Numerical Algorithms (2022) 91:749–772
https://doi.org/10.1007/s11075-022-01280-4

ORIGINAL PAPER

Cocoercivity, smoothness and bias in variance-reduced stochastic gradient methods

Martin Morin[1] · **Pontus Giselsson**[1]

Received: 22 December 2020 / Accepted: 10 February 2022 / Published online: 13 April 2022

Abstract
With the purpose of examining biased updates in variance-reduced stochastic gradient methods, we introduce SVAG, a SAG/SAGA-like method with adjustable bias. SVAG is analyzed in a cocoercive root-finding setting, a setting which yields the same results as in the usual smooth convex optimization setting for the ordinary proximal-gradient method. We show that the same is not true for SVAG when biased updates are used. The step-size requirements for when the operators are gradients are significantly less restrictive compared to when they are not. This highlights the need to not rely solely on cocoercivity when analyzing variance-reduced methods meant for optimization. Our analysis either match or improve on previously known convergence conditions for SAG and SAGA. However, in the biased cases they still do not correspond well with practical experiences and we therefore examine the effect of bias numerically on a set of classification problems. The choice of bias seem to primarily affect the early stages of convergence and in most cases the differences vanish in the later stages of convergence. However, the effect of the bias choice is still significant in a couple of cases.

Keywords Variance reduction · Stochastic gradient · Bias · SAG · SAGA · Monotone inclusion · Root finding

1 Introduction

Variance-reduced stochastic gradient (VR-SG) methods is a family of iterative optimization algorithms that combine the low per-iteration computational cost of the

✉ Martin Morin
martin.morin@control.lth.se

Pontus Giselsson
pontus.giselsson@control.lth.se

[1] Department of Automatic Control, Lund University, Lund, Sweden

ordinary stochastic gradient descent and the attractive convergence properties of gradient descent. Just as ordinary stochastic gradient descent, VR-SG methods solve smooth optimization problems on finite sum form,

$$\min_{x\in\mathbb{R}^N} \frac{1}{n}\sum_{j=1}^{n} f_i(x) \tag{1}$$

where, for all $i \in \{1,\dots,n\}$, $f_i : \mathbb{R}^N \to \mathbb{R}$ is a convex function that is L-smooth, i.e., f_i is differentiable with L-Lipschitz continuous gradient. These types of problems are common in model fitting, supervised learning, and empirical risk minimization which, together with the nice convergence properties of VR-SG methods, has lead to a great amount of research on VR-SG methods and the development of several different variants, e.g., [1, 15, 16, 21–25, 27, 30, 33, 40, 41, 43, 45].

Broadly speaking, VR-SG methods form a stochastic estimate of the objective gradient by combining one or a few newly evaluated terms of the gradient with all previously evaluated terms. Classic examples of this can be seen in the SAG [27, 40] and SAGA [15] algorithms. Given some initial iterates $x^0, y_1^0, \dots, y_n^0 \in \mathbb{R}^N$ and step-size $\lambda > 0$, SAGA samples i^k uniformly from $\{1,\dots,n\}$ and then updates the iterates as

$$\begin{aligned} x^{k+1} &= x^k - \lambda\Big(\nabla f_{i^k}(x^k) - y_{i^k}^k + \tfrac{1}{n}\textstyle\sum_{j=1}^{n} y_j^k\Big),\\ y_{i^k}^{k+1} &= \nabla f_{i^k}(x^k),\\ y_j^{k+1} &= y_j^k \quad \text{for all} \quad j \neq i^k, \end{aligned}$$

for $k \in \{0,1,\dots\}$. The update of x^{k+1} is said to be unbiased since the expected value of x^{k+1} at iteration k is equal to an ordinary gradient descent update. This is in contrast to the biased SAG, which is identical to SAGA except that the update of x^{k+1} is

$$x^{k+1} = x^k - \lambda\left(\tfrac{1}{n}\big(\nabla f_{i^k}(x^k) - y_{i^k}^k\big) + \tfrac{1}{n}\textstyle\sum_{j=1}^{n} y_j^k\right)$$

and the expected value of x^{k+1} now includes a term containing the old gradients $\frac{1}{n}\sum_{i=1}^{n} y_i^k$. Although SAG shows that unbiasedness is not essential for the convergence of VR-SG methods, the effects of this bias are unclear. The majority of VR-SG methods are unbiased but existing works have not established any clear advantage of either the biased SAG or the unbiased SAGA. This paper will examine the effect of bias and its interplay with different problem assumptions for SAG/SAGA-like methods.

1.1 Problem and algorithm

Instead of solving (1) directly, we consider a closely related but more general root-finding problem. Throughout the paper, we consider the Euclidean space $\mathbb{R}^N$ and the problem of finding $x \in \mathbb{R}^N$ such that

$$0 = Rx \coloneqq \tfrac{1}{n}\textstyle\sum_{i=1}^{n} R_i x \tag{2}$$

where $R_i : \mathbb{R}^N \to \mathbb{R}^N$ is $\frac{1}{L}$-cocoercive—see Section 2—for all $i \in \{1, \dots, n\}$. Since L-smoothness of a convex function is equivalent to $\frac{1}{L}$-cocoercivity of the gradient [2, Corollary 18.17], the smooth optimization problem in (1) can be recovered by setting $R_i = \nabla f_i$ for all $i \in \{1, \dots, n\}$ in (2). Problem (2) is also interesting in its own right with it and the closely related fixed point problem of finding $x \in \mathbb{R}^N$ such that $x = (Id - \alpha R)x$ where $\alpha \in (0, 2L^{-1})$ both having applications in for instance feasibility and non-linear signal recovery problems, see [8, 10, 13] and the references therein. To solve this problem, we present the *Stochastic Variance Adjusted Gradient* (SVAG) algorithm.

Algorithm 1 SVAG.

input operators $R_i : \mathbb{R}^N \to \mathbb{R}^N$, initial state $x^0 \in \mathbb{R}^N$ and $y_1^0, \dots, y_n^0 \in \mathbb{R}^N$, step-size $\lambda > 0$, innovation weight $\theta \in \mathbb{R}$
for $k = 0, 1, \dots$ **do**
  Sample i^k uniformly from $\{1, \dots, n\}$
  $x^{k+1} = x^k - \lambda\left(\frac{\theta}{n}\left(R_{i^k}x^k - y_{i^k}^k\right) + \frac{1}{n}\sum_{j=1}^{n} y_j^k\right)$
  $y_{i^k}^{k+1} = R_{i^k}x^k$
  $y_j^{k+1} = y_j^k$ for all $j \neq i^k$
end for

SVAG is heavily inspired by SAG and SAGA with both being special cases, $\theta = 1$ and $\theta = n$ respectively. Just like SAG and SAGA, in each iteration, SVAG evaluates one operator R_{i^k} and stores the results in $y_{i^k}^{k+1}$. An estimate of the full operator is then formed as

$$Rx^k \approx \widetilde{R}^k = \tfrac{\theta}{n}(R_{i^k}x^k - y_{i^k}^k) + \tfrac{1}{n}\textstyle\sum_{j=1}^{n} y_j^k. \tag{3}$$

