverse preconditioning. A limited memory variant of the stochastic block BFGS method has been used to develop new efficient stochastoc quasi-Newton methods for emprical risk minimization problems appearing in machine learning [16].

It is possible to approach the problem $\mathbf{AX} = \mathbf{I}$ in the same way we approach the system (1) in our paper, writing down stochastic reformulations, and then developing new variants of the sketch and project method [20]: a basic method with a stepsize, and parallel and accelerated methods. This would lead to the development of new variants of stochastic quasi-Newton rules, notably parallel and accelerated block BFGS. We conjecture that these rules will have superior performance to classical BFGS in practice.

Similar extensions and improvements can be done in relation to the problem of computing the pseudoinverse of very large rectangular matrices [19].

3 Stochastic Reformulations of Linear Systems

In this section we formally derive the four stochastic formulations outlined in the introduction: *stochastic optimization*, *stochastic linear system*, *stochastic fixed point problem* and *probabilistic intersection*. Along the way we collect a number of results and observations which will be useful in the complexity analysis of our methods.

3.1 Projections

For a closed convex set $\mathcal{Y} \subseteq \mathbb{R}^n$, we let $\Pi_{\mathcal{Y}}^{\mathbf{B}}$ denote the projection operator onto $\mathcal{Y}$, in the $\mathbf{B}$-norm. That is, $\Pi_{\mathcal{Y}}^{\mathbf{B}}(x) \overset{\text{def}}{=} \arg\min_{y\in\mathbb{R}^n} \{\|y-x\|_{\mathbf{B}} \ : \ y \in \mathcal{Y}\}$. The $\mathbf{B}$-pseudoinverse of a matrix $\mathbf{M}$ is defined as

$$\mathbf{M}^{\dagger_{\mathbf{B}}} \overset{\text{def}}{=} \mathbf{B}^{-1}\mathbf{M}^\top(\mathbf{M}\mathbf{B}^{-1}\mathbf{M}^\top)^\dagger. \tag{20}$$

The projection onto $\mathcal{L} = \{x \ : \ \mathbf{A}x = b\}$ is given by

$$\Pi_{\mathcal{L}}^{\mathbf{B}}(x) = x - \mathbf{B}^{-1}\mathbf{A}^\top(\mathbf{A}\mathbf{B}^{-1}\mathbf{A}^\top)^\dagger(\mathbf{A}x-b) \overset{(20)}{=} x - \mathbf{A}^{\dagger_{\mathbf{B}}}(\mathbf{A}x-b). \tag{21}$$

Note that for $\mathbf{B} = \mathbf{I}$, we get $\mathbf{A}^{\dagger_{\mathbf{I}}} = \mathbf{A}^\top(\mathbf{A}\mathbf{A}^\top)^\dagger = \mathbf{A}^\dagger$, and hence the $\mathbf{I}$-pseudoinverse reduces to the standard Moore-Penrose pseudoinverse. The $\mathbf{B}$-pseudoinverse satisfies $\mathbf{A}^{\dagger_{\mathbf{B}}}b = \Pi_{\mathcal{L}}^{\mathbf{B}}(0) = \arg\min_x\{\|x\|_{\mathbf{B}} \ : \ \mathbf{A}x = b\}$.

3.2 Stochastic functions

Let $\mathcal{D}$ be an arbitrary distribution over $m \times q$ matrices, which we shall denote as $\mathbf{S}$. We shall write $\mathbf{S} \sim \mathcal{D}$ to say that $\mathbf{S}$ is drawn from $\mathcal{D}$. We shall often refer to matrix expressions involving $\mathbf{S}, \mathbf{A}$ and

$\mathbf{B}$. In order to keep the expressions brief throughout the paper, it will be useful to define

$$\mathbf{H} \overset{\text{def}}{=} \mathbf{S}(\mathbf{S}^\top \mathbf{A}\mathbf{B}^{-1}\mathbf{A}^\top \mathbf{S})^\dagger \mathbf{S}^\top, \tag{22}$$

and

$$\mathbf{Z} \overset{\text{def}}{=} \mathbf{A}^\top \mathbf{H}\mathbf{A} \overset{(22)}{=} \mathbf{A}^\top \mathbf{S}(\mathbf{S}^\top \mathbf{A}\mathbf{B}^{-1}\mathbf{A}^\top \mathbf{S})^\dagger \mathbf{S}^\top \mathbf{A}. \tag{23}$$

Notice that $\mathbf{B}^{-1}\mathbf{Z}$ is the projection, in the $\mathbf{B}$-norm, onto $\mathrm{Range}\left(\mathbf{B}^{-1}\mathbf{A}^\top \mathbf{S}\right)$. In particular,

$$(\mathbf{B}^{-1}\mathbf{Z})^2 = \mathbf{B}^{-1}\mathbf{Z} \qquad \text{and} \qquad \mathbf{Z}\mathbf{B}^{-1}\mathbf{Z} = \mathbf{Z}. \tag{24}$$

Given $\mathbf{S} \sim \mathcal{D}$, we define the *stochastic (random) function*

$$f_{\mathbf{S}}(x) \overset{\text{def}}{=} \frac{1}{2}\|\mathbf{A}x - b\|^2_{\mathbf{H}} = \frac{1}{2}(\mathbf{A}x - b)^\top \mathbf{H}(\mathbf{A}x - b). \tag{25}$$

By combining (25) and (23), this can be also written in the form

$$f_{\mathbf{S}}(x) = \frac{1}{2}(x - x_*)^\top \mathbf{Z}(x - x_*), \qquad x \in \mathbb{R}^n,\ x_* \in \mathcal{L}. \tag{26}$$

For each $x, h \in \mathbb{R}^n$ we have the expansion $f_{\mathbf{S}}(x+h) = f_{\mathbf{S}}(x) + \langle \nabla f_{\mathbf{S}}(x), h\rangle_{\mathbf{B}} + \frac{1}{2}\left\langle (\nabla^2 f_{\mathbf{S}})h, h\right\rangle_{\mathbf{B}}$, where

$$\nabla f_{\mathbf{S}}(x) = \mathbf{B}^{-1}\mathbf{A}^\top \mathbf{H}(\mathbf{A}x - b) \qquad \text{and} \qquad \nabla^2 f_{\mathbf{S}} = \mathbf{B}^{-1}\mathbf{Z} \tag{27}$$

are the gradient and Hessian of $f_{\mathbf{S}}$ with respect to the $\mathbf{B}$-inner product, respectively.[1] In view of (23) and (27), the gradient can also be written as

$$\nabla f_{\mathbf{S}}(x) = \mathbf{B}^{-1}\mathbf{Z}(x - x_*), \qquad x \in \mathbb{R}^n,\ x_* \in \mathcal{L}. \tag{28}$$

Identities (29) in the following lemma explain why algorithm (6) can be equivalently written as stochastic gradient descent (7), stochastic Newton method (8), stochastic fixed point method (10), and stochastic projection method (11). For instance, the identity $(\nabla^2 f_{\mathbf{S}})\nabla f_{\mathbf{S}}(x) = \nabla f_{\mathbf{S}}(x)$ means that the stochastic gradients of $f_{\mathbf{S}}$ are eigenvectors of the stochastic Hessian $\nabla^2 f_{\mathbf{S}}$, corresponding to eigenvalue one.

Lemma 3.1. For all $x \in \mathbb{R}^n$, we have

$$\nabla f_{\mathbf{S}}(x) = (\nabla^2 f_{\mathbf{S}})\nabla f_{\mathbf{S}}(x) = (\nabla^2 f_{\mathbf{S}})^{\dagger_{\mathbf{B}}}\nabla f_{\mathbf{S}}(x) = x - \Pi^{\mathbf{B}}_{\mathcal{L}_{\mathbf{S}}}(x) = \mathbf{B}^{-1}\mathbf{A}^\top \mathbf{H}(\mathbf{A}x - b). \tag{29}$$

Moreover,

$$f_{\mathbf{S}}(x) = \frac{1}{2}\|\nabla f_{\mathbf{S}}(x)\|^2_{\mathbf{B}}. \tag{30}$$

If $\mathcal{L}_{\mathbf{S}}$ is the set of minimizers of $f_{\mathbf{S}}$, then $\mathcal{L} \subseteq \mathcal{L}_{\mathbf{S}}$, and

(i) $\mathcal{L}_{\mathbf{S}} = \{x \ :\ f_{\mathbf{S}}(x) = 0\} = \{x \ :\ \nabla f_{\mathbf{S}}(x) = 0\}$

(ii) $\mathcal{L}_{\mathbf{S}} = x_* + \mathrm{Null}\left(\mathbf{B}^{-1}\mathbf{Z}\right)$ for all $x_* \in \mathcal{L}$

(iii) $\mathcal{L}_{\mathbf{S}} = \{x \ :\ \mathbf{B}^{-1}\mathbf{A}^\top \mathbf{H}\mathbf{A}x = \mathbf{B}^{-1}\mathbf{A}^\top \mathbf{H}b\}$ (see (3))

[1] If $\mathbf{B} = \mathbf{I}$, then $\langle\cdot,\cdot\rangle_{\mathbf{B}}$ is the standard Euclidean inner product, and we recover formulas for the standard gradient and Hessian. Note that $\mathbf{B}^{-1}\mathbf{Z}$ is both self-adjoint and positive semidefinite with respect to the $\mathbf{B}$-inner product. Indeed, for all $x, y \in \mathbb{R}^n$ we have $\left\langle \mathbf{B}^{-1}\mathbf{Z}x, y\right\rangle_{\mathbf{B}} = \langle \mathbf{Z}x, y\rangle_{\mathbf{I}} = \langle x, \mathbf{Z}y\rangle_{\mathbf{I}} = \left\langle x, \mathbf{B}^{-1}\mathbf{Z}y\right\rangle_{\mathbf{B}}$, and $\left\langle \mathbf{B}^{-1}\mathbf{Z}x, x\right\rangle_{\mathbf{B}} = \langle \mathbf{Z}x, x\rangle_{\mathbf{I}} \geq 0$.

(iv) $\mathcal{L}_{\mathbf{S}} = \{x \ : \ \mathbf{S}^\top \mathbf{A} x = \mathbf{S}^\top b\}$ (see (5))

Finally, for all $x \in \mathbb{R}^n$ we have the identity

$$f_{\mathbf{S}}(x - \nabla f_{\mathbf{S}}(x)) = 0. \tag{31}$$

Proof. Pick any $x_* \in \mathcal{L}$. First, we have $\Pi^{\mathbf{B}}_{\mathcal{L}_{\mathbf{S}}}(x) \overset{(21)}{=} x - \mathbf{B}^{-1}\mathbf{A}^\top \mathbf{H}(\mathbf{A}x - b) \overset{(27)}{=} x - \nabla f_{\mathbf{S}}(x)$. To establish (29), it now only remains to consider the two expressions involving the Hessian. We have

$$\nabla^2 f_{\mathbf{S}} \nabla f_{\mathbf{S}}(x) \overset{(27)+(28)}{=} \mathbf{B}^{-1}\mathbf{Z}\mathbf{B}^{-1}\mathbf{Z}(x - x_*) \overset{(24)}{=} \mathbf{B}^{-1}\mathbf{Z}(x - x_*) \overset{(28)}{=} \nabla f_{\mathbf{S}}(x),$$

and

$$\begin{aligned}
(\nabla^2 f_{\mathbf{S}})^{\dagger_{\mathbf{B}}} \nabla f_{\mathbf{S}}(x) &\overset{(20)}{=} \mathbf{B}^{-1}(\nabla^2 f_{\mathbf{S}})^\top \left((\nabla^2 f_{\mathbf{S}})\mathbf{B}^{-1}(\nabla^2 f_{\mathbf{S}})^\top\right)^\dagger \nabla f_{\mathbf{S}}(x) \\
&\overset{(27)}{=} \mathbf{B}^{-1}(\mathbf{B}^{-1}\mathbf{Z})^\top \left((\mathbf{B}^{-1}\mathbf{Z})\mathbf{B}^{-1}(\mathbf{B}^{-1}\mathbf{Z})^\top\right)^\dagger \mathbf{B}^{-1}\mathbf{Z}(x - x_*) \\
&= \mathbf{B}^{-1}\mathbf{Z}\mathbf{B}^{-1}\left(\mathbf{B}^{-1}\mathbf{Z}\mathbf{B}^{-1}\mathbf{Z}\mathbf{B}^{-1}\right)^\dagger \mathbf{B}^{-1}\mathbf{Z}(x - x_*) \\
&\overset{(24)}{=} \left(\mathbf{B}^{-1}\mathbf{Z}\mathbf{B}^{-1}\right)\left(\mathbf{B}^{-1}\mathbf{Z}\mathbf{B}^{-1}\right)^\dagger \left(\mathbf{B}^{-1}\mathbf{Z}\mathbf{B}^{-1}\right)\mathbf{B}(x - x_*) \\
&= \mathbf{B}^{-1}\mathbf{Z}(x - x_*) \\
&\overset{(28)}{=} \nabla f_{\mathbf{S}}(x).
\end{aligned}$$

Identity (30) follows from

$$\frac{1}{2}\|\nabla f_{\mathbf{S}}(x)\|^2_{\mathbf{B}} \overset{(28)}{=} \frac{1}{2}(x - x_*)^\top \mathbf{Z}\mathbf{B}^{-1}\mathbf{Z}(x - x_*) \overset{(24)}{=} \frac{1}{2}(x - x_*)^\top \mathbf{Z}(x - x_*) \overset{(26)}{=} f_{\mathbf{S}}(x).$$

If $x \in \mathcal{L}$, then by picking $x_* = x$ in (28), we see that $x \in \mathcal{L}_{\mathbf{S}}$. It remains to show that the sets defined in (i)–(iv) are identical. Equivalence between (i) and (ii) follows from (28). Now consider (ii) and (iii). Any $x_* \in \mathcal{L}$ belongs to the set defined in (iii), which follows immediately by substituting $b = \mathbf{A}x_*$. The rest follows after observing the nullspaces are identical. In order to show that (iii) and (iv) are equivalent, it suffices to compute $\Pi^{\mathbf{B}}_{\mathcal{L}_{\mathbf{S}}}(x)$ and observe that $\Pi^{\mathbf{B}}_{\mathcal{L}_{\mathbf{S}}}(x) = x$ if and only if x belongs to the set defined in (iii).

It remains to establish (31). In view of (i), it suffices to show that $x - \nabla f_{\mathbf{S}}(x) \in \mathcal{L}_{\mathbf{S}}$. However, from (29) we know that $x - \nabla f_{\mathbf{S}}(x) = \Pi^{\mathbf{B}}_{\mathcal{L}_{\mathbf{S}}}(x) \in \mathcal{L}_{\mathbf{S}}$. □

3.3 Stochastic reformulation

In order to proceed, we shall enforce a basic assumption on $\mathcal{D}$.

Assumption 3.2 (Finite mean)**.** The random matrix $\mathbf{H}$ has a mean. That is, the matrix $\mathrm{E}_{\mathbf{S}\sim\mathcal{D}}[\mathbf{H}] = \mathrm{E}_{\mathbf{S}\sim\mathcal{D}}\left[\mathbf{S}(\mathbf{S}^\top \mathbf{A}\mathbf{B}^{-1}\mathbf{A}^\top \mathbf{S})^\dagger \mathbf{S}^\top\right]$ has finite entries.

This is an assumption on $\mathcal{D}$ since a suitable distribution satisfying it exists for all $\mathbf{A} \in \mathbb{R}^{m\times n}$ and $\mathbf{B} \succ 0$. Note that if the assumption holds, then $\mathrm{E}[\mathbf{H}]$ is symmetric and positive semidefinite. We shall enforce this assumption throughout the paper and hence will not henceforth refer to it.