The scalar θ determine how much weight should be put on the new information gained from evaluating $R_{i^k}x^k$. If the innovation, $R_{i^k}x^k - y_{i^k}^k$, is highly correlated with the total innovation, $Rx^k - \frac{1}{n}\sum_{j=1}^{n} y_j^k$, a large innovation weight θ can be chosen and vice versa. The innovation weight θ also determines the bias of SVAG. Taking the expected value $\widetilde{R}^k$ given the information at iteration k gives

$$\mathbb{E}[\widetilde{R}^k | x^k, y_1^k, \dots, y_n^k] = \tfrac{\theta}{n}Rx^k + \left(1 - \tfrac{\theta}{n}\right)\tfrac{1}{n}\textstyle\sum_{j=1}^{n} y_j^k$$

which reveals that $\widetilde{R}^k$ is an unbiased estimate of Rx^k if $\theta = n$, i.e., in the SAGA case. Any other choice, for instance SAG where $\theta = 1$, yields a bias towards $\frac{1}{n}\sum_{j=1}^{n} y_j^k$.

1.2 Contribution

The theory behind finding roots of monotone operators in general, and cocoercive operators in particular, has been put to good use when analyzing first-order optimization methods, examples include [2, 4, 14, 26, 38, 44]. For instance can both the proximal-gradient and ADMM methods be seen as instances of classic root-finding fixed point iterations and analyzed as such, namely forward-backward and Douglas–Rachford respectively. The resulting analyses can often be simple and intuitive and even though the root-finding formulation is more general—not all cocoercive operators are gradients of convex functions—the analyses are not necessarily more conservative. For example, analyzing proximal-gradient as forward-backward splitting yields the same rates and step-size conditions as analyzing it as a minimization method in the smooth/cocoercive setting, see for instance [32, Theorem 2.1.14] and [2, Example 5.18 and Proposition 4.39]. However, the main contribution of this paper is to show that the same is not true for VR-SG methods, in particular it is not true for SVAG when it is biased.

The results consist of two main convergence theorems for SVAG: one in the cocoercive operator case and one in the cocoercive gradient case, the later being equivalent to the minimization of a smooth and convex finite sum. Both of these theorems match or improve upon previously known results for the SAG and SAGA special cases. Comparing the two settings reveal that SVAG can use significantly larger step-sizes, with faster convergence as a result, in the cocoercive gradient case compared to the general cocoercive operator case. In the operator case, an upper bound on the step-size that scales as $\mathcal{O}(n^{-1})$ is found where n is the number of terms in (2). However, the restrictions on the step-size loosen with reduced bias and the unfavorable $\mathcal{O}(n^{-1})$ scaling disappears completely when SVAG is unbiased. In the gradient case, this bad scaling never occurs, regardless of bias. We provide examples in which SVAG diverges with step-sizes larger than the theoretical upper bounds in the operator case. Since the gradient case is proven to converge with much larger step-sizes, this verifies the difference between the convergence behavior of cocoercive operators and gradients.

These results indicate that it is inadvisable to only rely on the more general monotone operator theory and not explicitly use the gradient property when analyzing VR-SG methods meant for optimization. However, the large impact of bias in the cocoercive operator setting also raises the question regarding its importance in other non-gradient settings as well. One such setting of interest, where the operators are not gradients of convex functions, is the case of saddle-point problems. These problems are of importance in optimization due to their use in primal-dual methods but recently they have also gained a lot of attention due to their applications in the training of GANs in machine learning. Because of this, and due to the attractive properties of VR-SG methods in the convex optimization setting, efforts have gone into applying VR-SG methods to saddle-point problems as well [5, 7, 34, 42, 46]. Most of these efforts have been unbiased, something our analysis suggests is wise. With that said,

it is important to note that our analysis is often not directly applicable due the fact that saddle-point problems rarely are cocoercive.

The main reason for the recent rise in popularity of variance-reduced stochastic methods is their use in the optimization setting, but, although bias plays a big role in the cocoercive operator case, our results are not as clear in this setting. For instance, the theoretical results for the SAG and SAGA special cases yield identical rates and step-size conditions with no clear advantage to either special case. Further experiments are therefore performed where several different choices of bias in SVAG are examined on a set of logistic regression and SVM optimization problems. However, the results of these experiments are in line with existing works with no significant advantage of any particular bias choice in SVAG. Although the performance difference is significant in some cases, no single choice of bias performs best for all problems and all bias choices eventually converge with the same rate in the majority of the cases. Furthermore, the theoretical maximal step-size can routinely be exceeded in these experiments, indicating that there is room for further theoretical improvements.

1.3 Related work

There is a large array of options for solving (2). For $n \in \{1, 2, 3, 4\}$, several operator splitting methods exist with varying assumptions on the operator properties, see for instance [4, 18, 19, 28, 29, 44] and the references therein. However, while these methods also can be applied for larger n by simply regrouping the terms, they do not utilize the finite sum structure of the problem. Algorithms have therefore been designed to utilize this structure for arbitrary large n with the hopes of reducing the total computational costs, e.g., [9–11, 36]. In particular the problem and method in [10] is closely related to the root-finding problem and algorithm considered in this paper.

Using the notation of [10], when $T_0 = \mathrm{Id}$, the fixed point problem of [10] can be mapped to (2) via $R_i = \omega_i(\mathrm{Id} - T_i)$ and vice verse. [1] Many applications considered in [10] can therefore, at least in part, be tackled with our algorithm as well. In particular, the problem of finding common fixed points of firmly nonexpansive operators can directly be solved by our algorithm. However, [10] is more general in that it allows for $T_0 \neq Id$ and works in general real Hilbert spaces. Looking at the algorithm of [10] we see that, just as our algorithm is a generalization of SAG/SAGA, it can be seen as a generalization of Finito [16], another classic VR-SG method. It generalize Finito in several way, for instance it allows for an additional proximal/backward step and it replaces the stochastic selection with a different selection criteria. However, in the optimization setting it still suffers from the same drawback as Finito when compared to SAG/SAGA-like algorithms. It still needs to store a full copy of the iterate for each term in objective. Since SAG, SAGA, and SVAG only need to store the gradient of each term, they can utilize any potential structure of the gradients to reduce the storage requirements [27]. Although the differences above are interesting

[1] If T_i is α_i-averaged, as assumed in [10], R_i is $(2\alpha_i\omega_i)^{-1}$-cocoercive.

in their own right, the notion of bias we examine in this paper is not applicable to Finito-like algorithms.

SAG and SAGA were compared in [15] but with no direct focus on the effects of bias. Other examples of research on SAG and SAGA include acceleration, sampling strategy selection, and ways to reduce the memory requirement [20, 22, 31, 35, 39, 47]. However, none of these works, including [31] that was written by the authors, analyze the biased case we consider in this paper. Even the works considering non-uniform sampling of gradients [20, 31, 35, 39] perform some sort of bias correction in order to remain unbiased. Furthermore, in order to keep the focus on the effects of the bias we have refrained from bringing in such generalizations into this work, making it distinct from the above research. To the authors' knowledge, the only theoretical convergence result for biased VR-SG methods are the ones for SAG [27, 40]. But, since they only consider SAG, they fail to capture the breadth of SVAG and our proof is the first to simultaneously capture SAG, SAGA, and more.

Since the release of the first preprint of this paper, [17] has also provided a proof covering the gradient case of both SAG and SAGA, and some choices of bias in SVAG. All though [17] does not consider cocoercive operators, it is some sense more general with them considering a general biased stochastic estimator of the gradient. This generality comes at the cost of a more conservative analysis with their step-size scaling with $\mathcal{O}(n^{-1})$ in all cases.

2 Preliminaries and notation

Let $\mathbb{R}$ denote the real numbers and let the natural numbers be denoted $\mathbb{N} = \{0, 1, 2, \dots\}$. Let $\langle \cdot, \cdot \rangle$ denote the standard Euclidean inner product and $\|\cdot\| = \sqrt{\langle \cdot, \cdot \rangle}$ the standard 2-norm. The scaled inner product and norm we denote as $\langle \cdot, \cdot \rangle_\Sigma = \langle \Sigma(\cdot), \cdot \rangle$ and $\|\cdot\|_\Sigma = \sqrt{\langle \cdot, \cdot \rangle_\Sigma}$ where Σ is a positive definite matrix. If Σ is not positive definite, $\|\cdot\|_\Sigma$ is not a norm but we keep the notation for convenience.

Let n be the number of operators in (2). The vector $\mathbf{1}$ is the vector of all ones in $\mathbb{R}^n$ and e_i is the vector in $\mathbb{R}^n$ of all zeros except the i:th element which contains a 1. The matrix I is an identity matrix with the size derived from context and $E_i = e_i e_i^T$.

The symbol $\otimes$ denotes the Kronecker product of two matrices. The Kronecker product is linear in both arguments and the following properties hold

$$(A \otimes B)^T = A^T \otimes B^T, \qquad (A \otimes B)(C \otimes D) = (AC) \otimes (BD).$$

In the last property it is assumed that the dimensions are such that the matrix multiplications are well defined. The eigenvalues of $A \otimes B$ are given by

$$\tau_i \mu_j \text{ for all } i \in \{1, \dots, m\},\ j \in \{1, \dots, l\} \tag{4}$$

where τ_i and μ_j are the eigenvalues of A and B respectively.

The Cartesian product of two sets C_1 and C_2 is defined as

$$C_1 \times C_2 = \{(c_1, c_2) \mid c_1 \in C_1, c_2 \in C_2\}.$$

From this definition we see that if C_1 and C_2 are closed and convex, so is $C_1 \times C_2$.

Let $X^\star$ be the set of all solutions of (2),

$$X^\star = \{x \mid 0 = \tfrac{1}{n}\textstyle\sum_{i=1}^n R_i x\}$$

and define $Z^\star$ as the set of primal-dual solutions

$$Z^\star = \{(x, R_1 x, \ldots, R_n x) \mid 0 = \tfrac{1}{n}\textstyle\sum_{i=1}^n R_i x\}.$$

Assuming they exists, $x^\star$ denotes a solution to (2) and $z^\star$ denotes a primal-dual solution, i.e., $x^\star \in X^\star$ and $z^\star \in Z^\star$.

A single valued operator $R : \mathbb{R}^N \to \mathbb{R}^N$ is $\frac{1}{L}$-cocoercive if

$$\langle Rx - Ryx - y\rangle \geq \frac{1}{L}\|Rx - Ry\|^2 \tag{5}$$

holds for all $x, y \in \mathbb{R}^N$. An operator that is $\frac{1}{L}$-cocoercive is L-Lipschitz continuous. The set of zeros of a cocoercive operator R is closed and convex.

A differentiable convex function $f : \mathbb{R}^N \to \mathbb{R}$ is called L-smooth if the gradient is $\frac{1}{L}$-cocoercive. Equivalently, a differentiable convex function is L-smooth if

$$f(y) \leq f(x) + \langle \nabla f(x) y - x\rangle + \frac{L}{2}\|y - x\|^2 \tag{6}$$

holds for all $x, y \in \mathbb{R}^N$.

If $f_i : \mathbb{R}^N \to \mathbb{R}$ is a differentiable convex function for each $i \in \{1, \ldots, n\}$, the minimization of $\sum_{i=1}^n f_i(x)$ is equivalent to (2) with $R_i = \nabla f_i$.

For more details regarding monotone operators and convex functions see [2, 32].

To establish almost sure sequence convergence of the stochastic algorithm, the following propositions will be used. The first is from [37] and establishes convergence of non-negative almost super-martingales. The second is based on [12] and provides the tool to show almost sure sequence convergence.

Proposition 2.1 *Let $(\Omega, \mathcal{F}, P)$ be a probability space and $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \ldots$ be a sequence of sub-σ-algebras of $\mathcal{F}$. For all $k \in \mathbb{N}$, let z^k, β^k, ξ^k and ζ^k be non-negative $\mathcal{F}_k$-measurable random variables. If $\sum_{i=0}^\infty \beta^i < \infty$, $\sum_{i=0}^\infty \xi^i < \infty$ and*

$$\mathbb{E}[z^{k+1}|\mathcal{F}_k] \leq (1 + \beta^k) z^k + \xi^k - \zeta^k$$

hold almost surely for all $k \in \mathbb{N}$, then z^k converges a.s. to a finite valued random variable and $\sum_{i=0}^\infty \zeta^i < \infty$ almost surely.

Proof See [37, Theorem 1]. □

Proposition 2.2 *Let Z be a non-empty closed subset of a finite dimensional Hilbert space H, let $\phi : [0, \infty) \to [0, \infty)$ be a strictly increasing function such that $\phi(t) \to \infty$ as $t \to \infty$, and let $(x^k)_{k\in\mathbb{N}}$ be a sequence of H-valued random variables. If $\phi(\|x^k - z\|)$ converges a.s. to a finite valued non-negative random variable for all $z \in Z$, then the following hold:*

1. *$(x^k)_{k\in\mathbb{N}}$ is bounded almost surely.*
2. *Suppose the cluster points of $(x^k)_{k\in\mathbb{N}}$ are a.s. in Z, then $(x^k)_{k\in\mathbb{N}}$ converge a.s. to a Z-valued random variable.*

Proof In finite dimensional Hilbert spaces, these two statements are the same as statements (ii) and (iv) of [12, Proposition 2.3]. Hence, consider the proof of [12, Proposition 2.3] restricted to finite dimensional Hilbert spaces. The proof of (ii) in [12, Proposition 2.3] only relies on the a.s. convergence of $\phi(\|x^k - z\|)$ and hence is implied by the assumptions of this proposition. This proves our first statement. The proof of (iv) in [12, Proposition 2.3] only relies on (iii) of [12, Proposition 2.3] which in turn is implied by (ii) of [12, Proposition 2.3], i.e., our first statement. This proves our second statement. □

3 Convergence

Throughout the analysis we will use the following two assumptions on the operators in (2).