Example 1. Let $\mathcal{D}$ be the uniform distribution over unit basis vectors in $\mathbb{R}^m$. That is, $\mathbf{S} = e_i$ (the ith unit basis vector in $\mathbb{R}^m$) with probability $1/m$. Then

$$\mathrm{E}\left[\mathbf{H}\right] = \sum_{i=1}^{m} \frac{1}{m} e_i (\mathbf{A}_{i:} \mathbf{B}^{-1} \mathbf{A}_{i:}^\top)^\dagger e_i^\top = \frac{1}{m} \mathbf{Diag}\left(\alpha_1, \ldots, \alpha_m\right),$$

where $\alpha_i = 1/\|\mathbf{A}_{i:}^\top\|^2_{\mathbf{B}^{-1}}$ for $i = 1, 2, \ldots, m$. If $\mathbf{A}$ has nonzero rows, then $\mathrm{E}\left[\mathbf{H}\right] \succ 0$.

In this paper we reformulate the linear system (1) as the *stochastic optimization problem*

$$\boxed{\min_{x \in \mathbb{R}^n} \left\{ f(x) \overset{\text{def}}{=} \mathrm{E}_{\mathbf{S} \sim \mathcal{D}} \left[f_{\mathbf{S}}(x) \right] \right\}} \tag{32}$$

Under Assumption 3.2, the expectation in (32) is finite for all x, and hence f is well defined. The following is a direct consequence of Lemma 3.1. We shall use these formulas throughout the paper.

Lemma 3.3 (Representations of f)**.** Function f defined in (32) can be represented in multiple ways:

$$f(x) = \frac{1}{2}\mathrm{E}\left[\|x - \Pi^{\mathbf{B}}_{\mathcal{L}_{\mathbf{S}}}(x)\|^2_{\mathbf{B}}\right] = \frac{1}{2}\mathrm{E}\left[\|\nabla f_{\mathbf{S}}(x)\|^2_{\mathbf{B}}\right].$$

Moreover,

$$f(x) \overset{(25)}{=} \frac{1}{2}\|\mathbf{A}x - b\|^2_{\mathrm{E}[\mathbf{H}]} = \frac{1}{2}(\mathbf{A}x - b)^\top \mathrm{E}\left[\mathbf{H}\right](\mathbf{A}x - b), \tag{33}$$

and for any $x_* \in \mathcal{L}$ we can write

$$f(x) = \frac{1}{2}(x - x_*)^\top \mathrm{E}\left[\mathbf{Z}\right](x - x_*). \tag{34}$$

Since $\mathrm{E}\left[\mathbf{H}\right] \succeq 0$, f is a convex quadratic function. Moreover, f is nonnegative. The gradient and Hessian of f (with respect to the $\mathbf{B}$-inner product) are given by

$$\nabla f(x) = \mathrm{E}_{\mathbf{S} \sim \mathcal{D}}\left[\nabla f_{\mathbf{S}}(x)\right] = \mathbf{B}^{-1}\mathrm{E}\left[\mathbf{Z}\right](x - x_*), \qquad \text{and} \qquad \nabla^2 f = \mathbf{B}^{-1}\mathrm{E}\left[\mathbf{Z}\right], \tag{35}$$

respectively, where x_* is any point in $\mathcal{L}$.

The set of minimizers of f, denoted $\mathcal{X}$, can be represented in several ways, as captured by our next result. It immediately follows that the four stochastic formulations mentioned in the introduction are equivalent.

Theorem 3.4 (Equivalence of stochastic formulations)**.** Let $x_* \in \mathcal{L}$. The following sets are identical:

(i) $\mathcal{X} = \arg\min f(x) = \{x \;:\; f(x) = 0\} = \{x \;:\; \nabla f(x) = 0\}$ (see (2))

(ii) $\mathcal{X} = \{x \;:\; \mathbf{B}^{-1}\mathbf{A}^\top \mathrm{E}\left[\mathbf{H}\right]\mathbf{A}x = \mathbf{B}^{-1}\mathbf{A}^\top \mathrm{E}\left[\mathbf{H}\right] b\} = x_* + \mathrm{Null}\left(\mathrm{E}\left[\mathbf{Z}\right]\right)$ (see (3))

(iii) $\mathcal{X} = \{x \;:\; \mathrm{E}\left[\Pi^{\mathbf{B}}_{\mathcal{L}_{\mathbf{S}}}(x)\right] = x\}$ (see (4))

(iv) $\mathcal{X} = \{x \;:\; \mathrm{Prob}(x \in \mathcal{L}_{\mathbf{S}}) = 1\}$ (see (5))

As a consequence, the stochastic problems (2), (3), (4), and (5) are equivalent (i.e., their solutions sets are identical). Moreover, the set $\mathcal{X}$ does not depend on $\mathbf{B}$.

Proof. As f is convex, nonnegative and achieving the value of zero (since $\mathcal{L} \neq \emptyset$), the sets in (i) are all identical. We shall now show that the sets defined in (ii)–(iv) are equal to that defined in (i). Using the formula for the gradient from (35), we see that $\{x \ : \ \nabla f(x) = 0\} = \{x \ : \ \mathbf{B}^{-1}\mathrm{E}\left[\mathbf{Z}\right](x - x_*) = 0\} = \{x \ : \ \mathrm{E}\left[\mathbf{Z}\right](x - x_*) = 0\} = x_* + \{h \ : \ \mathrm{E}\left[\mathbf{Z}\right]h = 0\} = x_* + \mathrm{Null}\left(\mathrm{E}\left[\mathbf{Z}\right]\right)$, which shows that (i) and (ii) are the same. Equivalence of (i) and (iii) follows by taking expectations in (29) to obtain $\nabla f(x) = \mathrm{E}\left[\nabla f_{\mathbf{S}}(x)\right] \overset{(29)}{=} \mathrm{E}\left[x - \Pi^{\mathbf{B}}_{\mathcal{L}_{\mathbf{S}}}(x)\right]$.

It remains to establish equivalence between (i) and (iv). Let

$$\mathcal{X} = \{x \ : \ f(x) = 0\} \overset{\text{(Lemma 3.3)}}{=} \left\{x \ : \ \mathrm{E}\left[\left\|x - \Pi^{\mathbf{B}}_{\mathcal{L}_{\mathbf{S}}}(x)\right\|^2_{\mathbf{B}}\right] = 0\right\} \tag{36}$$

and let $\mathcal{X}'$ be the set from (iv). For easier reference, let $\xi_{\mathbf{S}}(x) \overset{\text{def}}{=} \left\|x - \Pi^{\mathbf{B}}_{\mathcal{L}_{\mathbf{S}}}(x)\right\|^2_{\mathbf{B}}$. The following three probabilistic events are identical:

$$\left[x \in \mathcal{L}_{\mathbf{S}}\right] = \left[x = \Pi^{\mathbf{B}}_{\mathcal{L}_{\mathbf{S}}}(x)\right] = \left[\xi_{\mathbf{S}}(x) = 0\right]. \tag{37}$$

Therefore, if $x \in \mathcal{X}'$, then the random variable $\xi_{\mathbf{S}}(x)$ is equal to zero with probability 1, and hence $x \in \mathcal{X}$. Let us now establish the reverse inclusion. First, let $1_{[\xi_{\mathbf{S}}(x) \geq t]}$ be the indicator function of the event $[\xi_{\mathbf{S}}(x) \geq t]$. Note that since $\xi_{\mathbf{S}}(x)$ is a nonnegative random variable, for all $t \in \mathbb{R}$ we have the inequality

$$\xi_{\mathbf{S}}(x) \geq t 1_{\xi_{\mathbf{S}}(x) \geq t}. \tag{38}$$

Now take $x \in \mathcal{X}$ and consider $t > 0$. By taking expectations in (38), we obtain

$$0 = \mathrm{E}\left[\xi_{\mathbf{S}}(x)\right] \geq \mathrm{E}\left[t 1_{\xi_{\mathbf{S}}(x) \geq t}\right] = t\mathrm{E}\left[1_{\xi_{\mathbf{S}}(x) \geq t}\right] = t\,\mathrm{Prob}(\xi_{\mathbf{S}}(x) \geq t),$$

which implies that $\mathrm{Prob}(\xi_{\mathbf{S}}(x) \geq t) = 0$. Now choose $t_i = 1/i$ for $i = 1, 2, \ldots$ and note that the event $[\xi_{\mathbf{S}}(x) > 0]$ can be written as

$$[\xi_{\mathbf{S}}(x) > 0] = \bigcup_{i=1}^{\infty} [\xi_{\mathbf{S}}(x) \geq t_i].$$

Therefore, by the union bound, $\mathrm{Prob}(\xi_{\mathbf{S}}(x) > 0) = 0$, which immediately implies that $\mathrm{Prob}(\xi_{\mathbf{S}}(x) = 0) = 1$. From (37) we conclude that $x \in \mathcal{X}'$.

That $\mathcal{X}$ does not depend on $\mathbf{B}$ follows from representation (iv). □

3.4 Exactness of the Reformulations

In this section ask the following question: when are the stochastic formulations (2), (3), (4), (5) equivalent to the linear system (1)? This leads to the concept of *exactness*, captured by the following assumption.

Assumption 3.5 (Exactness)**.** Stochastic reformulations (2), (3), (4), (5) of problem (1) are *exact*. That is, $\mathcal{X} = \mathcal{L}$.

We do not need this assumption for all our results, and hence we will specifically invoke it when needed. For future reference in the paper, it will be useful to be able to draw upon several equivalent characterizations of exactness.

Theorem 3.6 (Exactness)**.** The following statements are equivalent:

(i) Assumption 3.5 holds

(ii) $\mathrm{Null}\left(\mathrm{E}\left[\mathbf{Z}\right]\right) = \mathrm{Null}\left(\mathbf{A}\right)$

(iii) $\mathrm{Null}\left(\mathbf{B}^{-1/2}\mathrm{E}\left[\mathbf{Z}\right]\mathbf{B}^{-1/2}\right) = \mathrm{Null}\left(\mathbf{A}\mathbf{B}^{-1/2}\right)$

(iv) $\mathrm{Range}\left(\mathbf{A}\right) \cap \mathrm{Null}\left(\mathrm{E}\left[\mathbf{H}\right]\right) = \{0\}$

Proof. Choose any $x_* \in \mathcal{L}$. We know that $\mathcal{L} = x_* + \mathrm{Null}\left(\mathbf{A}\right)$. On the other hand, Theorem 3.4 says that $\mathcal{X} = x_* + \mathrm{Null}\left(\mathrm{E}\left[\mathbf{Z}\right]\right)$. This establishes equivalence of (i) and (ii). If (ii) holds, then $\mathrm{Null}\left(\mathbf{A}\right) = \mathrm{Null}\left(\mathrm{E}\left[\mathbf{Z}\right]\right) = \mathrm{Null}\left(\mathbf{B}^{-1/2}\mathrm{E}\left[\mathbf{Z}\right]\right)$, and (iii) follows. If (iii) holds, then $\mathrm{Null}\left(\mathbf{A}\right) = \mathrm{Null}\left(\mathbf{B}^{-1/2}\mathrm{E}\left[\mathbf{Z}\right]\right) = \mathrm{Null}\left(\mathrm{E}\left[\mathbf{Z}\right]\right)$, proving (ii). We now show that (ii) and (iv) are equivalent. First, note that $\mathrm{E}\left[\mathbf{Z}\right] = \mathbf{A}^\top(\mathrm{E}\left[\mathbf{H}\right])^{1/2}(\mathrm{E}\left[\mathbf{H}\right])^{1/2}\mathbf{A}$. Therefore, $\mathrm{Null}\left(\mathrm{E}\left[\mathbf{Z}\right]\right) = \mathrm{Null}\left((\mathrm{E}\left[\mathbf{H}\right])^{1/2}\mathbf{A}\right)$. Moreover, we know that a) $\mathrm{Null}\left((\mathrm{E}\left[\mathbf{H}\right])^{1/2}\mathbf{A}\right) = \mathrm{Null}\left(\mathbf{A}\right)$ if and only if $\mathrm{Range}\left(\mathbf{A}\right) \cap \mathrm{Null}\left((\mathrm{E}\left[\mathbf{H}\right])^{1/2}\right) = \{0\}$, and b) $\mathrm{Null}\left((\mathrm{E}\left[\mathbf{H}\right])^{1/2}\right) = \mathrm{Null}\left(\mathrm{E}\left[\mathbf{H}\right]\right)$. It remains to combine these observations. □

We now list two sufficient conditions for exactness.

Lemma 3.7 (Sufficient conditions)**.** Any of these conditions implies that Assumption 3.5 is satisfied:

(i) $\mathrm{E}\left[\mathbf{H}\right] \succ 0$

(ii) $\mathrm{Null}\left(\mathrm{E}\left[\mathbf{H}\right]\right) \subseteq \mathrm{Null}\left(\mathbf{A}^\top\right)$

Proof. If (i) holds, then $\mathrm{Null}\left(\mathrm{E}\left[\mathbf{Z}\right]\right) = \mathrm{Null}\left(\mathbf{A}^\top\mathrm{E}\left[\mathbf{H}\right]\mathbf{A}\right) = \mathrm{Null}\left(\mathbf{A}\right)$, we have exactness by applying Theorem 3.6. Finally, (ii) implies statement (iv) in Theorem 3.6, and hence exactness follows. □

4 Basic Method

We propose solving (33) by Algorithm 1. In the rest of this section we offer several equivalent interpretations of the method.

Algorithm 1 Basic Method

1: **Parameters:** distribution $\mathcal{D}$ from which to sample matrices; positive definite matrix $\mathbf{B} \in \mathbb{R}^{n\times n}$; stepsize/relaxation parameter $\omega \in \mathbb{R}$
2: Choose $x_0 \in \mathbb{R}^n$ ▷ Initialization
3: **for** $k = 0, 1, 2, \ldots$ **do**
4: Draw a fresh sample $\mathbf{S}_k \sim \mathcal{D}$
5: Set $x_{k+1} = x_k - \omega\mathbf{B}^{-1}\mathbf{A}^\top\mathbf{S}_k(\mathbf{S}_k^\top\mathbf{A}\mathbf{B}^{-1}\mathbf{A}^\top\mathbf{S}_k)^\dagger\mathbf{S}_k^\top(\mathbf{A}x_k - b)$

Remark 1. Since $\mathbf{S}$ is random, the matrices $\mathbf{H} = \mathbf{S}(\mathbf{S}^\top\mathbf{A}\mathbf{B}^{-1}\mathbf{A}^\top\mathbf{S})^\dagger\mathbf{S}^\top$ and $\mathbf{Z} = \mathbf{A}^\top\mathbf{H}\mathbf{A}$ are also random. At iteration k of our algorithm, we sample matrix $\mathbf{S}_k \sim \mathcal{D}$ and perform an update step. It will be useful to to use the notation $\mathbf{H}_k = \mathbf{S}_k(\mathbf{S}_k^\top\mathbf{A}\mathbf{B}^{-1}\mathbf{A}^\top\mathbf{S}_k)^\dagger\mathbf{S}_k^\top$ and $\mathbf{Z}_k = \mathbf{A}^\top\mathbf{H}_k\mathbf{A}$. In this notation, Algorithm 1 can be written in the form $x_{k+1} = x_k - \omega\mathbf{B}^{-1}\mathbf{A}^\top\mathbf{H}_k(\mathbf{A}x_k - b)$.

In the rest of this section we analyze Algorithm 1.