Assumption 3.1 *For each $i \in \{1, \dots, n\}$, let R_i be $\frac{1}{L}$-cocoercive and $X^\star \neq \emptyset$, i.e., (2) has at least one solution.*

Assumption 3.2 *For each $i \in \{1, \dots, n\}$, let $R_i = \nabla f_i$ for some differentiable function f_i and define $F = \frac{1}{n}\sum_{i=1}^n f_i$. Furthermore, let Assumption 3.1 hold, i.e., f_i is L-smooth and convex and* $\operatorname{argmin} F(x)$ *exists.*

3.1 Reformulation

We begin by formalizing and reformulating Algorithm 1 into a more convenient form. Let $(\Omega, \mathcal{F}, P)$ be the underlying probability space of Algorithm 1. The index selected at iteration k is then a uniformly distributed random variable $i^k : \Omega \to \{1, \dots, n\}$. For each $k \in \mathbb{N}$, define the random variable $z^k : \Omega \to \mathbb{R}^{N(n+1)}$ as $z^k = (x^k, y_1^k, \dots, y_n^k)$ where x^k and y_i^k for $i \in \{1, \dots, n\}$ are the iterates of Algorithm 1. Let $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \dots$ be a sequence of sub-σ-algebras of $\mathcal{F}$ such that z^k are $\mathcal{F}_k$-measurable and i^k is independent of $\mathcal{F}_k$. With the operator $\mathbf{B} : \mathbb{R}^{N(n+1)} \to \mathbb{R}^{2Nn}$ defined as $\mathbf{B}(x, y_1, \dots, y_n) = (R_1 x, \dots, R_n x, y_1, \dots, y_n)$, one iteration of Algorithm 1 can be written as

$$z^{k+1} = z^k - (U_{i^k} \otimes I)\mathbf{B}z^k \tag{7}$$

where $z^0 \in \mathbb{R}^{N(n+1)}$ is given and

$$U_i = \begin{bmatrix} \frac{\lambda}{n}\theta e_i^T & -\frac{\lambda}{n}\theta e_i^T + \frac{\lambda}{n}\mathbf{1}^T \\ -E_i & E_i \end{bmatrix}$$

for all $i \in \{1, \dots, n\}$. The vector e_i and the matrix E_i are defined in Section 2.

The following lemma characterizes the zeros of $(U_i \otimes I)\mathbf{B}$ and hence the fixed points of (7) and Algorithm 1.

Lemma 3.3 *Let Assumption 3.1 hold, each $z^\star$ in $Z^\star$ is then a zero of $(U_i \otimes I)\mathbf{B}$ for all $i \in \{1, \dots, n\}$, i.e.,*

$$\forall z^\star \in Z^\star, \forall i \in \{1, \dots, n\}: \quad 0 = (U_i \otimes I)\mathbf{B}z^\star.$$

Furthermore, the set $Z^\star$ is closed and convex and $R_i x^\star = R_i \bar{x}^\star$ for all $x^\star, \bar{x}^\star \in X^\star$ and for all $i \in \{1, \dots, n\}$.

Proof of Lemma 3.3 The zero statement, $0 = (U_i \otimes I)\mathbf{B}z^\star$, follows from definition of $z^\star$. For closedness and convexity of $Z^\star$, we first prove that $R_i x^\star$ is unique for each $i \in \{1, \dots, n\}$. Taking $x, y \in X^\star$, which implies $\sum_{i=1}^n R_i x = \sum_{i=1}^n R_i y = 0$, and using cocoercivity (5) of each R_i gives

$$\begin{aligned} 0 = \langle \textstyle\sum_{i=1}^n R_i x - \sum_{i=1}^n R_i y, x - y \rangle &= \textstyle\sum_{i=1}^n \langle R_i x - R_i y x - y \rangle \\ &\geq \textstyle\sum_{i=1}^n \frac{1}{L} \|R_i x - R_i y\|^2 \geq 0, \end{aligned}$$

hence must $R_i x = R_i y$ for all $i \in \{1, \dots, n\}$. The set $Z^\star$ is a Cartesian product of $X^\star$ and the points $r_i = R_i x^\star$ for $i \in \{1, \dots, n\}$ for any $x^\star \in X^\star$. A set consisting of only one point is closed and convex and $X^\star$ is closed and convex since $\frac{1}{n}\sum_{i=1}^n R_i$ is cocoercive [2, Proposition 23.39], hence is $Z^\star$ closed and convex. □

The operator $\mathbf{B}$ in the reformulated algorithm can be used to enforce the following property on the sequence $(z^k)_{k\in\mathbb{N}}$.

Lemma 3.4 *Let $(\Omega, \mathcal{F}, P)$ be a probability space and $(z^k)_{k\in\mathbb{N}}$ be a sequence of random variables $z^k : \Omega \to \mathbb{R}^{N(n+1)}$. If $\mathbf{B}z^k \to \mathbf{B}z^\star$ a.s. where $z^\star \in Z^\star$, then any cluster point of $(z^k)_{k\in\mathbb{N}}$ will almost surely be in $Z^\star$.*

Proof of Lemma 3.4 Let z be a cluster point of $(z^k)_{k\in\mathbb{N}}$. Take an $\omega \in \Omega$ such that $\mathbf{B}z^k(\omega) \to \mathbf{B}z^\star$. For this ω and for all $k \in \mathbb{N}$, we define the realizations of z and z^k as

$$z(\omega) = (\bar{x}, \bar{y}_1, \dots, \bar{y}_n), \quad z^k(\omega) = (\bar{x}^k, \bar{y}_1^k, \dots, \bar{y}_n^k)$$

where $\bar{x}, \bar{y}_1, \dots, \bar{y}_n \in \mathbb{R}^N$ and $\bar{x}^k, \bar{y}_1^k, \dots, \bar{y}_n^k \in \mathbb{R}^N$ for all $k \in \mathbb{N}$.

Since $\mathbf{B}\bar{z}^k \to \mathbf{B}z^\star$ we directly have $\bar{y}_i^k \to R_i x^\star$ for $x^\star \in X^\star$ and hence must $\bar{y}_i = R_i x^\star$ for all $i \in \{1, \dots, n\}$. Note, $R_i x^\star$ is independent of which $x^\star \in X^\star$ was chosen, see Lemma 3.3. Furthermore, $\mathbf{B}\bar{z}^k \to \mathbf{B}\bar{z}^\star$ implies that $R_i \bar{x}^k \to R_i x^\star$ for all $i \in \{1, \dots, n\}$. Let $(\bar{x}^{k(l)})_{l\in\mathbb{N}}$ be a subsequence converging to $\bar{x}$, then