4.1 Condition number of the stochastic reformulation

Recall that the Hessian of f is given by

$$\nabla^2 f = \mathrm{E}_{\mathbf{S}\sim\mathcal{D}}\left[\nabla^2 f_{\mathbf{S}}\right] = \mathbf{B}^{-1}\mathrm{E}\left[\mathbf{Z}\right]. \tag{39}$$

Since $\mathbf{B}^{-1}\mathrm{E}\left[\mathbf{Z}\right]$ is not symmetric (although it is self-adjoint with respect to the $\mathbf{B}$-inner product), it will be more convenient to instead study the spectral properties of the related matrix $\mathbf{B}^{-1/2}\mathrm{E}\left[\mathbf{Z}\right]\mathbf{B}^{-1/2}$. Note that this matrix is symmetric, and has the same spectrum as $\mathbf{B}^{-1}\mathrm{E}\left[\mathbf{Z}\right]$. Let

$$\mathbf{W} \overset{\text{def}}{=} \mathbf{B}^{-1/2}\mathrm{E}\left[\mathbf{Z}\right]\mathbf{B}^{-1/2} = \mathbf{U}\Lambda\mathbf{U}^\top = \sum_{i=1}^n \lambda_i u_i u_i^\top \tag{40}$$

be the eigenvalue decomposition of $\mathbf{W}$, where $\mathbf{U} = [u_1, \ldots, u_n] \in \mathbb{R}^{n\times n}$ is an orthonormal matrix composed of eigenvectors, and $\Lambda = \mathbf{Diag}\left(\lambda_1, \lambda_2, \ldots, \lambda_n\right)$ is a diagonal matrix of eigenvalues. Assume without loss of generality that the eigenvalues are ordered from largest to smallest: $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0$. We shall often write $\lambda_{\max} = \lambda_1$ to denote the largest eigenvalue, and $\lambda_{\min} = \lambda_n$ for the smallest eigenvalue.

Lemma 4.1. $0 \leq \lambda_i \leq 1$ for all i.

Proof. Since $\mathbf{B}^{-1/2}\mathbf{Z}\mathbf{B}^{-1/2}$ is symmetric positive semidefinite, so is its expectation $\mathbf{W}$, implying that $\lambda_i \geq 0$ for all i. Further, note that $\mathbf{B}^{-1/2}\mathbf{Z}\mathbf{B}^{-1/2}$ is a projection matrix. Indeed, it is the projection (in the standard $\mathbf{I}$-norm) onto $\mathrm{Range}\left(\mathbf{B}^{-1/2}\mathbf{A}^\top\mathbf{S}\right)$. Therefore, its eigenvalues are all zeros or ones. Since the map $\mathbf{X} \mapsto \lambda_{\max}(\mathbf{X})$ is convex, by Jensen's inequality we get

$$\lambda_{\max}(\mathbf{W}) = \lambda_{\max}\left(\mathrm{E}\left[\mathbf{B}^{-1/2}\mathbf{Z}\mathbf{B}^{-1/2}\right]\right) \leq \mathrm{E}\left[\lambda_{\max}(\mathbf{B}^{-1/2}\mathbf{Z}\mathbf{B}^{-1/2})\right] \leq 1.$$

□

It follows from Assumption 3.5 that $\lambda_{\max} > 0$. Indeed, if we assume that $\lambda_i = 0$ for all i, then from Theorem 3.6 and the fact that $\mathrm{Null}\left(\mathbf{W}\right) = \mathrm{Range}\left(u_i \,:\, \lambda_i = 0\right)$ we conclude that $\mathrm{Null}\left(\mathbf{A}\mathbf{B}^{-1/2}\right) = \mathbb{R}^n$, which in turn implies that $\mathrm{Null}\left(\mathbf{A}\right) = \mathbb{R}^n$. This can only happen if $\mathbf{A} = 0$, which is a trivial case we excluded from consideration in this paper by assumption.

Now, let j be the largest index for which $\lambda_j > 0$. We shall often write $\lambda_{\min}^+ = \lambda_j$. If all eigenvalues $\{\lambda_i\}$ are positive, then $j = n$.

We now define the *condition number* of problem (32) to be the quantity

$$\zeta \overset{\text{def}}{=} \|\mathbf{W}\|\|\mathbf{W}^\dagger\| = \frac{\lambda_{\max}}{\lambda_{\min}^+}. \tag{41}$$

Lemma 4.2 (Quadratic bounds)**.** For all $x \in \mathbb{R}^n$ and $x_* \in \mathcal{L}$ we have

$$\lambda_{\min}^+ \cdot f(x) \leq \frac{1}{2}\|\nabla f(x)\|_{\mathbf{B}}^2 \leq \lambda_{\max} \cdot f(x). \tag{42}$$

and

$$f(x) \leq \frac{\lambda_{\max}}{2}\|x - x_*\|_{\mathbf{B}}^2. \tag{43}$$

Moreover, if Assumption 3.5 holds, then for all $x \in \mathbb{R}^n$ and $x_* = \Pi_{\mathcal{L}}^{\mathbf{B}}(x)$ we have

$$\frac{\lambda_{\min}^+}{2}\|x - x_*\|_{\mathbf{B}}^2 \leq f(x). \tag{44}$$

Proof. In view of (34) and (40), we obtain a spectral characterization of f:

$$f(x) = \frac{1}{2}\sum_{i=1}^{n} \lambda_i \left(u_i^\top \mathbf{B}^{1/2}(x - x_*)\right)^2, \tag{45}$$

where x_* is any point in $\mathcal{L}$. On the other hand, in view of (35) and (40), we have

$$\begin{aligned}
\|\nabla f(x)\|_{\mathbf{B}}^2 &= \|\mathbf{B}^{-1}\mathrm{E}\left[\mathbf{Z}\right](x - x_*)\|_{\mathbf{B}}^2 = (x - x_*)^\top \mathrm{E}\left[\mathbf{Z}\right]\mathbf{B}^{-1}\mathrm{E}\left[\mathbf{Z}\right](x - x_*) \\
&= (x - x_*)^\top \mathbf{B}^{1/2}(\mathbf{B}^{-1/2}\mathrm{E}\left[\mathbf{Z}\right]\mathbf{B}^{-1/2})(\mathbf{B}^{-1/2}\mathrm{E}\left[\mathbf{Z}\right]\mathbf{B}^{-1/2})\mathbf{B}^{1/2}(x - x_*) \\
&= (x - x_*)^\top \mathbf{B}^{1/2}\mathbf{U}(\mathbf{U}^\top\mathbf{B}^{-1/2}\mathrm{E}\left[\mathbf{Z}\right]\mathbf{B}^{-1/2}\mathbf{U})(\mathbf{U}^\top\mathbf{B}^{-1/2}\mathrm{E}\left[\mathbf{Z}\right]\mathbf{B}^{-1/2}\mathbf{U})\mathbf{U}^\top\mathbf{B}^{1/2}(x - x_*) \\
&\overset{(40)}{=} (x - x_*)^\top \mathbf{B}^{1/2}\mathbf{U}\Lambda^2\mathbf{U}^\top\mathbf{B}^{1/2}(x - x_*) \\
&= \sum_{i=1}^{n} \lambda_i^2 \left(u_i^\top \mathbf{B}^{1/2}(x - x_*)\right)^2.
\end{aligned} \tag{46}$$

Inequality (42) follows by comparing (45) and (46), using the bounds $\lambda_{\min}^+\lambda_i \leq \lambda_i^2 \leq \lambda_{\max}\lambda_i$, which hold for i for which $\lambda_i > 0$.

We now move to the bounds involving norms. First, note that for any $x_* \in \mathcal{L}$ we have

$$f(x) \overset{(34)}{=} \frac{1}{2}(x - x_*)^\top \mathrm{E}\left[\mathbf{Z}\right](x - x_*) = \frac{1}{2}(\mathbf{B}^{1/2}(x - x_*))^\top(\mathbf{B}^{-1/2}\mathrm{E}\left[\mathbf{Z}\right]\mathbf{B}^{-1/2})\mathbf{B}^{1/2}(x - x_*). \tag{47}$$

The upper bound follows by applying the inequality $\mathbf{B}^{-1/2}\mathrm{E}\left[\mathbf{Z}\right]\mathbf{B}^{-1/2} \preceq \lambda_{\max}\mathbf{I}$. If $x_* = \Pi_{\mathcal{L}}^{\mathbf{B}}(x)$, then in view of (21), we have $\mathbf{B}^{1/2}(x - x_*) \in \mathrm{Range}\left(\mathbf{B}^{-1/2}\mathbf{A}^\top\right)$. Applying Lemma B.1 to (47), we get the lower bound. □

Remark 2. Bounds such as those in Lemma 4.2 are often seen in convex optimization. In particular, if $\phi : \mathbb{R}^n \to \mathbb{R}$ is a μ-strongly convex and L-smooth function, then $\mu(\phi(x) - \phi^*) \leq \frac{1}{2}\|\nabla\phi(x)\|^2 \leq L(\phi(x) - \phi^*)$ for all $x \in \mathbb{R}^n$, where $\phi^* = \min_x \phi(x)$. In our case, the optimal objective value is zero. The presence of $\mathbf{B}$-norm is due to us defining gradients using the $\mathbf{B}$-inner product. Moreover, it is the case that f is $\lambda_{\max}$-smooth, which explains the upper bound. However, f is not necessarily μ-strongly convex for any $\mu > 0$, since $\mathrm{E}\left[\mathbf{Z}\right]$ is not necessarily positive definite. However, we still obtain a nontrivial lower bound.

4.2 Convergence of expected iterates

We now present a fundamental theorem precisely describing the evolution of the expected iterates of the basic method.

Theorem 4.3 (Convergence of expected iterates)**.** Choose any $x_0 \in \mathbb{R}^n$ and let $\{x_k\}$ be the random iterates produced by Algorithm 1.

1. Let $x_* \in \mathcal{L}$ be chosen arbitrarily. Then

$$\mathrm{E}\left[x_{k+1} - x_*\right] = \left(\mathbf{I} - \omega \mathbf{B}^{-1}\mathrm{E}\left[\mathbf{Z}\right]\right)\mathrm{E}\left[x_k - x_*\right]. \tag{48}$$

Moreover, by transforming the error via the linear mapping $h \to \mathbf{U}^\top \mathbf{B}^{1/2} h$, this can be written in the form

$$\mathrm{E}\left[\mathbf{U}^\top \mathbf{B}^{1/2}(x_k - x_*)\right] = (\mathbf{I} - \omega\Lambda)^k \mathbf{U}^\top \mathbf{B}^{1/2}(x_0 - x_*), \tag{49}$$

which is separable in the coordinates of the transformed error:

$$\mathrm{E}\left[u_i^\top \mathbf{B}^{1/2}(x_k - x_*)\right] = (1 - \omega\lambda_i)^k u_i^\top \mathbf{B}^{1/2}(x_0 - x_*), \qquad i = 1, 2, \ldots, n. \tag{50}$$

Finally,

$$\|\mathrm{E}\left[x_k - x_*\right]\|_{\mathbf{B}}^2 = \sum_{i=1}^{n} (1 - \omega\lambda_i)^{2k} \left(u_i^\top \mathbf{B}^{1/2}(x_0 - x_*)\right)^2. \tag{51}$$

2. Assumption 3.5 hold and let $x_* = \Pi_{\mathcal{L}}^{\mathbf{B}}(x_0)$. Then for all $i = 1, 2, \ldots, n$,

$$\mathrm{E}\left[u_i^\top \mathbf{B}^{1/2}(x_k - x_*)\right] = \begin{cases} 0 & \text{if} \quad \lambda_i = 0, \\ (1 - \omega\lambda_i)^k u_i^\top \mathbf{B}^{1/2}(x_0 - x_*) & \text{if} \quad \lambda_i > 0. \end{cases} \tag{52}$$

Moreover,

$$\|\mathrm{E}\left[x_k - x_*\right]\|_{\mathbf{B}}^2 \leq \rho^k(\omega)\|x_0 - x_*\|_{\mathbf{B}}^2, \tag{53}$$

where the rate is given by

$$\rho(\omega) \overset{\text{def}}{=} \max_{i:\lambda_i > 0} (1 - \omega\lambda_i)^2. \tag{54}$$

Note that all eigenvalues of $\mathbf{W}$ play a role, governing the convergence speeds of individual elements of the transformed error vector. Under exactness, and relative to the particular solution $x_* = \Pi_{\mathcal{L}}^{\mathbf{B}}(x_0)$, the expected error $\mathrm{E}\left[x_k - x_*\right]$ converges to zero at a linear rate. The proof of the theorem is provided in Section 4.3.

Remark 3. Having established (48), perhaps the the most obvious way of analyzing the method is by taking $\mathbf{B}$-norms on both sides of identity (48). This way we obtain the estimate $\|\mathrm{E}\left[x_{k+1} - x_*\right]\|_{\mathbf{B}}^2 \leq \tilde{\rho}(\omega)\|\mathrm{E}\left[x_k - x_*\right]\|_{\mathbf{B}}^2$, where $\tilde{\rho}(\omega) = \|\mathbf{I} - \omega\mathbf{B}^{-1}\mathrm{E}\left[\mathbf{Z}\right]\|_{\mathbf{B}}^2 = \sigma_{\max}^2(\mathbf{I} - \omega\mathbf{B}^{-1/2}\mathrm{E}\left[\mathbf{Z}\right]\mathbf{B}^{-1/2})$, $\|\mathbf{M}\|_{\mathbf{B}} \overset{\text{def}}{=} \max\{\|\mathbf{M}x\|_{\mathbf{B}} \ : \ \|x\|_{\mathbf{B}} \leq 1\}$, and $\sigma_{\max}(\cdot)$ denotes the largest singular value. This gives the inequality

$$\|\mathrm{E}\left[x_k - x_*\right]\|_{\mathbf{B}}^2 \leq \tilde{\rho}^k(\omega)\|x_0 - x_*\|_{\mathbf{B}}^2, \tag{55}$$

which can be directly compared with (51). We now highlight two differences between these two bounds. The first approach gives a more detailed, information, as the identity in (51) is an exact formula for the norm of the expected error. Moreover, while in view of (54), we have $\rho(\omega) = \max_{i:\lambda_i>0}(1 - \omega\lambda_i)^2$, it can be shown that $\tilde{\rho}(\omega) = \max_i(1 - \omega\lambda_i)^2$. The two bounds are identical if $\lambda_{\min} > 0$, but they differ otherwise. In particular, as long as $\lambda_{\min} = 0$, we have $\tilde{\rho}(\omega) \geq 1$ for all ω, which means that the bound (55) does not guarantee convergence.

The following result, characterizing convergence of the expected errors to zero, is a straightforward corollary of Theorem 4.3.

Corollary 4.4 (Necessary and sufficient conditions for convergence)**.** Let Assumption 3.5 hold. Choose any $x_0 \in \mathbb{R}^n$ and let $x_* = \Pi_{\mathcal{L}}^{\mathbf{B}}(x_0)$. If $\{x_k\}$ are the random iterates produced by Algorithm 1, then the following statements are equivalent:

(i) $|1 - \omega\lambda_i| < 1$ for all i for which $\lambda_i > 0$

(ii) $0 < \omega < 2/\lambda_{\max}$

(iii) $\mathrm{E}\left[u_i^\top \mathbf{B}^{1/2}(x_k - x_*)\right] \to 0$ for all i

(iv) $\|\mathrm{E}\left[x_k - x_*\right]\|_{\mathbf{B}}^2 \to 0$

4.3 Proof of Theorem 4.3

We first start with a lemma.

Lemma 4.5. Let Assumption 3.5 hold. Consider arbitrary $x \in \mathbb{R}^n$ and let $x_* = \Pi_{\mathcal{L}}^{\mathbf{B}}(x)$. If $\lambda_i = 0$, then $u_i^\top \mathbf{B}^{1/2}(x - x_*) = 0$.

Proof. From (21) we see that $x - x_* = \mathbf{B}^{-1}\mathbf{A}^\top w$ for some $w \in \mathbb{R}^m$. Therefore, $u_i^\top \mathbf{B}^{1/2}(x - x_*) = u_i^\top \mathbf{B}^{-1/2}\mathbf{A}^\top w$. By Theorem 3.6, we have $\mathrm{Range}\left(u_i \;:\; \lambda_i = 0\right) = \mathrm{Null}\left(\mathbf{A}\mathbf{B}^{-1/2}\right)$, from which it follows that $u_i^\top \mathbf{B}^{-1/2}\mathbf{A} = 0$. □

We now proceed with the proof of Theorem 4.3. The iteration of Algorithm 1 can be written in the form

$$e_{k+1} = (\mathbf{I} - \omega\mathbf{B}^{-1}\mathbf{Z}_k)e_k, \tag{56}$$

where $e_k = x_k - x_*$. Multiplying both sides of this equation by $\mathbf{B}^{1/2}$ from the left, and taking expectation conditional on e_k, we obtain $\mathrm{E}\left[\mathbf{B}^{1/2}e_{k+1} \mid e_k\right] = (\mathbf{I} - \omega\mathbf{B}^{-1/2}\mathrm{E}\left[\mathbf{Z}\right]\mathbf{B}^{-1/2})\mathbf{B}^{1/2}e_k$. Taking expectations on both sides and using the tower property, we get

$$\mathrm{E}\left[\mathbf{B}^{1/2}e_{k+1}\right] = \mathrm{E}\left[\mathrm{E}\left[\mathbf{B}^{1/2}e_{k+1} \mid e_k\right]\right] = (\mathbf{I} - \omega\mathbf{B}^{-1/2}\mathrm{E}\left[\mathbf{Z}\right]\mathbf{B}^{-1/2})\mathrm{E}\left[\mathbf{B}^{1/2}e_k\right].$$

We now replace $\mathbf{B}^{-1/2}\mathrm{E}\left[\mathbf{Z}\right]\mathbf{B}^{-1/2}$ by its eigenvalue decomposition $\mathbf{U}\Lambda\mathbf{U}^\top$ (see (40)), multiply both sides of the last inequality by $\mathbf{U}^\top$ from the left, and use linearity of expectation to obtain

$$\mathrm{E}\left[\mathbf{U}^\top\mathbf{B}^{1/2}e_{k+1}\right] = (\mathbf{I} - \omega\Lambda)\mathrm{E}\left[\mathbf{U}^\top\mathbf{B}^{1/2}e_k\right].$$

Unrolling the recurrence, we get (49). When this is written coordinate-by-coordinate, (50) follows. Identity (51) follows immediately by equating standard Euclidean norms of both sides of (49). If $x_* = \Pi_{\mathcal{L}}^{\mathbf{B}}(x_0)$, then from Lemma 4.5 we see that $\lambda_i = 0$ implies $u_i^\top \mathbf{B}^{1/2}(x_0 - x_*) = 0$. Using this in

(50) gives (52). Finally, inequality (53) follows from

$$\begin{aligned}
\|\mathrm{E}\left[x_k - x_*\right]\|_{\mathbf{B}}^2 &\overset{(51)}{=} \sum_{i=1}^{n}(1-\omega\lambda_i)^{2k}\left(u_i^\top \mathbf{B}^{1/2}(x_0-x_*)\right)^2 \\
&= \sum_{i:\lambda_i>0}(1-\omega\lambda_i)^{2k}\left(u_i^\top \mathbf{B}^{1/2}(x_0-x_*)\right)^2 \\
&\overset{(54)}{\leq} \rho^k(\omega)\sum_{i:\lambda_i>0}\left(u_i^\top \mathbf{B}^{1/2}(x_0-x_*)\right)^2 \\
&= \rho^k(\omega)\sum_{i:\lambda_i>0}\left(u_i^\top \mathbf{B}^{1/2}(x_0-x_*)\right)^2 + \rho^k(\omega)\sum_{i:\lambda_i=0}\left(u_i^\top \mathbf{B}^{1/2}(x_0-x_*)\right)^2 \\
&= \rho^k(\omega)\sum_{i}\left(u_i^\top \mathbf{B}^{1/2}(x_0-x_*)\right)^2 \\
&= \rho^k(\omega)\sum_{i}(x_0-x_*)^\top \mathbf{B}^{1/2}u_iu_i^\top \mathbf{B}^{1/2}(x_0-x_*) \\
&= \rho^k(\omega)\sum_{i}(x_0-x_*)^\top \mathbf{B}^{1/2}\left(\sum_i u_iu_i^\top\right)\mathbf{B}^{1/2}(x_0-x_*) \\
&= \rho^k(\omega)\|x_0-x_*\|_{\mathbf{B}}^2.
\end{aligned}$$

The last identity follows from the fact that $\sum_i u_iu_i^\top = \mathbf{U}\mathbf{U}^\top = \mathbf{I}$.

4.4 Choice of the stepsize / relaxation parameter

We now consider the problem of choosing the stepsize (relaxation) parameter ω. In view of (53) and (54), the optimal relaxation parameter is the one solving the following optimization problem:

$$\min_{\omega\in\mathbb{R}}\left\{\rho(\omega) = \max_{i:\lambda_i>0}(1-\omega\lambda_i)^2\right\}. \tag{57}$$

In the next result we solve the above problem.

Theorem 4.6 (Stepsize)**.** The objective of (57) is given by

$$\rho(\omega) = \begin{cases}(1-\omega\lambda_{\max})^2 & \text{if} \quad \omega \leq 0 \\ (1-\omega\lambda_{\min}^+)^2 & \text{if} \quad 0\leq\omega\leq\omega^*\,, \\ (1-\omega\lambda_{\max})^2 & \text{if} \quad \omega\geq\omega^*\end{cases} \tag{58}$$

where $\omega^* \overset{\text{def}}{=} 2/(\lambda_{\min}^+ + \lambda_{\max})$. Moreover, ρ is decreasing on $(-\infty,\omega^*]$ and increasing on $[\omega^*,+\infty)$, and hence the optimal solution of (57) is ω^*. Further, we have:

(i) If we choose $\omega = 1$ (no over-relaxation), then

$$\rho(1) = (1-\lambda_{\min}^+)^2. \tag{59}$$

(ii) If we choose $\omega = 1/\lambda_{\max}$ (over-relaxation), then

$$\rho(1/\lambda_{\max}) = \left(1 - \frac{\lambda_{\min}^+}{\lambda_{\max}}\right)^2 \overset{(41)}{=} \left(1 - \frac{1}{\zeta}\right)^2. \tag{60}$$

(iii) If we choose $\omega = \omega^*$ (optimal over-relaxation), then the optimal rate is

$$\rho(\omega^*) = \left(1 - \frac{2\lambda_{\min}^+}{\lambda_{\min}^+ + \lambda_{\max}}\right)^2 = \left(1 - \frac{2}{\zeta + 1}\right)^2. \tag{61}$$

Proof. Recall that $\lambda_{\max} \leq 1$. Letting $\rho_i(\omega) = (1-\omega\lambda_i)^2$, it is easy to see that $\rho(\omega) = \max\{\rho_j(\omega), \rho_n(\omega)\}$, where j is such that $\lambda_j = \lambda_{\min}^+$. Note that $\rho_j(\omega) = \rho_n(\omega)$ for $\omega \in \{0, \omega^*\}$. From this we deduce that $\rho_j \geq \rho_n$ on $(-\infty, 0]$, $\rho_j \leq \rho_n$ on $[0, \omega^*]$, and $\rho_j \geq \rho_n$ on $[\omega^*, +\infty)$, obtaining (58). We see that ρ is decreasing on $(-\infty, \omega^*]$, and increasing on $[\omega^*, +\infty)$. The remaining results follow directly by plugging specific values of ω into (58). □

Theorem 4.6 can be intuitively understood in the following way. By design, we know that $\lambda_{\max} \leq 1$. If we do not have a better bound on the largest eigenvalue, we can simply choose $\omega = 1$ to ensure convergence. If we have a stronger bound available, say $\lambda_{\max} \leq U < 1$, we can pick $\omega = 1/U$, and the convergence rate will improve. The better the bound, the better the rate. However, using a stepsize of the form $\omega = 1/U$ where U is not an upper bound on $\lambda_{\max}$ is risky: if we underestimate the eigenvalue by a factor of 2 or more, we can not guarantee convergence. Indeed, if $U \leq \lambda_{\max}/2$, then $1/U \geq 2/\lambda_{\max}$ and hence $\rho(\omega) \geq 1$. Beyond this point, information about $\lambda_{\min}^+$ is useful. However, the best possible improvement beyond this only leads to a further factor of 2 speedup in terms of the number of iterations. Therefore, one needs to be careful about underestimating $\lambda_{\max}$.

Example 2 (Random vectors). An important class of methods is obtained by restricting $\mathbf{S}$ to random vectors. In this case,

$$\begin{aligned} \lambda_{\min}^+ + \lambda_{\max} &\leq \sum_{i=1}^n \lambda_i = \mathrm{Trace}\left(\mathbf{B}^{-1/2}\mathrm{E}\left[\mathbf{Z}\right]\mathbf{B}^{-1/2}\right) = \mathrm{E}\left[\mathrm{Trace}\left(\mathbf{B}^{-1/2}\mathbf{Z}\mathbf{B}^{-1/2}\right)\right] \\ &= \mathrm{E}\left[\mathrm{Trace}\left(\mathbf{B}^{-1}\mathbf{Z}\right)\right] \quad = \quad \mathrm{E}\left[\mathsf{dim}(\mathrm{Range}\left(\mathbf{B}^{-1}\mathbf{Z}\right))\right] = 1, \end{aligned}$$

and thus $\omega^* = 2/(\lambda_{\min}^+ + \lambda_{\max}) \geq 2$. This means that in this case we can always safely choose the relaxation parameter to be $\omega = 2$. This results in faster rate than the choice $\omega = 1$.

4.5 L2 convergence

In this section we establish a bound on $\mathrm{E}\left[\|x_k - x_*\|_{\mathbf{B}}^2\right]$, i.e., we prove $L2$ convergence. This is a stronger type of convergence than what we get by bounding $\|\mathrm{E}\left[x_k - x_*\right]\|_{\mathbf{B}}^2$. Indeed, for any random vector x_k we have the inequality (see Lemma 4.1 in [17])

$$\mathrm{E}\left[\|x_k - x_*\|_{\mathbf{B}}^2\right] = \|\mathrm{E}\left[x_k - x_*\right]\|_{\mathbf{B}}^2 + \mathrm{E}\left[\|x_k - \mathrm{E}\left[x_k\right]\|_{\mathbf{B}}^2\right].$$

Hence, L2 convergence also implies that the quantity $\mathrm{E}\left[\|x_k - \mathrm{E}\left[x_k\right]\|_{\mathbf{B}}^2\right]$—the *total variance*[2] of x_k—converges to zero.

[2] Total variance of a random vector is the trace of its covariance matrix.

We shall first establish an insightful lemma. The lemma connects two important measures of success: $\|x_k - x_*\|_{\mathbf{B}}^2$ and $f(x_k)$.

Lemma 4.7. Choose $x_0 \in \mathbb{R}^n$ and let $\{x_k\}_{k=0}^{\infty}$ be the random iterates produced by Algorithm 1, with an arbitrary relaxation parameter $\omega \in \mathbb{R}$. Let $x_* \in \mathcal{L}$. Then we have the identities $\|x_{k+1} - x_k\|_{\mathbf{B}}^2 = 2\omega^2 f_{\mathbf{S}_k}(x_k)$, and

$$\|x_{k+1} - x_*\|_{\mathbf{B}}^2 = \|x_k - x_*\|_{\mathbf{B}}^2 - 2\omega(2-\omega) f_{\mathbf{S}_k}(x_k). \tag{62}$$

Moreover, $\mathrm{E}\left[\|x_{k+1} - x_k\|_{\mathbf{B}}^2\right] = 2\omega^2 \mathrm{E}\left[f(x_k)\right]$, and

$$\mathrm{E}\left[\|x_{k+1} - x_*\|_{\mathbf{B}}^2\right] = \mathrm{E}\left[\|x_k - x_*\|_{\mathbf{B}}^2\right] - 2\omega(2-\omega)\mathrm{E}\left[f(x_k)\right]. \tag{63}$$

Proof. Recall that Algorithm 1 performs the update $x_{k+1} = x_k - \omega \mathbf{B}^{-1}\mathbf{Z}_k(x_k - x_*)$. From this we get

$$\|x_{k+1} - x_k\|_{\mathbf{B}}^2 = \omega^2 \|\mathbf{B}^{-1}\mathbf{Z}_k(x_k - x_*)\|_{\mathbf{B}}^2 \overset{(24)}{=} \omega^2 (x_k - x_*)^\top \mathbf{Z}_k (x_k - x_*) \overset{(26)}{=} 2\omega^2 f_{\mathbf{S}_k}(x_k), \tag{64}$$

In a similar vein,

$$\begin{aligned}
\|x_{k+1} - x_*\|_{\mathbf{B}}^2 &= \|(\mathbf{I} - \omega \mathbf{B}^{-1}\mathbf{Z}_k)(x_k - x_*)\|_{\mathbf{B}}^2 \\
&= (x_k - x_*)^\top (\mathbf{I} - \omega \mathbf{Z}_k \mathbf{B}^{-1}) \mathbf{B} (\mathbf{I} - \omega \mathbf{B}^{-1}\mathbf{Z}_k)(x_k - x_*) \\
&\overset{(24)}{=} (x_k - x_*)^\top (\mathbf{B} - \omega(2-\omega)\mathbf{Z}_k)(x_k - x_*) \\
&\overset{(26)}{=} \|x_k - x_*\|_{\mathbf{B}}^2 - 2\omega(2-\omega) f_{\mathbf{S}_k}(x_k),
\end{aligned}$$

establishing (62). Taking expectation in (64), we get

$$\mathrm{E}\left[\|x_{k+1} - x_k\|_{\mathbf{B}}^2\right] = \mathrm{E}\left[\mathrm{E}\left[\|x_{k+1} - x_k\|_{\mathbf{B}}^2 \mid x_k\right]\right] = 2\omega^2 \mathrm{E}\left[\mathrm{E}\left[f_{\mathbf{S}_k}(x_k) \mid x_k\right]\right] = 2\omega^2 \mathrm{E}\left[f(x_k)\right].$$

Taking expectation in (62), we get $\mathrm{E}\left[\|x_{k+1} - x_*\|_{\mathbf{B}}^2 \,|\, x_k\right] = \|x_k - x_*\|_{\mathbf{B}}^2 - 2\omega(2-\omega) f(x_k)$. It remains to take expectation again. □

In our next result we utilize Lemma 4.7 to establish L2 convergence of the basic method.

Theorem 4.8 ($L2$ convergence)**.** Let Assumption 3.5 hold and set $x_* = \Pi_{\mathcal{L}}^{\mathbf{B}}(x_0)$. Let $\{x_k\}$ be the random iterates produced by Algorithm 1, where the relaxation parameter satisfies $0 < \omega < 2$.