$$\begin{aligned} \|\tfrac{1}{n}\textstyle\sum_{i=1}^n R_i \bar{x}\| &\leq \|\tfrac{1}{n}\textstyle\sum_{i=1}^n R_i \bar{x}^{k(l)} - \tfrac{1}{n}\sum_{i=1}^n R_i \bar{x}\| + \|\tfrac{1}{n}\textstyle\sum_{i=1}^n R_i \bar{x}^{k(l)}\| \\ &\leq L\|\bar{x}^{k(l)} - \bar{x}\| + \|\tfrac{1}{n}\textstyle\sum_{i=1}^n R_i \bar{x}^{k(l)}\| \to \|\tfrac{1}{n}\textstyle\sum_{i=1}^n R_i x^\star\| = 0 \end{aligned}$$

as $l \to \infty$ where L-Lipschitz continuity of $\frac{1}{n}\sum_{i=1}^n R_i$ was used. This concludes that $\bar{x} \in X^\star$ and since $\bar{y}_i = R_i x^\star = R_i \bar{x}$ for all $i \in \{1, \dots, n\}$ by Lemma 3.3, we have that $z(\omega) \in Z^\star$. Since this hold for any ω such that $\mathbf{B}z^k(\omega) \to \mathbf{B}z^\star$ and the set in $\mathcal{F}$ of all such ω have probability one due to the almost sure convergence of $\mathbf{B}z^k \to \mathbf{B}z^\star$, we have $z \in Z^\star$ almost surely. □

The reformulation (7) further allows us to concisely formulate two Lyapunov inequalities.

Lemma 3.5 *Let Assumption 3.1 hold, the update (7) then satisfies*

$$\begin{aligned}\mathbb{E}[\|z^{k+1}-z^\star\|^2_{H\otimes I}|\mathcal{F}_k] \\ \leq \|z^k-z^\star\|^2_{H\otimes I}-\|\mathbf{B}z^k-\mathbf{B}z^\star\|^2_{(2M-\mathbb{E}[U_{i^k}^T H U_{i^k}]-\xi I)\otimes I} \\ -\xi nL\langle Rx^kx^k-x^\star\rangle\end{aligned}$$

for all $k\in\mathbb{N}$ and $\xi\in[0,\frac{2\lambda}{nL}]$, where the matrices H and M are given by

$$H=\begin{bmatrix}1 & -\frac{\lambda}{n}(n-\theta)\mathbf{1}^T\\ -\frac{\lambda}{n}(n-\theta)\mathbf{1} & \frac{\lambda}{L}I+\frac{\lambda^2}{n^2}(n-\theta)^2\mathbf{1}\mathbf{1}^T\end{bmatrix}$$

and

$$M=\begin{bmatrix}2 & -1\\ -1 & 2\end{bmatrix}\otimes\frac{1}{2n}\frac{\lambda}{L}I-\begin{bmatrix}0 & 1\\ 1 & 0\end{bmatrix}\otimes\frac{\lambda^2}{2n^2}(n-\theta)\mathbf{1}\mathbf{1}^T.$$

Lemma 3.6 *Let Assumption 3.2 hold, the update (7) then satisfies*

$$\mathbb{E}[F((K\otimes I)z^{k+1})|\mathcal{F}_k]\leq F((K\otimes I)z^k)-\|\mathbf{B}z^k-\mathbf{B}z^\star\|^2_{\frac{1}{2}S\otimes I}$$

for all $k\in\mathbb{N}$, where $K=\begin{bmatrix}1 & \frac{\lambda}{n}\mathbf{1}^T\end{bmatrix}$ and

$$\begin{aligned}S &= \begin{bmatrix}2 & -1\\ -1 & 0\end{bmatrix}\otimes(\theta-1)\frac{\lambda}{n^3}\mathbf{1}\mathbf{1}^T-\begin{bmatrix}1 & -1\\ -1 & 1\end{bmatrix}\otimes(\theta-1)^2\frac{L\lambda^2}{n^3}I\\ &+\begin{bmatrix}0 & 1\\ 1 & 0\end{bmatrix}\otimes\frac{\lambda}{n^2}\mathbf{1}\mathbf{1}^T.\end{aligned}$$

Proof of Lemma 3.5 Take $k\in\mathbb{N}$, note that since U_{i^k} is independent of $\mathcal{F}_k$ and z_k is $\mathcal{F}_k$-measurable we have

$$\begin{aligned}\mathbb{E}[\langle(U_{i^k}\otimes I)(\mathbf{B}z^k-\mathbf{B}z^\star)z^k-z^\star\rangle_{H\otimes I}|\mathcal{F}_k]\\ =\langle(H\mathbb{E}[U_{i^k}]\otimes I)(\mathbf{B}z^k-\mathbf{B}z^\star)z^k-z^\star\rangle.\end{aligned}$$

The matrix $H\mathbb{E}[U_{i^k}]$ is given by

$$H\mathbb{E}[U_{i^k}]=\begin{bmatrix}\frac{\lambda}{n}\mathbf{1}^T & 0\\ -\frac{\lambda^2}{n^2}(n-\theta)\mathbf{1}\mathbf{1}^T-\frac{\lambda}{nL}I & \frac{\lambda}{nL}I\end{bmatrix},$$

see the supplementary material for verification of this and other matrix identities. We also note that

$$\langle Rx^k-Rx^\star x^k-x^\star\rangle=\langle(\begin{bmatrix}\frac{1}{n}\mathbf{1}^T & 0\\ 0 & 0\end{bmatrix}\otimes I)(\mathbf{B}z^k-\mathbf{B}z^\star)z^k-z^\star\rangle.$$

Taking $\xi\in[0,\frac{2\lambda}{nL}]$ and putting these two expression together yield

$$\begin{aligned}\mathbb{E}[\langle(U_{i^k}\otimes I)(\mathbf{B}z^k-\mathbf{B}z^\star)z^k-z^\star\rangle_{H\otimes I}|\mathcal{F}_k]-\frac{\xi nL}{2}\langle Rx^k-Rx^\star x^k-x^\star\rangle\\ =\langle(\begin{bmatrix}(\frac{\lambda}{n}-\frac{\xi L}{2})\mathbf{1}^T & 0\\ -\frac{\lambda^2}{n^2}(n-\theta)\mathbf{1}\mathbf{1}^T-\frac{\lambda}{nL}I & \frac{\lambda}{nL}I\end{bmatrix}\otimes I)(\mathbf{B}z^k-\mathbf{B}z^\star)z^k-z^\star\rangle.\end{aligned}$$

Using $\frac{1}{L}$-cocoercivity of R_i for each $i \in \{1, \dots, n\}$ gives

$$\mathbb{E}[\langle (U_{i^k} \otimes I)(\mathbf{B}z^k - \mathbf{B}z^\star)z^k - z^\star\rangle_{H\otimes I}|\mathcal{F}_k] - \frac{\xi nL}{2}\langle Rx^k - Rx^\star x^k - x^\star\rangle$$
$$\geq \langle (\begin{bmatrix} (\frac{\lambda}{nL} - \frac{\xi}{2})I & 0 \\ -\frac{\lambda^2}{n^2}(n-\theta)\mathbf{1}\mathbf{1}^T - \frac{\lambda}{nL}I & \frac{\lambda}{nL}I \end{bmatrix} \otimes I)(\mathbf{B}z^k - \mathbf{B}z^\star)\mathbf{B}z^k - \mathbf{B}z^\star\rangle$$