(i) For $k \geq 0$ we have

$$(1 - \omega(2-\omega)\lambda_{\max})^k \|x_0 - x_*\|_{\mathbf{B}}^2 \leq \mathrm{E}\left[\|x_k - x_*\|_{\mathbf{B}}^2\right] \leq (1 - \omega(2-\omega)\lambda_{\min}^+)^k \|x_0 - x_*\|_{\mathbf{B}}^2. \tag{65}$$

(ii) The average iterate $\hat{x}_k \overset{\text{def}}{=} \frac{1}{k}\sum_{t=0}^{k-1} x_t$ for all $k \geq 1$ satisfies

$$\mathrm{E}\left[\|\hat{x}_k - x_*\|_{\mathbf{B}}^2\right] \leq \frac{\|x_0 - x_*\|_{\mathbf{B}}^2}{2\omega(2-\omega)\lambda_{\min}^+ k}. \tag{66}$$

The best rate is achieved when $\omega = 1$.

Proof. Let $\phi_k = \mathrm{E}\left[f(x_k)\right]$ and $r_k = \mathrm{E}\left[\|x_k - x_*\|_{\mathbf{B}}^2\right]$.

(i) We have $r_{k+1} \overset{(63)}{=} r_k - 2\omega(2-\omega)\phi_k \overset{(44)}{\leq} r_k - \omega(2-\omega)\lambda_{\min}^+ r_k$, and $r_{k+1} \overset{(63)}{=} r_k - 2\omega(2-\omega)\phi_k \overset{(43)}{\geq} r_k - \omega(2-\omega)\lambda_{\max} r_k$. Inequalities (65) follow from this by unrolling the recurrences.

(ii) By summing up the identities from (63), we get $2\omega(2-\omega)\sum_{t=0}^{k-1}\phi_t = r_0 - r_k$. Therefore,

$$\begin{aligned} \mathrm{E}\left[\|\hat{x}_k - x_*\|_{\mathbf{B}}^2\right] &= \mathrm{E}\left[\left\|\frac{1}{k}\sum_{t=0}^{k-1}(x_t - x_*)\right\|_{\mathbf{B}}^2\right] \leq \mathrm{E}\left[\frac{1}{k}\sum_{t=0}^{k-1}\|x_t - x_*\|_{\mathbf{B}}^2\right] \\ &= \frac{1}{k}\sum_{t=0}^{k-1} r_t \overset{(44)}{\leq} \frac{1}{\lambda_{\min}^+ k}\sum_{t=0}^{k-1}\phi_t \leq \frac{r_0}{2\omega(2-\omega)\lambda_{\min}^+ k}. \end{aligned}$$

□

Not that in part (i) we give both an upper and a *lower* bound on $\mathrm{E}\left[\|x_k - x_*\|_{\mathbf{B}}^2\right]$.

4.6 Convergence of expected function values

In this section we establish a linear convergence rate for the decay of $\mathrm{E}[f(x_k)]$ to zero. We prove two results, with different quantitative (speed) and qualitative (assumptions and insights gained) qualities.

The complexity in the first result (Theorem 4.9) is disappointing: it is (slightly) worse than quadratic in the condition number ζ. However, we do not need to invoke Assumption 3.5 (exactness). In addition, this result implies that the expected function values decay *monotonically* to zero.

Theorem 4.9 (Convergence of expected function values)**.** Choose any $x_0 \in \mathbb{R}^n$ and let $\{x_k\}$ be the random iterates produced by Algorithm 1, where $0 \leq \omega \leq 2/\zeta$ (note that $2/\zeta = 2\lambda_{\min}^+/\lambda_{\max} \leq 2$). Then

$$\mathrm{E}[f(x_k)] \leq (1 - 2\lambda_{\min}^+\omega + \lambda_{\max}\omega^2)^k f(x_0). \tag{67}$$

The optimal rate is achieved for $\omega = 1/\zeta$, in which case we get the bound

$$\mathrm{E}[f(x_k)] \leq \left(1 - \frac{(\lambda_{\min}^+)^2}{\lambda_{\max}}\right)^k f(x_0).$$

Proof. Let $\mathbf{S} \sim \mathcal{D}$ be independent from $\mathbf{S}_0, \mathbf{S}_1, \ldots, \mathbf{S}_k$ and fix any $x_* \in \mathcal{L}$. Then we have

$$\begin{aligned} f(x_{k+1}) &\overset{(32)}{=} \mathrm{E}_{\mathbf{S}\sim\mathcal{D}}\left[f_{\mathbf{S}}(x_{k+1})\right] \\ &\overset{(7)}{=} \mathrm{E}_{\mathbf{S}\sim\mathcal{D}}\left[f_{\mathbf{S}}(x_k - \omega\nabla f_{\mathbf{S}_k}(x_k))\right] \\ &\overset{(26)}{=} \frac{1}{2}\mathrm{E}_{\mathbf{S}\sim\mathcal{D}}\left[(x_k - x_* - \omega\nabla f_{\mathbf{S}_k}(x_k))^\top \mathbf{Z}(x_k - x_* - \omega\nabla f_{\mathbf{S}_k}(x_k))\right] \\ &= \frac{1}{2}(x_k - x_* - \omega\nabla f_{\mathbf{S}_k}(x_k))^\top \mathrm{E}[\mathbf{Z}](x_k - x_* - \omega\nabla f_{\mathbf{S}_k}(x_k)) \\ &= \frac{1}{2}(x_k - x_*)^\top \mathrm{E}[\mathbf{Z}](x_k - x_*) - \omega(\nabla f_{\mathbf{S}_k}(x_k))^\top \mathrm{E}[\mathbf{Z}](x_k - x_*) + \frac{\omega^2}{2}\|\nabla f_{\mathbf{S}_k}(x_k)\|_{\mathrm{E}[\mathbf{Z}]}^2 \\ &\overset{(34)}{=} f(x_k) - \omega(\nabla f_{\mathbf{S}_k}(x_k))^\top \mathrm{E}[\mathbf{Z}](x_k - x_*) + \frac{\omega^2}{2}\|\nabla f_{\mathbf{S}_k}(x_k)\|_{\mathrm{E}[\mathbf{Z}]}^2. \end{aligned}$$

Taking expectations, conditioned on x_k (that is, the expectation is with respect to $\mathbf{S}_k$), we can further write

$$\mathrm{E}\left[f(x_{k+1}) \mid x_k\right] = f(x_k) - \omega\alpha_k + \omega^2\beta_k, \tag{68}$$

where

$$\alpha_k \overset{\text{def}}{=} \left(\mathrm{E}_{\mathbf{S}_k\sim\mathcal{D}}\left[\nabla f_{\mathbf{S}_k}(x_k)\right]\right)^\top \mathrm{E}\left[\mathbf{Z}\right](x_k - x_*), \qquad \beta_k \overset{\text{def}}{=} \frac{1}{2}\mathrm{E}_{\mathbf{S}_k\sim\mathcal{D}}\left[\|\nabla f_{\mathbf{S}_k}(x_k)\|^2_{\mathrm{E}[\mathbf{Z}]}\right]. \tag{69}$$

We shall now bound α_k from below and β_k from above in terms of $f(x_k)$. Using the inequality $\mathrm{E}\left[\mathbf{Z}\right] \preceq \lambda_{\max}\mathbf{B}$ (this follows from $\mathbf{B}^{-1/2}\mathrm{E}\left[\mathbf{Z}\right]\mathbf{B}^{-1/2} \preceq \lambda_{\max}\mathbf{I}$), we get

$$\beta_k \overset{(69)}{\leq} \frac{\lambda_{\max}}{2}\mathrm{E}_{\mathbf{S}_k\sim\mathcal{D}}\left[\|\nabla f_{\mathbf{S}_k}(x_k)\|^2_{\mathbf{B}}\right] \overset{(30)}{=} \lambda_{\max}\mathrm{E}_{\mathbf{S}_k\sim\mathcal{D}}\left[f_{\mathbf{S}_k}(x_k)\right] \overset{(32)}{=} \lambda_{\max}f(x_k).$$

On the other hand,

$$\alpha_k \overset{(69)+(35)}{=} (x_k - x_*)^\top \mathrm{E}\left[\mathbf{Z}\right]\mathbf{B}^{-1}\mathrm{E}\left[\mathbf{Z}\right](x_k - x_*) \overset{(35)}{=} \|\nabla f(x_k)\|^2_{\mathbf{B}} \overset{(42)}{\geq} 2\lambda^+_{\min}f(x_k).$$

Substituting the bounds for α_k and β_k into (68), we get $\mathrm{E}\left[f(x_{k+1}) \mid x_k\right] \leq (1 - 2\lambda^+_{\min}\omega + \lambda_{\max}\omega^2)f(x_k)$. Taking expectations again gives

$$\mathrm{E}\left[f(x_{k+1})\right] = \mathrm{E}\left[\mathrm{E}\left[f(x_{k+1}) \mid x_k\right]\right] \leq (1 - 2\lambda^+_{\min}\omega + \lambda_{\max}\omega^2)\mathrm{E}\left[f(x_k)\right].$$

It remains to unroll the recurrence. □

We now present an alternative convergence result (Theorem 4.10), one in which we do not bound the decrease in terms of the initial function value, $f(x_0)$, but in terms of a somewhat larger quantity. This allows us to provide a better convergence rate. For this result to hold, however, we need to invoke Assumption 3.5. Note also that this result does not imply that the expected function values decay monotonically.

Theorem 4.10 (Convergence of expected function values)**.** Choose $x_0 \in \mathbb{R}^n$, and let $\{x_k\}_{k=0}^\infty$ be the random iterates produced by Algorithm 1, where the relaxation parameter satisfies $0 < \omega < 2$.

(i) Let $x_* \in \mathcal{L}$. The average iterate $\hat{x}_k \overset{\text{def}}{=} \frac{1}{k}\sum_{t=0}^{k-1} x_t$ for all $k \geq 1$ satisfies

$$\mathrm{E}\left[f(\hat{x}_k)\right] \leq \frac{\|x_0 - x_*\|^2_{\mathbf{B}}}{2\omega(2-\omega)k}. \tag{70}$$

(ii) Now let Assumption 3.5 hold. For $x_* = \Pi^{\mathbf{B}}_{\mathcal{L}}(x_0)$ and $k \geq 0$ we have

$$\mathrm{E}\left[f(x_k)\right] \leq \left(1 - \omega(2-\omega)\lambda^+_{\min}\right)^k \frac{\lambda_{\max}\|x_0 - x_*\|^2_{\mathbf{B}}}{2}. \tag{71}$$

The best rate is achieved when $\omega = 1$.

Proof. (i) Let $\phi_k = \mathrm{E}\left[f(x_k)\right]$ and $r_k = \mathrm{E}\left[\|x_k - x_*\|^2_{\mathbf{B}}\right]$. By summing up the identities from (63), we get $2\omega(2-\omega)\sum_{t=0}^{k-1}\phi_t = r_0 - r_k$. Therefore, using Jensen's inequality,

$$\mathrm{E}\left[f(\hat{x}_k)\right] \;\leq\; \mathrm{E}\left[\frac{1}{k}\sum_{t=0}^{k-1} f(x_t)\right] = \frac{1}{k}\sum_{t=0}^{k-1}\phi_t = \frac{r_0 - r_k}{2\omega(2-\omega)k} \leq \frac{r_0}{2\omega(2-\omega)k}.$$

(ii) Combining inequality (43) with Theorem 4.8, we get

$$\mathrm{E}\left[f(x_k)\right] \leq \frac{\lambda_{\max}}{2}\mathrm{E}\left[\|x_k - x_*\|_{\mathbf{B}}^2\right] \overset{(65)}{\leq} \left(1-\omega(2-\omega)\lambda_{\min}^{+}\right)^k \frac{\lambda_{\max}\|x_0 - x_*\|_{\mathbf{B}}^2}{2}.$$

□

Remark 4. Theorems 4.9 and 4.10 are complementary. In particular, the complexity result given in Theorem 4.9 (for the last iterate) holds under weaker assumptions. Moreover, Theorem 4.9 implies monotonicity of expected function values. On the other hand, the rate is substantially better in Theorem 4.10. Also, Theorem 4.10 applies to a wider range of stepsizes.

It is also possible to obtain other convergence results as a corollary. For instance, one can get a linear rate for the decay of the norms of the gradients as a corollary of Theorems 4.9 and 4.10 using the upper bound in Lemma 4.2.

5 Parallel and Accelerated Methods

In this section we propose and analyze *parallel* and *accelerated* variants of Algorithm 1.

5.1 Parallel method

The parallel method (12) is formalized in this section as Algorithm 2.

Algorithm 2 Parallel Method

Parameters: distribution $\mathcal{D}$ from which to sample matrices; positive definite matrix $\mathbf{B} \in \mathbb{R}^{n\times n}$; stepsize/relaxation parameter $\omega \in \mathbb{R}$; parallelism parameter τ
Choose $x_0 \in \mathbb{R}^n$ ▷ Initialization
for $k = 0, 1, 2, \ldots$ **do**
  for $i = 1, 2, \ldots, \tau$ **do**
    Draw $\mathbf{S}_{ki} \sim \mathcal{D}$
    Set $z_{k+1,i} = x_k - \omega\mathbf{B}^{-1}\mathbf{A}^\top\mathbf{S}_{ki}(\mathbf{S}_{ki}^\top\mathbf{A}\mathbf{B}^{-1}\mathbf{A}^\top\mathbf{S}_{ki})^\dagger\mathbf{S}_{ki}^\top(\mathbf{A}x_k - b)$
  Set $x_{k+1} = \frac{1}{\tau}\sum_{i=1}^{\tau} z_{k+1,i}$ ▷ Average the results

For brevity, we only prove L2 convergence results. However, various other results can be obtained as well, as was the case for the basic method, such as convergence of expected iterates, expected function values and average iterates.

Theorem 5.1. Let Assumption 3.5 hold and set $x_* = \Pi_{\mathcal{L}}^{\mathbf{B}}(x_0)$. Let $\{x_k\}_{k=0}^{\infty}$ be the random iterates produced by Algorithm 2, where the relaxation parameter satisfies $0 < \omega < 2/\xi(\tau)$, and $\xi(\tau) \overset{\text{def}}{=} \frac{1}{\tau} + \left(1-\frac{1}{\tau}\right)\lambda_{\max}$. Then

$$\mathrm{E}\left[\|x_{k+1} - x_*\|_{\mathbf{B}}^2\right] \leq \rho(\omega,\tau)\cdot\mathrm{E}\left[\|x_k - x_*\|_{\mathbf{B}}^2\right],$$

and

$$\mathrm{E}\left[f(x_k)\right] \leq \rho(\omega,\tau)^k\frac{\lambda_{\max}}{2}\|x_0 - x_*\|_{\mathbf{B}}^2,$$

where $\rho(\omega,\tau) \overset{\text{def}}{=} 1-\omega\left[2-\omega\xi(\tau)\right]\lambda_{\min}^{+}$. For any fixed $\tau \geq 1$, the optimal stepsize choice is $\omega(\tau) \overset{\text{def}}{=} 1/\xi(\tau)$ and the associated optimal rate is

$$\rho(\omega(\tau),\tau) = 1 - \frac{\lambda_{\min}^{+}}{\frac{1}{\tau}+\left(1-\frac{1}{\tau}\right)\lambda_{\max}}. \tag{72}$$

Proof. Recall that Algorithm 2 performs the update $x_{k+1} = x_k - \omega \mathbf{B}^{-1}\tilde{\mathbf{Z}}_k(x_k - x_*)$, where $\tilde{\mathbf{Z}}_k \overset{\text{def}}{=} \frac{1}{\tau}\sum_{i=1}^{\tau}\mathbf{Z}_{ki}$. We have

$$\begin{aligned}
\mathrm{E}\left[\|x_{k+1}-x_*\|_{\mathbf{B}}^2 \mid x_k\right] &= \mathrm{E}\left[\|(\mathbf{I}-\omega\mathbf{B}^{-1}\tilde{\mathbf{Z}}_k)(x_k-x_*)\|_{\mathbf{B}}^2\right] \\
&= \mathrm{E}\left[(x_k-x_*)^\top(\mathbf{I}-\omega\tilde{\mathbf{Z}}_k\mathbf{B}^{-1})\mathbf{B}(\mathbf{I}-\omega\mathbf{B}^{-1}\tilde{\mathbf{Z}}_k)(x_k-x_*)\right] \\
&\overset{(24)}{=} \mathrm{E}\left[(x_k-x_*)^\top(\mathbf{B}-2\omega\tilde{\mathbf{Z}}_k+\omega^2\tilde{\mathbf{Z}}_k\mathbf{B}^{-1}\tilde{\mathbf{Z}}_k(x_k-x_*)\right] \\
&= (x_k-x_*)^\top\left(\mathbf{B}-2\omega\mathrm{E}\left[\tilde{\mathbf{Z}}_k\right]+\omega^2\mathrm{E}\left[\tilde{\mathbf{Z}}_k\mathbf{B}^{-1}\tilde{\mathbf{Z}}_k\right]\right)(x_k-x_*). \qquad (73)
\end{aligned}$$

Next, we can write

$$\tilde{\mathbf{Z}}_k\mathbf{B}^{-1}\tilde{\mathbf{Z}}_k = \frac{1}{\tau^2}\left(\sum_{i=1}^{\tau}\mathbf{Z}_{ki}\mathbf{B}^{-1}\mathbf{Z}_{ki} + \sum_{(i,j)\,:\,i\neq j}\mathbf{Z}_{ki}\mathbf{B}^{-1}\mathbf{Z}_{kj}\right).$$

Since $\mathbf{Z}_{ki}\mathbf{B}^{-1}\mathbf{Z}_{ki} = \mathbf{Z}_{ki}$, and because $\mathbf{Z}_{ki}$ and $\mathbf{Z}_{kj}$ are independent for $i \neq j$, we can further write

$$\mathrm{E}\left[\tilde{\mathbf{Z}}_k\mathbf{B}^{-1}\tilde{\mathbf{Z}}_k\right] = \frac{1}{\tau^2}\left(\tau\mathrm{E}\left[\mathbf{Z}\right]+(\tau^2-\tau)\mathrm{E}\left[\mathbf{Z}\right]\mathbf{B}^{-1}\mathrm{E}\left[\mathbf{Z}\right]\right) \preceq \left(\frac{1}{\tau}+\left(1-\frac{1}{\tau}\right)\lambda_{\max}\right)\mathrm{E}\left[\mathbf{Z}\right], \tag{74}$$

where we have used the estimate $\mathrm{E}\left[\mathbf{Z}\right]\mathbf{B}^{-1}\mathrm{E}\left[\mathbf{Z}\right] \preceq \lambda_{\max}\mathrm{E}\left[\mathbf{Z}\right]$, which follows from the bound $\mathbf{W}^2 \leq \lambda_{\max}\mathbf{W}$. Plugging (74) into (73), and noting that $\|x_k-x_*\|_{\mathrm{E}[\mathbf{Z}]}^2 = 2f(x_k)$, we obtain:

$$\begin{aligned}
\mathrm{E}\left[\|x_{k+1}-x_*\|_{\mathbf{B}}^2 \mid x_k\right] &\leq \|x_k-x_*\|_{\mathbf{B}}^2 - \left[2\omega-\omega^2\left(\frac{1}{\tau}+\left(1-\frac{1}{\tau}\right)\lambda_{\max}\right)\right]2f(x_k) \qquad (75) \\
&\overset{(44)}{\leq} \rho(\omega,\tau)\|x_k-x_*\|_{\mathbf{B}}^2. \qquad (76)
\end{aligned}$$

The inequality involving f is shown in the same way as in Theorem 4.10.

□

As $\tau \to \infty$, we have $\rho(\omega,\tau) \to 1-\omega(2-\omega\lambda_{\max})\lambda_{\min}^{+}$. The optimal stepsize is $\omega = 1/\lambda_{\max}$, which leads to the optimal rate

$$\rho(1/\lambda_{\max},+\infty) = 1 - \frac{\lambda_{\min}^{+}}{\lambda_{\max}} = 1 - \frac{1}{\zeta}.$$

In this asymptotic regime, we again recover linear dependence on the condition number.

5.2 Accelerated method

In this section we develop an accelerated variant of Algorithm 1. Recall that a single iteration of Algorithm 1 takes the form $x_{k+1} = \phi_\omega(x_k, \mathbf{S}_k)$, where

$$\phi_\omega(x, \mathbf{S}) \overset{\text{def}}{=} x - \omega \mathbf{B}^{-1}\mathbf{A}^\top \mathbf{S}(\mathbf{S}^\top \mathbf{A}\mathbf{B}^{-1}\mathbf{A}^\top \mathbf{S})^\dagger \mathbf{S}^\top(\mathbf{A}x - b). \tag{77}$$

We have seen that the convergence rate progressively improves as we increase ω from 1 to ω_*, which is the optimal choice. In particular, with $\omega = 1$ we have the complexity $\tilde{\mathcal{O}}(1/\lambda_{\min}^+)$, while choosing $\omega = 1/\lambda_{\max} = 1/\lambda_{\max}$ or $\omega = \omega_*$ leads to the improved complexity $\tilde{\mathcal{O}}(\lambda_{\max}/\lambda_{\min}^+) = \tilde{\mathcal{O}}(\zeta)$.

In order to obtain further acceleration, we suggest to perform an update step in which x_{k+1} depends on both x_k and x_{k-1}. In particular, we take two *dependent* steps of Algorithm 1, one from x_k and one from x_{k-1}, and take an affine combination of the results. This, the process is started with $x_0, x_1 \in \mathbb{R}^n$, and for $k \geq 1$ involves an iteration of the form

$$x_{k+1} = \gamma\phi_\omega(x_k, \mathbf{S}_k) + (1-\gamma)\phi_\omega(x_{k-1}, \mathbf{S}_{k-1}), \tag{78}$$

where the matrices $\{\mathbf{S}_k\}$ are independent samples from $\mathcal{D}$, and $\gamma \in \mathbb{R}$ is an *acceleration parameter*. Note that by choosing $\gamma = 1$ (no acceleration), we recover Algorithm 1. This method is formalized as Algorithm 3.

Algorithm 3 Accelerated Method

Parameters: distribution $\mathcal{D}$ from which to sample matrices; positive definite matrix $\mathbf{B} \in \mathbb{R}^{n\times n}$; stepsize/relaxation parameter $\omega > 0$; acceleration parameter $\gamma > 0$
Choose $x_0, x_1 \in \mathbb{R}^n$ such that $x_0 - x_1 \in \text{Range}\left(\mathbf{B}^{-1}\mathbf{A}^\top\right)$ (for instance, choose $x_0 = x_1$)
Draw $\mathbf{S}_0 \sim \mathcal{D}$
Set $z_0 = \phi_\omega(x_0, \mathbf{S}_0)$
for $k = 1, 2, \ldots$ **do**
    Draw a fresh sample $\mathbf{S}_k \sim \mathcal{D}$
    Set $z_k = \phi_\omega(x_k, \mathbf{S}_k)$
    Set $x_{k+1} = \gamma z_k + (1-\gamma)z_{k-1}$ ▷ Main update step
Output x_k

As we shall see, by a proper combination of overrelaxation (choice of ω) with acceleration (choice of γ), Algorithm 3 enjoys the accelerated complexity of $\tilde{\mathcal{O}}(\sqrt{\zeta})$.

We start with a lemma describing the evolution of the expected iterates.

Lemma 5.2 (Expected iterates)**.** Let x_* be any solution of $\mathbf{A}x = b$ and let $r_k \overset{\text{def}}{=} \mathrm{E}\left[x_k - x_*\right]$. Then for all k we have the recursion

$$r_{k+1} = \gamma(\mathbf{I} - \omega\mathbf{B}^{-1}\mathrm{E}\left[\mathbf{Z}\right])r_k + (1-\gamma)(\mathbf{I} - \omega\mathbf{B}^{-1}\mathrm{E}\left[\mathbf{Z}\right])r_{k-1}. \tag{79}$$

Proof. Taking expectation on both sides of (78), we get

$$\mathrm{E}\left[x_{k+1}\right] = \gamma\mathrm{E}\left[\phi_\omega(x_k, \mathbf{S}_k)\right] + (1-\gamma)\mathrm{E}\left[\phi_\omega(x_{k-1}, \mathbf{S}_{k-1})\right].$$

After subtracting x_* from both sides, using (77), and replacing b by $\mathbf{A}x_*$, we get

$$r_{k+1} \quad = \quad \gamma\mathrm{E}\left[(\mathbf{I} - \omega\mathbf{B}^{-1}\mathbf{Z}_k)(x_k - x_*)\right] + (1-\gamma)\mathrm{E}\left[(\mathbf{I} - \omega\mathbf{B}^{-1}\mathbf{Z}_{k-1})(x_{k-1} - x_*)\right],$$

where $\mathbf{Z}_k = \mathbf{A}^\top \mathbf{S}_k (\mathbf{S}_k^\top \mathbf{A}\mathbf{B}^{-1}\mathbf{A}^\top \mathbf{S}_k)^\dagger \mathbf{S}_k^\top \mathbf{A}$. We now use the tower property and linearity of expectation, we finally obtain:

$$\begin{aligned} r_{k+1} &= \gamma \mathrm{E}\left[\mathrm{E}\left[(\mathbf{I} - \omega \mathbf{B}^{-1}\mathbf{Z}_k)(x_k - x_*) \mid x_k\right]\right] + (1-\gamma)\mathrm{E}\left[\mathrm{E}\left[(\mathbf{I} - \omega \mathbf{B}^{-1}\mathbf{Z}_{k-1})(x_{k-1} - x_*) \mid x_{k-1}\right]\right] \\ &= \gamma \mathrm{E}\left[(\mathbf{I} - \omega \mathbf{B}^{-1}\mathrm{E}\left[\mathbf{Z}\right])(x_k - x_*)\right] + (1-\gamma)\mathrm{E}\left[(\mathbf{I} - \omega \mathbf{B}^{-1}\mathrm{E}\left[\mathbf{Z}\right])(x_{k-1} - x_*)\right] \\ &= \gamma(\mathbf{I} - \omega \mathbf{B}^{-1}\mathrm{E}\left[\mathbf{Z}\right])r_k + (1-\gamma)(\mathbf{I} - \omega \mathbf{B}^{-1}\mathrm{E}\left[\mathbf{Z}\right])r_{k-1}. \end{aligned}$$

□

We can now state our main complexity result. Note that the optimal choice of parameters, covered in case (i), leads to a rate which depends on the square root of the condition number.

Theorem 5.3 (Complexity of Algorithm 3)**.** Let Assumption 3.5 be satisfied and let $\{x_k\}_{k=0}^\infty$ be the sequence of random iterates produced by Algorithm 3, started with $x_0, x_1 \in \mathbb{R}^n$ satisfying the relation $x_0 - x_1 \in \mathrm{Range}\left(\mathbf{B}^{-1}\mathbf{A}^\top\right)$, with relaxation parameter $0 < \omega \leq 1/\lambda_{\max}$ and acceleration parameter $\gamma = 2/(1+\sqrt{\mu})$, where $\mu \in (0, \omega\lambda_{\min}^+)$. Let $x_* = \Pi_{\mathcal{L}}^{\mathbf{B}}(x_0)$. Then there exists a constant $C > 0$, such that for all $k \geq 2$ we have

$$\left\|\mathrm{E}\left[x_k - x_*\right]\right\|_{\mathbf{B}}^2 \leq (1-\sqrt{\mu})^{2k} C. \tag{80}$$

(i) If we choose $\omega = 1/\lambda_{\max}$ (overrelaxation), then we can pick $\mu = 0.99/\zeta$ (recall that $\zeta = \lambda_{\max}/\lambda_{\min}^+$ is the condition number), which leads to the rate

$$\left\|\mathrm{E}\left[x_k - x_*\right]\right\|_{\mathbf{B}}^2 \leq \left(1 - \sqrt{0.99\lambda_{\min}^+/\lambda_{\max}}\right)^{2k} C.$$

(ii) If we choose $\omega = 1$ (no overrelaxation), then we can pick $\mu = 0.99\lambda_{\min}^+$, which leads to the rate

$$\left\|\mathrm{E}\left[x_k - x_*\right]\right\|_{\mathbf{B}}^2 \leq \left(1 - \sqrt{0.99\lambda_{\min}^+}\right)^{2k} C.$$

Proof. Multiplying the identity in Lemma 5.2 from the left by $\mathbf{B}^{1/2}$, we obtain

$$\mathbf{B}^{1/2}r_{k+1} = \gamma\left(\mathbf{I} - \omega\mathbf{B}^{-1/2}\mathrm{E}\left[\mathbf{Z}\right]\mathbf{B}^{-1/2}\right)\mathbf{B}^{1/2}r_k + (1-\gamma)\left(\mathbf{I} - \omega\mathbf{B}^{-1/2}\mathrm{E}\left[\mathbf{Z}\right]\mathbf{B}^{-1/2}\right)\mathbf{B}^{1/2}r_{k-1}.$$

Plugging the eigenvalue decomposition $\mathbf{U}\Lambda\mathbf{U}^\top$ of $\mathbf{B}^{-1/2}\mathrm{E}\left[\mathbf{Z}\right]\mathbf{B}^{-1/2}$ into the above, and multiplying both sides from the left by $\mathbf{U}^\top$, we get

$$\mathbf{U}^\top\mathbf{B}^{1/2}r_{k+1} = \gamma\left(\mathbf{I} - \omega\Lambda\right)\mathbf{U}^\top\mathbf{B}^{1/2}r_k + (1-\gamma)(\mathbf{I} - \omega\Lambda)\mathbf{U}^\top\mathbf{B}^{1/2}r_{k-1}. \tag{81}$$

Now if we denote $w_k = \mathbf{U}^\top\mathbf{B}^{1/2}r_k \in \mathbb{R}^n$, (81) becomes separable in the coordinates of w:

$$w_{k+1} = \gamma(\mathbf{I} - \omega\Lambda)w_k + (1-\gamma)(\mathbf{I} - \omega\Lambda)w_{k-1}. \tag{82}$$

Writing this coordinate-by-coordinate (with w_k^i indicating the ith coordinate of w_k), we get

$$w_{k+1}^i = \gamma\left(1 - \omega\lambda_i\right)w_k^i + (1-\gamma)(1-\omega\lambda_i)w_{k-1}^i, \quad i = 1, 2, \ldots, n. \tag{83}$$

We now fix i and analyze recursion (83). We can use Lemma C.1 with $E = \gamma(1-\omega\lambda_i)$ and $F = (1-\gamma)(1-\omega\lambda_i)$. Now recall that $0 \leq \lambda_i \leq 1$ for all i, and $\lambda_{\min}^+ > 0$. Since we assume that $0 < \omega < 1/\lambda_{\max}$, we know that $0 < \omega\lambda_i \leq 1$ for all i for which $\lambda_i > 0$, and $\omega\lambda_i = 0$ for those i for which $\lambda_i = 0$. Therefore, it is enough to consider the following 3 cases:

(1) $\omega\lambda_i = 1$. In this case we see from (83) that $w_k^i = 0$ for all $k \geq 2$.