Setting

$$\bar{M} = \begin{bmatrix} \frac{\lambda}{nL}I & 0 \\ -\frac{\lambda^2}{n^2}(n-\theta)\mathbf{1}\mathbf{1}^T - \frac{\lambda}{nL}I & \frac{\lambda}{nL}I \end{bmatrix}$$

gives

$$\begin{aligned}
&\mathbb{E}[\langle (U_{i^k} \otimes I)(\mathbf{B}z^k - \mathbf{B}z^\star)z^k - z^\star\rangle_{H\otimes I}|\mathcal{F}_k] - \frac{\xi nL}{2}\langle Rx^k - Rx^\star x^k - x^\star\rangle \\
&\quad \geq \langle (\bar{M} \otimes I)(\mathbf{B}z^k - \mathbf{B}z^\star)\mathbf{B}z^k - \mathbf{B}z^\star\rangle \\
&\qquad -\langle (\begin{bmatrix} \frac{\xi}{2}I & 0 \\ 0 & 0 \end{bmatrix} \otimes I)(\mathbf{B}z^k - \mathbf{B}z^\star)\mathbf{B}z^k - \mathbf{B}z^\star\rangle \\
&\quad \geq \|\mathbf{B}z^k - \mathbf{B}z^\star\|^2_{\frac{1}{2}(\bar{M}+\bar{M}^T)\otimes I} - \frac{\xi}{2}\|\mathbf{B}z^k - \mathbf{B}z^\star\|^2 \\
&\quad = \|\mathbf{B}z^k - \mathbf{B}z^\star\|^2_{(M-\frac{\xi}{2}I)\otimes I}
\end{aligned}$$

where $M = \frac{1}{2}(\bar{M} + \bar{M}^T)$ is the matrix in the statement of the lemma. Finally, using this inequality and $0 = (U_{i^k} \otimes I)\mathbf{B}z^\star$ from Lemma 3.3 gives

$$\begin{aligned}
&\mathbb{E}[\|z^{k+1} - z^\star\|^2_{H\otimes I}|\mathcal{F}_k] \\
&\quad = \mathbb{E}[\|\big(z^k - (U_{i^k} \otimes I)\mathbf{B}z^k\big) - \big(z^\star - (U_{i^k} \otimes I)\mathbf{B}z^\star\big)\|^2_{H\otimes I}|\mathcal{F}_k] \\
&\quad = \|z^k - z^\star\|^2_{H\otimes I} + \mathbb{E}[\|(U_{i^k}\| \otimes I)(\mathbf{B}z^k - \mathbf{B}z^\star)\|^2_{H\otimes I}|\mathcal{F}_k] \\
&\qquad -2\mathbb{E}[\langle (U_{i^k} \otimes I)(\mathbf{B}z^k - \mathbf{B}z^\star)z^k - z^\star\rangle_{H\otimes I}|\mathcal{F}_k] \\
&\quad \leq \|z^k - z^\star\|^2_{H\otimes I} + \|\mathbf{B}z^k - \mathbf{B}z^\star\|^2_{\mathbb{E}[U_{i^k}^T H U_{i^k}]\otimes I} \\
&\qquad -\|\mathbf{B}z^k - \mathbf{B}z^\star\|^2_{(2M-\xi I)\otimes I} - \xi nL\langle Rx^k - Rx^\star x^k - x^\star\rangle \\
&\quad = \|z^k - z^\star\|^2_{H\otimes I} - \|\mathbf{B}z^k - \mathbf{B}z^\star\|^2_{(2M-\mathbb{E}[U_{i^k}^T H U_{i^k}]-\xi I)\otimes I} \\
&\qquad -\xi nL\langle Rx^k x^k - x^\star\rangle.
\end{aligned}$$

□

Proof of Lemma 3.6 Take $k \in \mathbb{N}$ and note that

$$(K \otimes I)z^{k+1} = (K \otimes I)(z^k - (U_{i^k} \otimes I)\mathbf{B}z^k) = x^k - (Q_{i^k} \otimes I)\mathbf{B}z^k$$

where $Q_{i^k} = \frac{\lambda}{n}\left[(\theta-1)e_{i^k}^T \; -(\theta-1)e_{i^k}^T\right]$. Furthermore, with $G = \frac{1}{n}\left[\mathbf{1}^T \; 0\right]$, we have $\nabla F(x^k) = (G \otimes I)\mathbf{B}z^k$. From the definition of $z^\star$ we have $0 = (G \otimes I)\mathbf{B}z^\star = (Q_{i^k} \otimes I)\mathbf{B}z^\star$. Using L-smoothness, (6), of F yields

$$\begin{aligned}
&\mathbb{E}[F((K \otimes I)z^{k+1})|\mathcal{F}_k] \\
&\quad = \mathbb{E}[F(x^k - (Q_{i^k} \otimes I)\mathbf{B}z^k)|\mathcal{F}_k] \\
&\quad \leq F(x^k) - \langle \nabla F(x^k)(\mathbb{E}[Q_{i^k}] \otimes I)\mathbf{B}z^k\rangle + \tfrac{L}{2}\mathbb{E}[\|(Q_{i^k} \otimes I)\mathbf{B}z^k\|^2|\mathcal{F}_k] \\
&\quad = F(x^k) - \langle (G \otimes I)\mathbf{B}z^k(\mathbb{E}[Q_{i^k}] \otimes I)\mathbf{B}z^k\rangle + \|\mathbf{B}z^k\|^2_{\frac{L}{2}\mathbb{E}[Q_{i^k}^T Q_{i^k}]\otimes I} \\
&\quad = F(x^k) - \|\mathbf{B}z^k\|^2_{\frac{1}{2}\mathbb{E}[Q_{i^k}^T G + G^T Q_{i^k}]\otimes I} + \|\mathbf{B}z^k\|^2_{\frac{L}{2}\mathbb{E}[Q_{i^k}^T Q_{i^k}]\otimes I} \\
&\quad = F(x^k) - \|\mathbf{B}z^k - \mathbf{B}z^\star\|^2_{\frac{1}{2}S_L \otimes I}
\end{aligned}$$

where $S_L = \mathbb{E}[Q_{i^k}^T G + G^T Q_{i^k} - L Q_{i^k}^T Q_{i^k}]$.