(2) $\omega\lambda_i = 0$. Since, by assumption, $x_0 - x_1 \in \mathrm{Range}\left(\mathbf{B}^{-1}\mathbf{A}^\top\right)$, it follows that $\Pi_{\mathcal{L}}^{\mathbf{B}}(x_0) = \Pi_{\mathcal{L}}^{\mathbf{B}}(x_1)$. All our arguments up to this point hold for arbitrary $x_* \in \mathcal{L}$. However, we now choose $x_* = \Pi_{\mathcal{L}}^{\mathbf{B}}(x_0) = \Pi_{\mathcal{L}}^{\mathbf{B}}(x_1)$. Invoking Lemma 4.5 twice, once for $x = x_0$ and then for $x = x_1$, we conclude that $w_0^i = u_i^\top \mathbf{B}^{1/2}(x_0 - x_*) = 0$ and $w_1^i = u_i^\top \mathbf{B}^{1/2}(x_1 - x_*) = 0$. In view of recursion (83), we conclude that $\omega_k^i = 0$ for all $k \geq 0$.

(3) $0 < \omega\lambda_i < 1$. In this case we have

$$E^2 + 4F = \gamma^2(1-\omega\lambda_i)^2 + 4(1-\gamma)(1-\omega\lambda_i) = (1-\omega\lambda_i)\left((2-\gamma)^2 - \omega\lambda_i\gamma^2\right)$$
$$= (1-\omega\lambda_i)\left(\left(\frac{2\sqrt{\mu}}{1+\sqrt{\mu}}\right)^2 - \omega\lambda_i\left(\frac{2}{1+\sqrt{\mu}}\right)^2\right) = 4\frac{(1-\omega\lambda_i)}{(1+\sqrt{\mu})^2}\left(\mu - \omega\lambda_i\right) < 0,$$

where the last inequality follows from the assumption $\mu < \omega\lambda_{\min}^+$. Therefore, we can apply Lemma C.1, using which we can deduce the bound

$$\begin{aligned} w_k^i &= 2M^k\left(C_0\cos(\theta k) + C_1\sin(\theta k)\right) \leq 2\left(\sqrt{\frac{E^2}{4} + \frac{-E^2-4F}{4}}\right)^k (|C_0| + |C_1|) \\ &= 2\left(\sqrt{-F}\right)^k (|C_0| + |C_1|) = 2\left(\sqrt{\frac{1-\sqrt{\mu}}{1+\sqrt{\mu}}(1-\omega\lambda_i)}\right)^k (|C_0| + |C_1|) \\ &\leq 2\left(\sqrt{\frac{1-\sqrt{\mu}}{1+\sqrt{\mu}}(1-\sqrt{\mu})(1+\sqrt{\mu})}\right)^k (|C_0| + |C_1|) = 2\left(1-\sqrt{\mu}\right)^k (|C_0| + |C_1|). \end{aligned}$$

As $|C_0| + |C_1|$ depends on i, we shall write $C^i = |C_0| + |C_1|$.

Putting everything together, for all $k \geq 2$ we have

$$\|r_k\|_{\mathbf{B}}^2 = \|\mathrm{E}\left[x_k - x_*\right]\|_{\mathbf{B}}^2 = \|\mathbf{U}^\top\mathbf{B}^{1/2}\mathrm{E}\left[x_k - x_*\right]\|^2 = \|w_k\|^2 = \sum_{i=1}^n (w_k^i)^2$$
$$\overset{(84)}{\leq} \sum_{i:\lambda_i=0} \underbrace{(w_0^i)^2}_{=0} + \sum_{i:\lambda_i>0} 4(1-\sqrt{\mu})^{2k} C^i = 4(1-\sqrt{\mu})^{2k} \sum_{i:\lambda_i>0} C^i,$$

finishing the proof.

□

Note that we do not have a result on L2 convergence. We have tried to obtain an accelerated rate in the L2 sense, but were not successful. We conjecture that such a result can be obtained.

6 Conclusion

We have developed a generic scheme for reformulating any linear system as a *stochastic problem*, which has several seemingly different but nevertheless equivalent interpretations: stochastic optimization problem, stochastic linear system, stochastic fixed point problem, and probabilistic intersection problem.

While stochastic optimization is a broadly studied field with rich history, the concepts of stochastic linear system, stochastic fixed point problem and probablistic intersection appear to be new.

We give sufficient, and necessary and sufficient conditions for the reformulation to be exact, i.e., for the solution set of the reformulation to exactly match the solution set of the linear system. To the bets of our knowledge, this is is first systematic study of stochastic reformulations of linear systems. Further, we have developed three algorithms—basic, parallel and accelerated methods—to solve the stochastic reformulations. We have studied the convergence of expected iterates, L2 convergence, converge of a Cesaro average of all iterates, and convergence of f. Our methods recover an array of existing randomized algorithms for solving linear systems in special cases, including several variants of the randomized Kaczmarz method [60], randomized coordinate descent [24], and all the methods developed in [17, 18].

Our work can be extended in several ways. One of the most promising of these is *stochastic preconditioning*, which refers to the generic problem of fine-tuning the formulations (by designing the distributions $\mathcal{D}$ and matrix $\mathbf{B}$) to the structure of $\mathbf{A}$. We conjecture that specific highly efficient methods can be designed in this way. Accelerated convergence in the L2 sense remains an important open problem.

Last but not least, we hope that this work provides bridge across several communities: numerical linear algebra, stochastic optimization, machine learning, computational geometry, fixed point theory, applied mathematics and probability theory. We hope that our work may inspire further progress at the boundaries of these fields.

References

[1] H. Avron, P. Maymounkov, and S. Toledo. Blendenpik: Supercharging LAPACK's least-squares solver. *SIAM Journal on Scientific Computing*, 32(3):1217–1236, 2010.

[2] HH Bauschke and JM Borwein. On projection algorithms for solving convex feasibility problems. *SIAM Review*, 38(3):367–426, 1996.

[3] J.P. Boyle and R.L. Dykstra. A method for finding projections onto the intersection of convex sets in Hilbert spaces. *Lecture Notes in Statistics*, 37:28–47, 1986.

[4] GC Calafiore, F Dabbene, and R Tempo. Randomized algorithms for probabilistic robustness with real and complex structured uncertainty. *IEEE Transactions on Automatic Control*, 45(12):2218–2235, 2000.

[5] GC Calafiore and BT Polyak. Stochastic algorithms for exact and approximate feasibility of robust LMIs. *IEEE Transactions on Automatic Control*, 46(11):1755–1759, 2001.

[6] 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)*, pages 693–703. Springer-Verlag London, 2002.

[7] Patrick Louis Combettes and Jean-Christophe Pesquet. Stochastic quasi-fejér block-coordinate fixed point iterations with random sweeping. *SIAM Journal on Optimization*, 25(2):1221–1248, 2015.

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

[9] Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. *arXiv:1407.0202*, 2014.

[10] Petros Drineas, Ravi Kannan, and Michael W. Mahoney. Fast monte carlo algorithms for matrices I: Approximating matrix multiplication. *SIAM Journal on Computing*, 36(1):132–157.

[11] Petros Drineas, Ravi Kannan, and Michael W. Mahoney. Fast monte carlo algorithms for matrices II: Computing a low-rank approximation to a matrix. *SIAM Journal on Computing*, 36(1):158–183.

[12] R.L. Dykstra. An algorithm for restricted least squares regression. *Journal of American Statistical Association*, 78:837–842, 1983.

[13] Saber Elaydi. *An Introduction to Difference Equations*. Undergraduate Texts in Mathematics. Springer, 2005.

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

[15] Jay P. Fillmore and Morris L. Marx. Linear recursive sequences. *SIAM Review*, 10(3):342–353, 1968.

[16] Robert Mansel 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*, pages 1869–1878, 2016.

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

[18] Robert Mansel Gower and Peter Richtárik. Stochastic dual ascent for solving linear systems. *arXiv preprint arXiv:1512.06890*, 2015.

[19] Robert Mansel Gower and Peter Richtárik. Linearly convergent randomized iterative methods for computing the pseudoinverse. *arXiv preprint arXiv:1612.06255*, 2016.

[20] Robert Mansel Gower and Peter Richtárik. Randomized quasi-Newton updates are linearly convergent matrix inversion algorithms. *arXiv preprint arXiv:1602.01768*, 2016.

[21] Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems 26*, pages 315–323. 2013.

[22] Jakub Konečný, Jie Lu, Peter Richtárik, and Martin Takáč. Mini-batch semi-stochastic gradient descent in the proximal setting. *IEEE Journal of Selected Topics in Signal Processing*, 10(2):242–255, 2016.

[23] Jakub Konečný and Peter Richtárik. S2GD: Semi-stochastic gradient descent methods. *Frontiers in Applied Mathematics and Statistics*, 2017.

[24] Dennis Leventhal and Adrian S. Lewis. Randomized methods for linear constraints: Convergence rates and conditioning. *Mathematics of Operations Research*, 35(3):641–654, 2010.

[25] Ji Liu and Stephen Wright. An accelerated randomized Kaczmarz algorithm. *Mathematics of Computation*, 2015.

[26] Nicolas Loizou and Peter Richtárik. A new perspective on randomized gossip algorithms. In *IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 440–444. IEEE, 2016.

[27] X. Meng, Michael A. Saunders, and Michael W. Mahoney. LSRN: a parallel iterative solver for strongly over- and under-determined systems. *SIAM Journal on Scientific Computing*, 36(2):95–118, 2014.

[28] Rajeev Motwani and Prabhakar Raghavan. *Randomized Algorithms*. Cambridge University Press, 1995.

[29] Mojmír Mutný and Peter Richtárik. Parallel stochastic Newton method. *arXiv preprint arXiv:1705.02005*, 2017.

[30] Angelia Nedić. Random algorithms for convex minimization problems. *Mathematical Programming*.

[31] Deana Needell. Randomized Kaczmarz solver for noisy linear systems. *BIT*, 50(2):395–403, 2010.

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

[33] Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

[34] Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course (Applied Optimization)*. Kluwer Academic Publishers, 2004.

[35] Yurii Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012.

[36] 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.

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

[38] Mert Pilanci and Martin J. Wainwright. A linear-time optimization algorithm with linear-quadratic convergence. *arXiv preprint arXiv:1505.02250*, 2015.

[39] Mert Pilanci and Martin J. Wainwright. Iterative Hessian sketch: Fast and accurate solution approximation for constrained least-squares. *Journal of Machine Learning Research*, 17(53):1–38, 2016.

[40] Zheng Qu and Richtárik. Coordinate descent with arbitrary sampling I: algorithms and complexity. *Optimization Methods and Software*, 31(5):829–857, 2016.

[41] Zheng Qu, Peter Richtárik, Martin Takáč, and Olivier Fercoq. SDNA: stochastic dual Newton ascent for empirical risk minimization. In *Proceedings of The 33rd International Conference on Machine Learning*, pages 1823–1832, 2016.

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

[43] Aaditya Ramdas. Rows vs columns for linear systems of equations - randomized Kaczmarz or coordinate descent ? *arXiv:1406.5295*, 2014.

[44] Peter Richtárik and Martin Takáč. Distributed coordinate descent method for learning with big data. *Journal of Machine Learning Research*, 17(75):1–25, 2016.

[45] Peter Richtárik and Martin Takáč. On optimal probabilities in stochastic coordinate descent methods. *Optimization Letters*, 10(6):1233–1243, 2016.

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

[47] Peter Richtárik and Martin Takáč. Parallel coordinate descent methods for big data optimization. *Mathematical Programming*, 156(1):433–484, 2016.

[48] H. Robbins and S. Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, 22:400–407, 1951.

[49] R. Tyrrell Rockafellar. Monotone operators and the proximal point algorithm. *SIAM Journal on Control and Optimization*, 14(5):877–898, 1976.

[50] V. Rokhlin and Tygert M. A fast randomized algorithm for overdetermined linear least-squares regression. *Proceedings of the National Academy of Sciences*, 105(36):13212–13218, 2008.

[51] Yousef Saad. *Iterative Methods for Sparse Linear Systems*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2nd edition, 2003.

[52] Mark Schmidt, Nicolas Le Roux, and Francis Bach. Minimizing finite sums with the stochastic average gradient. *arXiv:1309.2388*, 2013.

[53] Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

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

[55] Shai Shalev-Shwartz and Tong Zhang. Stochastic dual coordinate ascent methods for regularized loss. *Journal of Machine Learning Research*, 14(1):567–599, 2013.

[56] Jack Sherman and Winifred J. Morrison. Adjustment of an inverse matrix corresponding to changes in the elements of a given column or a given row of the original matrix (abstract). *Annals of Mathematical Statistics*, 20(4):621, 1949.

[57] Daniel A. Spielman and Shang-Hua Teng. Nearly-Linear Time Algorithms for Preconditioning and Solving Symmetric, Diagonally Dominant Linear Systems. 35(3):835–885, 2006.

[58] Nathan Srebro, Karthik Sridharan, and Ambuj Tewari. Optimistic rates for learning with a smooth loss. *arXiv preprint arXiv:1009.3896*, 2010.

[59] S. U. Stich, C. L. Müller, and B. Gärtner. Optimization of convex functions with random pursuit. *SIAM Journal on Optimization*, 23(2):1284–1309, 2014.

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

[61] 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.

[62] Roberto Tempo, Giuseppe Calafiore, and Fabrizio Dabbene. *Randomized Algorithms for Analysis and Control of Uncertain Systems*. Springer-Verlag, New York, 2013.

[63] Max A. Woodbury. The stability of out-input matrices. Technical report, 1949.

[64] Peilin Zhao and Tong Zhang. Stochastic optimization with importance sampling. *arXiv:1401.2753*, 2014.

[65] Anastasios Zouzias and Nikolaos Freris. Randomized extended Kaczmarz for solving least-squares. *arXiv:1205.5770*, page 19, 2012.

A Stochastic proximal point method

As claimed in the introduction, here we show (see Theorem A.3 below) that the stochastic proximal point method (9) is equivalent to stochastic gradient descent (7). First, we state a couple of lemmas, starting with the Sherman-Morrison-Woodbury matrix inversion formula [56, 63].

Lemma A.1 (Sherman-Morrison-Woodbury)**.** Let $\mathbf{M} \in \mathbb{R}^{n\times n}$, $\mathbf{C} \in \mathbb{R}^{n\times q}$, $\mathbf{N} \in \mathbb{R}^{q\times q}$ and $\mathbf{D} \in \mathbb{R}^{q\times n}$, with $\mathbf{M}$ and $\mathbf{N}$ being invertible. Then

$$(\mathbf{M}+\mathbf{CND})^{-1} = \mathbf{M}^{-1} - \mathbf{M}^{-1}\mathbf{C}\left(\mathbf{N}^{-1}+\mathbf{DM}^{-1}\mathbf{C}\right)^{-1}\mathbf{DM}^{-1}.$$

The next result, Lemma A.2, is trivially true if $\mathbf{M}$ is positive definite. Indeed, in that case, $(\mathbf{M}^\dagger)^{1/2}\mathbf{M}(\mathbf{M}^\dagger)^{1/2} = \mathbf{I}$, and the statement follows. However, in general, $(\mathbf{M}^\dagger)^{1/2}\mathbf{M}(\mathbf{M}^\dagger)^{1/2}$ is not equal to the identity; the lemma therefore says that the expression on the left hand side still behaves as if it was.

Lemma A.2. Let $\mathbf{M}$ be a symmetric positive semidefinite matrix. Then for all $\mu > 0$ we have the identity:

$$(\mathbf{M}^\dagger)^{1/2}\left(\mathbf{I}+\frac{1}{\mu}(\mathbf{M}^\dagger)^{1/2}\mathbf{M}(\mathbf{M}^\dagger)^{1/2}\right)^{-1}(\mathbf{M}^\dagger)^{1/2} = \frac{\mu}{1+\mu}\mathbf{M}^\dagger. \tag{84}$$

Proof. Let $\mathbf{M} = \mathbf{UDU}^\top$ be the eigenvalue decomposition of $\mathbf{M}$. Then $\mathbf{M}^\dagger = \mathbf{UD}^\dagger\mathbf{U}^\top$, and it is easy to show that identity (84) holds if it holds for $\mathbf{M}$ being diagonal. If $\mathbf{M}$ is diagonal, then matrices on both sides of (84) are diagonal, which means we can compare the individual diagonal entries. It is easy

to see that if $\mathbf{M}_{ii}=0$, then the ith diagonal element of the matrices on both sides of (84) is zero. If $\mathbf{M}_{ii}>0$, then ith diagonal element of the matrix on the left hand side of (84) is

$$\mathbf{M}_{ii}^{-1/2}\left(1+\frac{1}{\mu}\right)^{-1}\mathbf{M}_{ii}^{-1/2}=\mathbf{M}_{ii}^{-1}\left(1+\frac{1}{\mu}\right)^{-1}=\frac{\mu}{1+\mu}\mathbf{M}_{ii}^{-1}.$$

□

We are now ready to prove the equivalence result.

Theorem A.3. If $0<\omega\leq 1$, then Algorithms (9) and (7) are equivalent. That is, for every $x\in\mathbb{R}^n$, $\mu\geq 0$ and matrix $\mathbf{S}$ with m rows we have[a]

$$x-\omega\nabla f_{\mathbf{S}}(x)=\arg\min_{z\in\mathbb{R}^n} f_{\mathbf{S}}(z)+\frac{1-\omega}{2\omega}\|z-x\|_{\mathbf{B}}^2.$$

[a]Note that the identity trivially holds for $\omega=0$ if we understand the function on the right hand side in the limit sense: $\omega\to 0$ from the right. That is, $x=\arg\min_z \|z-x\|_{\mathbf{B}}^2$.

Proof. The identity holds[3] for $\omega=1$. This follows (31) in view of the fact that $f_{\mathbf{S}}$ is nonnegative. If $0<\omega<1$, then under the substitution $\mu=\frac{\omega-1}{\omega}$, the statement is equivalent to requiring that

$$x-\frac{1}{1+\mu}\nabla f_{\mathbf{S}}(x)=\arg\min_{z\in\mathbb{R}^n} f_{\mathbf{S}}(z)+\frac{\mu}{2}\|z-x\|_{\mathbf{B}}^2 \tag{85}$$

holds for any $\mu>0$.

The minimizer of the stochastic fixed point iteration (right hand side of (85)) can be computed by setting the gradient to zero: $0=\mathbf{A}^\top\mathbf{H}(\mathbf{A}z-b)+\mu\mathbf{B}(z-x)$, whence $z_*=(\mu\mathbf{B}+\mathbf{A}^\top\mathbf{H}\mathbf{A})^{-1}(\mathbf{A}^\top\mathbf{H}b+\mu\mathbf{B}x)$. In view of the formula for the stochastic gradient (27), our goal is therefore to show that

$$x-\frac{1}{1+\mu}\mathbf{B}^{-1}\mathbf{A}^\top\mathbf{H}(\mathbf{A}x-b)=(\mu\mathbf{B}+\mathbf{A}^\top\mathbf{H}\mathbf{A})^{-1}(\mathbf{A}^\top\mathbf{H}b+\mu\mathbf{B}x). \tag{86}$$

By comparing the terms involving x and those that do not in (86), it is sufficient to show that

$$\mu(\mu\mathbf{B}+\mathbf{A}^\top\mathbf{H}\mathbf{A})^{-1}\mathbf{B}=\mathbf{I}-\frac{1}{1+\mu}\mathbf{B}^{-1}\mathbf{A}^\top\mathbf{H}\mathbf{A}, \tag{87}$$

and

$$(\mu\mathbf{B}+\mathbf{A}^\top\mathbf{H}\mathbf{A})^{-1}\mathbf{A}^\top\mathbf{H}b=\frac{1}{1+\mu}\mathbf{B}^{-1}\mathbf{A}^\top\mathbf{H}b. \tag{88}$$

Let us now compute the inverse matrix in the expression defining z_*. First, we have

$$(\mu\mathbf{B}+\mathbf{A}^\top\mathbf{H}\mathbf{A})^{-1} \;=\; \mathbf{B}^{-1/2}\left(\mu\mathbf{I}+\mathbf{B}^{-1/2}\mathbf{A}^\top\mathbf{H}\mathbf{A}\mathbf{B}^{-1/2}\right)^{-1}\mathbf{B}^{-1/2}. \tag{89}$$

Let $\mathbf{K}$ be the symmetric square root of the symmetric positive semidefinite matrix $(\mathbf{S}^\top\mathbf{A}\mathbf{B}^{-1}\mathbf{A}^\top\mathbf{S})^\dagger$. This means that we can write $\mathbf{H}=\mathbf{S}\mathbf{K}^2\mathbf{S}^\top$. We now compute the inverse (89) by applying Lemma A.1 with $\mathbf{M}=\mu\mathbf{I}$, $\mathbf{C}=\mathbf{B}^{-1/2}\mathbf{A}^\top\mathbf{S}\mathbf{K}$, $\mathbf{N}=\mathbf{I}$ (of appropriate size) and $\mathbf{D}=\mathbf{C}^\top$:

$$\left(\mu\mathbf{I}+\mathbf{B}^{-1/2}\mathbf{A}^\top\mathbf{H}\mathbf{A}\mathbf{B}^{-1/2}\right)^{-1} \;=\; \frac{\mathbf{I}}{\mu}-\frac{1}{\mu^2}\mathbf{B}^{-1/2}\mathbf{A}^\top\mathbf{S}\mathbf{K}\left(\mathbf{I}+\frac{1}{\mu}\mathbf{K}\mathbf{S}^\top\mathbf{A}\mathbf{B}^{-1}\mathbf{A}^\top\mathbf{S}\mathbf{K}\right)^{-1}\mathbf{K}\mathbf{S}^\top\mathbf{A}\mathbf{B}^{-1/2}.$$

[3]In this case we interpret this identity as meaning that the vector on the left hand side is a minimizer of the function on the right hand side (as there may be multiple minimizers).

In view of (89), pre and post-multiplying both sides of the last identity by $\mathbf{B}^{-1/2}$, and subsequently applying Lemma A.2 with $\mathbf{M} = \mathbf{S}^\top \mathbf{A}\mathbf{B}^{-1}\mathbf{A}^\top \mathbf{S}$ yields

$$\begin{aligned}\left(\mu \mathbf{B} + \mathbf{A}^\top \mathbf{H}\mathbf{A}\right)^{-1} &= \frac{\mathbf{B}^{-1}}{\mu} - \frac{1}{\mu^2}\mathbf{B}^{-1}\mathbf{A}^\top \mathbf{S}\mathbf{K}\left(\mathbf{I} + \frac{1}{\mu}\mathbf{K}\mathbf{S}^\top \mathbf{A}\mathbf{B}^{-1}\mathbf{A}^\top \mathbf{S}\mathbf{K}\right)^{-1}\mathbf{K}\mathbf{S}^\top \mathbf{A}\mathbf{B}^{-1} \\ &\overset{(84)}{=} \frac{\mathbf{B}^{-1}}{\mu} - \frac{1}{\mu^2}\mathbf{B}^{-1}\mathbf{A}^\top \mathbf{S}\left(\frac{\mu \mathbf{K}^2}{1+\mu}\right)\mathbf{S}^\top \mathbf{A}\mathbf{B}^{-1} \\ &= \frac{\mathbf{B}^{-1}}{\mu} - \frac{\mathbf{B}^{-1}\mathbf{A}^\top \mathbf{H}\mathbf{A}\mathbf{B}^{-1}}{\mu(1+\mu)}.\end{aligned}$$

Given the above formula for the inverse, identity (87) follows immediately. Identity (88) follows using the facts that $b = \mathbf{A}x_*$ and $(\mathbf{B}^{-1}\mathbf{Z})^2 = \mathbf{B}^{-1}\mathbf{Z}$, where $\mathbf{Z} = \mathbf{A}^\top \mathbf{H}\mathbf{A}$. □

B Smallest nonzero eigenvalue

We are using the following inequality in the proof of Theorem 4.8.

Lemma B.1. If Assumption 3.5 holds, then for all $x \in \mathrm{Range}\left(\mathbf{B}^{-1/2}\mathbf{A}^\top\right)$ we have:

$$x^\top \mathbf{B}^{-1/2}\mathrm{E}\left[\mathbf{Z}\right]\mathbf{B}^{-1/2}x \geq \lambda_{\min}^+(\mathbf{B}^{-1/2}\mathrm{E}\left[\mathbf{Z}\right]\mathbf{B}^{-1/2})x^\top x \tag{90}$$

Proof. It is known that for any matrix $\mathbf{M} \in \mathbb{R}^{m\times n}$, the inequality $x^\top \mathbf{M}^\top \mathbf{M}x \geq \lambda_{\min}^+(\mathbf{M}^\top \mathbf{M})x^\top x$ holds for all $x \in \mathrm{Range}\left(\mathbf{M}^\top\right)$. Applying this with $\mathbf{M} = (\mathrm{E}\left[\mathbf{Z}\right])^{1/2}\mathbf{B}^{-1/2}$, we see that (90) holds for all $x \in \mathrm{Range}\left(\mathbf{B}^{-1/2}(\mathrm{E}\left[\mathbf{Z}\right])^{1/2}\right)$. However,

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

where the last identity follows by combining Assumption 3.5 and Theorem 3.6. □

C Linear difference equations

The proof of Theorem 5.3 uses the following standard linear recurrence relations result [15, 13].

Lemma C.1. Consider the following linear homogeneous recurrence relation of degree 2 with constant coefficients: $\xi_{k+1} = E\xi_k + F\xi_{k-1}$, with $\xi_0, \xi_1 \in \mathbb{R}$.

(i) $\xi_k \to 0$ if and only if both roots of the characteristic polynomial, $r^2 - Er^2 - F$, lie strictly inside the unit complex circle.

(ii) Assume that $E^2 + 4F < 0$, i.e., that both roots are complex (the roots are $\alpha + i\beta$ and $\alpha - i\beta$, where $\alpha = E/2$ and $\beta = \sqrt{-E^2 - 4F}/2$). Then there are (complex) constants C_0, C_1, depending on the initial conditions ξ_0, ξ_1, such that $\xi_k = 2M^k\left(C_0\cos(\theta k) + C_1\sin(\theta k)\right)$, where $M = \sqrt{\alpha^2 + \beta^2}$, and θ is such that $\alpha = M\cos(\theta)$ and $\beta = M\sin(\theta)$.

D Notation glossary

The Basics		
$\mathbf{A}, b$	$m \times n$ matrix and $m \times 1$ vector defining the system $\mathbf{A}x = b$	
$\mathcal{L}$	$\{x \,:\, \mathbf{A}x = b\}$ (solution set of the linear system)	
$\mathbf{B}$	$n \times n$ symmetric positive definite matrix	
$\langle x, y \rangle_{\mathbf{B}}$	$x^\top \mathbf{B} y$ ($\mathbf{B}$-inner product)	
$\|x\|_{\mathbf{B}}$	$\sqrt{\langle x, x \rangle_{\mathbf{B}}}$ ($\mathbf{B}$-norm)	
$\mathbf{M}^\dagger$	Moore-Penrose pseudoinverse of matrix $\mathbf{M}$	
$\mathbf{S}$	a random real matrix with m rows	
$\mathcal{D}$	distribution from which matrix $\mathbf{S}$ is drawn ($\mathbf{S} \sim \mathcal{D}$)	
$\mathbf{H}$	$\mathbf{S}(\mathbf{S}^\top \mathbf{A}\mathbf{B}^{-1}\mathbf{A}^\top \mathbf{S})^\dagger \mathbf{S}^\top$	(22)
$\mathbf{Z}$	$\mathbf{A}^\top \mathbf{H} \mathbf{A}$	(23)
$\mathrm{Range}\,(\mathbf{M})$	range space of matrix $\mathbf{M}$	
$\mathrm{Null}\,(\mathbf{M})$	null space of matrix $\mathbf{M}$	
$\mathrm{Trace}\,(\mathbf{M})$	trace of matrix $\mathbf{M}$	
$\mathrm{Prob}(\cdot)$	probability of an event	
$\mathrm{E}\,[\cdot]$	expectation	
Projections		
$\Pi_{\mathcal{L}}^{\mathbf{B}}(x)$	projection of x onto $\mathcal{L}$ in the $\mathbf{B}$-norm	(21)
$\mathbf{M}^{\dagger_{\mathbf{B}}}$	$\mathbf{B}^{-1}\mathbf{M}^\top(\mathbf{M}\mathbf{B}^{-1}\mathbf{M}^\top)^\dagger$ ($\mathbf{B}$-pseudoinverse of $\mathbf{M}$)	(20)
$\mathbf{B}^{-1}\mathbf{Z}$	projection matrix, in the $\mathbf{B}$-norm, onto $\mathrm{Range}\left(\mathbf{B}^{-1}\mathbf{A}^\top \mathbf{S}\right)$	(27)
Optimization		
$\mathcal{X}$	set of minimizers of f	Thm 3.4
x_*	a point in $\mathcal{L}$	
$f_{\mathbf{S}}$, $\nabla f_{\mathbf{S}}$, $\nabla^2 f_{\mathbf{S}}$	stochastic function, its gradient and Hessian	(25)–(30)
$\mathcal{L}_{\mathbf{S}}$	$\{x \,:\, \mathbf{S}^\top \mathbf{A}x = \mathbf{S}^\top b\}$ (set of minimizers of $f_{\mathbf{S}}$)	Lem 3.1
f	$\mathrm{E}\,[f_{\mathbf{S}}]$	(32), Lem 3.3
∇f	gradient of f with respect to the $\mathbf{B}$-inner product	
$\nabla^2 f$	$\mathbf{B}^{-1}\mathrm{E}\,[\mathbf{Z}]$ (Hessian of f in the $\mathbf{B}$-inner product)	
Eigenvalues		
$\mathbf{W}$	$\mathbf{B}^{-1/2}\mathrm{E}\,[\mathbf{Z}]\,\mathbf{B}^{-1/2}$ (psd matrix with the same spectrum as $\nabla^2 f$)	
$\lambda_1, \ldots, \lambda_n$	eigenvalues of $\mathbf{W}$	
Λ	$\mathbf{Diag}\,(\lambda_1, \ldots, \lambda_n)$ (diagonal matrix of eigenvalues)	
$\mathbf{U}$	$[u_1, \ldots, u_n]$ (eigenvectors of $\mathbf{W}$)	
$\mathbf{U}\Lambda\mathbf{U}^\top$	eigenvalue decomposition of $\mathbf{W}$	(40)
$\lambda_{\max}, \lambda_{\min}^+$	largest and smallest nonzero eigenvalues of $\mathbf{W}$	
ζ	$\lambda_{\max}/\lambda_{\min}^+$ (condition number of $\mathbf{W}$)	(14), (41)
Algorithms		
ω	relaxation parameter / stepsize	Alg 1–3
τ	parallelism parameter	Alg 2
γ	acceleration parameter	Alg 3

Table 2: Frequently used notation.