With $D = \begin{bmatrix} 0 & \mathbf{1}^T \end{bmatrix}$ we have $(K \otimes I)z^k = x^k + \frac{\lambda}{n}(D \otimes I)\mathbf{B}z^k$. Using the first-order convexity condition on F and $0 = (D \otimes I)\mathbf{B}z^\star = (G \otimes I)\mathbf{B}z^\star$ yields

$$\begin{aligned}
F((K \otimes I)z^k) &= F(x^k + \frac{\lambda}{n}(D \otimes I)\mathbf{B}z^k) \\
&\geq F(x^k) + \langle \nabla F(x^k)\frac{\lambda}{n}(D \otimes I)\mathbf{B}z^k\rangle \\
&= F(x^k) + \langle (G \otimes I)\mathbf{B}z^k\frac{\lambda}{n}(D \otimes I)\mathbf{B}z^k\rangle \\
&= F(x^k) + \|\mathbf{B}z^k\|^2_{\frac{1}{2}\frac{\lambda}{n}(D^T G + G^T D)\otimes I} \\
&= F(x^k) + \|\mathbf{B}z^k - \mathbf{B}z^\star\|^2_{\frac{1}{2}S_C \otimes I}
\end{aligned} \tag{8}$$

where $S_C = \frac{\lambda}{n}(D^T G + G^T D)$. Combining these two inequalities gives

$$\mathbb{E}[F((K \otimes I)z^{k+1})|\mathcal{F}_k] \leq F((K \otimes I)z^k) - \|\mathbf{B}z^k - \mathbf{B}z^\star\|^2_{\frac{1}{2}S \otimes I}$$

where $S = S_L + S_C$. □

3.2 Convergence theorems

We are now ready to state the main convergence theorems for SVAG. They are stated with the notation from Algorithm 1 but are proved at the end of this section with the help of the reformulation in (7) and the lemmas above.

Theorem 3.7 *For all $i \in \{1, \ldots, n\}$, let $(x^k)_{k\in\mathbb{N}}$ and $(y_i^k)_{k\in\mathbb{N}}$ be the sequences generated by Algorithm 1. If Assumption 3.1 hold and the step-size, $\lambda > 0$, and innovation weight, $\theta \in \mathbb{R}$, satisfy*

$$\frac{1}{L(2 + |n - \theta|)} > \lambda,$$

then $x^k \to x^\star$ and $y_i^k \to R_i x^\star$ almost surely for all $i \in \{1, \dots, n\}$, where $x^\star$ is a solution to (2). For all $i \in \{1, \dots, n\}$, the residuals converge a.s. as

$$\min_{k\in\{0,\dots,t\}} \mathbb{E}[\|R_i x^k - R_i x^\star\|^2] \leq \frac{n}{\lambda(L^{-1} - \lambda c)} \frac{1}{t+1} C_R,$$
$$\min_{k\in\{0,\dots,t\}} \mathbb{E}[\|y_i^k - R_i x^\star\|^2] \leq \frac{n}{\lambda(L^{-1} - \lambda c)} \frac{1}{t+1} C_R$$

where $c = 2 + |n - \theta|$ and

$$C_R = \min_{x\in X^\star} \|x^0 - x\|^2 + \frac{\lambda}{L}\textstyle\sum_{i=1}^n \|y_i^0\| - R_i x^{\star 2} + \lambda^2 (n-\theta)^2 \|\frac{1}{n}\sum_{i=1}^n y_i^0\|^2$$
$$-2\lambda(n-\theta)\langle x^0 - x \frac{1}{n}\textstyle\sum_{i=1}^n y_i^0\rangle$$

for any $x^\star \in X^\star$.

Theorem 3.8 *For all $i \in \{1, \dots, n\}$, let $(x^k)_{k\in\mathbb{N}}$ and $(y_i^k)_{k\in\mathbb{N}}$ be the sequences generated by Algorithm 1. If Assumption 3.2 hold and the step-size, $\lambda > 0$, and innovation weight, $\theta \in [0, n]$, satisfy*

$$\frac{1}{L} \frac{1}{2 + (n-\theta)\frac{\theta-1}{n}\left(\frac{\theta-1}{n} - 1 + \frac{\theta-1}{|\theta-1|}\sqrt{2}\right)} > \lambda,$$

then $x^k \to x^\star$ and $y_i^k \to \nabla f_i(x^\star)$ almost surely, where $x^\star$ is a solution to (2). For all $i \in \{1, \dots, n\}$, the residuals converge a.s. as

$$\min_{k\in\{0,\dots,t\}} \mathbb{E}[\|\nabla f_i(x^k) - \nabla f_i(x^\star)\|^2] \leq \frac{n}{\lambda(L^{-1} - \lambda c)} \frac{1}{t+1} (C_R + C_F),$$
$$\min_{k\in\{0,\dots,t\}} \mathbb{E}[\|y_i^k - \nabla f_i(x^\star)\|^2] \leq \frac{n}{\lambda(L^{-1} - \lambda c)} \frac{1}{t+1} (C_R + C_F),$$
$$\min_{k\in\{0,\dots,t\}} \mathbb{E}[F(x^k) - F(x^\star)] \leq \frac{1}{\lambda(1 - L\lambda c)} \frac{1}{t+1} (C_R + C_F)$$

where

$$c = 2 + (n-\theta)\frac{\theta-1}{n}\left(\frac{\theta-1}{n} - 1 + \frac{\theta-1}{|\theta-1|}\sqrt{2}\right),$$
$$C_R = \min_{x\in X^\star} \|x^0 - x\|^2 + \frac{\lambda}{L}\textstyle\sum_{i=1}^n \|y_i^0\| - R_i x^{\star 2} + \lambda^2 (n-\theta)^2 \|\frac{1}{n}\sum_{i=1}^n y_i^0\|^2$$
$$- 2\lambda(n-\theta)\langle x^0 - x \frac{1}{n}\textstyle\sum_{i=1}^n y_i^0\rangle,$$
$$C_F = 2\lambda(n-\theta)\left(F(x^0 + \frac{\lambda}{n}\textstyle\sum_{i=1}^n y_i^0) - F(x^\star)\right)$$

for any $x^\star \in X^\star$.

Both Theorems 3.7 and 3.8 give the step-size condition $\lambda \in (0, \frac{1}{2L})$ for the SAGA special case, i.e., $\theta = n$. This is the same as the largest upper bound found in the

literature [15] and appears to be tight [31]. Theorem 3.8 also give this step-size condition when $\theta = 1$, i.e., SAG in the optimization case. This bound improves on upper bound of $\frac{1}{16L} \leq \lambda$ presented in [40].

In the cocoercive operator setting with $\theta \neq n$, Theorem 3.7 gives a step-size condition that scales with n^{-1}. This step-size scaling is significantly worse compared to the gradient case in Theorem 3.8 in which the step-size's dependence on n is $\mathcal{O}(1)$ for all θ. This difference is indeed real and not an artifact of the analysis since we in Section 4 present a problem for which the cocoercivity result appears to be tight. A consequence of this unfavorable step-size scaling in the operator setting is slow convergence. There is therefore little reason to use anything else than $\theta = n$ in SVAG when R_i is not a gradient of a smooth function for all $i \in \{1, \dots, n\}$.

The rates of Theorem 3.7 and 3.8 are of $\mathcal{O}(\frac{1}{t+1})$ type with two sets of multiplicative factors. One factor which only depend on the algorithm parameters, $\frac{n}{\lambda(L^{-1}-\lambda c)}$, and one set which depend on how the algorithm initialization relates to the solution set, C_R and $C_R + C_F$. The initialization dependent factors also depend on the algorithm parameters, but, since knowing the exact dependency requires knowing the solution set, we will not attempt to tune the parameters to decrease this factor. Only considering the first factor, the rate becomes better if c is decreased and, since c is independent of λ, the best choice of step-size is $\lambda = (2Lc)^{-1}$. This means that $\lambda = (4L)^{-1}$ and $\theta = n$ are the best parameter choices in the cocoercive operator setting. In the optimization case the best step-size is also $\lambda = (4L)^{-1}$ but the innovation weight can be selected as either $\theta = n$ or $\theta = 1$.

However, in the optimization case we do not believe that these theoretical rates reflects real world performance and parameter choices based on them might therefore not perform particularly well. We base this belief on our experience with numerical experiments. For $\theta \neq n$ and $\theta \neq 1$, we have not found any optimization problem where the step-size condition in Theorem 3.8 appears to be tight. Also, using $\lambda = (2Lc)^{-1}$ as suggested by Theorem 3.8 can in some cases lead to impractically small step-sizes. For instance, if $\lambda = (2Lc)^{-1}$ was used in the experiments in Section 4, a couple of the experiments would have step-sizes over 1000 times smaller than the ones used now. One can of course not disprove a worst case analysis with experiments but we still feel they indicate a conservative analysis, even though the analysis improves on the previous best results.

Proof of Theorem 3.7 Apply Lemma 3.5 with $\xi = 0$, the iterates given by (7) then satisfy the following for all $z^\star \in Z^\star$,

$$\begin{aligned} &\mathbb{E}[\|z^{k+1} - z^\star\|^2_{H\otimes I} | \mathcal{F}_k] \\ &\quad \leq \|z^k - z^\star\|^2_{H\otimes I} - \|\mathbf{B}z^k - \mathbf{B}z^\star\|^2_{(2M - \mathbb{E}[U_{i^k}^T H U_{i^k}])\otimes I}. \end{aligned} \tag{9}$$

Assuming $H \succ 0$ and $2M - \mathbb{E}[U_{i^k}^T H U_{i^k}] \succ 0$, Proposition 2.1 can be applied. We will later prove that this assumption indeed does hold. Proposition 2.1 gives a.s. summability of $\|\mathbf{B}z^k - \mathbf{B}z^\star\|^2_{(2M-\mathbb{E}[U_{i^k}^T H U_{i^k}])\otimes I}$ and hence will $\mathbf{B}z^k \to \mathbf{B}z^\star$ almost surely. Lemma 3.4 then gives that all cluster points of $(z^k)_{k\in\mathbb{N}}$ are in $Z^\star$ almost surely. Finally, since Proposition 2.1 ensures the a.s. convergence of $\|z^k - z^\star\|^2_{H\otimes I}$ and since

$\mathbb{R}^{N(n+1)}$ with the inner product $\langle (H \otimes I)\cdot, \cdot \rangle$ is a finite dimensional Hilbert space, Proposition 2.2 gives the almost sure convergence of $z^k \to z^\star \in Z^\star$.

There always exists a λ such that $2M - \mathbb{E}[U_{i^k}^T H U_{i^k}]$ and H are positive definite. First we show that $H \succ 0$ always holds for $\lambda > 0$. Taking the Schur complement of 1 in H gives

$$\frac{\lambda}{L} I + \frac{\lambda^2}{n^2}(n-\theta)^2 \mathbf{1}\mathbf{1}^T - \frac{\lambda^2}{n^2}(n-\theta)^2 \mathbf{1}\mathbf{1}^T = \frac{\lambda}{L} I \succ 0.$$

Hence is $H \succ 0$ since the Schur complement is positive definite.

We now show $2M - \mathbb{E}[U_{i^k}^T H U_{i^k}] \succ 0$. Straightforward algebra, see the supplementary material, yields

$$\begin{aligned} 2M - \mathbb{E}[U_{i^k}^T H U_{i^k}] \;=\; & \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \otimes \frac{\lambda}{nL} I - \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \otimes \frac{\lambda^2}{n} I + \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \otimes \frac{\lambda^2}{n}(I - \frac{1}{n}\mathbf{1}\mathbf{1}^T) \\ & - \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \otimes (n-\theta)\frac{\lambda^2}{n^2}\mathbf{1}\mathbf{1}^T + \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \otimes \frac{\lambda^2}{n^2}\mathbf{1}\mathbf{1}^T. \end{aligned}$$

Positive definiteness of this matrix is established by ensuring positivity of the smallest eigenvalue $\sigma_{\min}$. The smallest eigenvalue $\sigma_{\min}$ is greater than the sum of the smallest eigenvalue of each term. For the eigenvalues of the Kronecker products, see (4). This gives that

$$\sigma_{\min} \geq \frac{\lambda}{nL} - \frac{\lambda^2}{n} - \frac{\lambda^2}{n} - \frac{\lambda^2}{n}|n-\theta| + 0 = \frac{\lambda}{n}\big(L^{-1} - \lambda(2 + |n-\theta|)\big).$$

Since $\lambda > 0$ by assumption, if

$$\frac{1}{L(2+|n-\theta|)} > \lambda.$$

we have $\sigma_{\min} > 0$ and $2M - \mathbb{E}[U_{i^k}^T H U_{i^k}]$ is positive definite.

Rates are gotten by taking the total expectation of (9) and adding together the inequalities from $k = 0$ to $k = t$, yielding

$$\begin{aligned} \|z^0 - z^\star\|_{H\otimes I}^2 &= \mathbb{E}[\|z^0 - z^\star\|_{H\otimes I}^2] - \mathbb{E}[\|z^{t+1} - z^\star\|_{H\otimes I}^2] \\ &\geq \textstyle\sum_{k=0}^t \mathbb{E}\big[\|\mathbf{B}z^k - \mathbf{B}z^\star\|^2_{(2M - \mathbb{E}[U_{i^k}^T H U_{i^k}])\otimes I}\big] \\ &\geq \textstyle\sum_{k=0}^t \sigma_{\min}\mathbb{E}[\|\mathbf{B}z^k - \mathbf{B}z^\star\|^2] \\ &\geq \sigma_{\min}(t+1) \min_{k\in\{0,\dots,t\}} \mathbb{E}[\|\mathbf{B}z^k - \mathbf{B}z^\star\|^2]. \end{aligned}$$

Putting in the lower bound on $\sigma_{\min}$ and rearranging yield

$$\min_{k\in\{0,\dots,t\}} \mathbb{E}[\|\mathbf{B}z^k - \mathbf{B}z^\star\|^2] \leq \frac{n}{\lambda(L^{-1} - \lambda(2+|n-\theta|))(t+1)} \|z^0 - z^\star\|_{H\otimes I}^2.$$

From the definition of H in Lemma 3.5 we have

$$\begin{aligned} \|z^0 - z^\star\|_{H\otimes I}^2 = \|x^0 - x^\star\|^2 + \tfrac{\lambda}{L}\textstyle\sum_{i=1}^n \|y_i^0 - R_i x^\star\|^2 + \lambda^2(n-\theta)^2 \|\frac{1}{n}\sum_{i=1}^n y_i^0\|^2 \\ -2\lambda(n-\theta)\langle x^0 - x^\star, \frac{1}{n}\textstyle\sum_{i=1}^n y_i^0\rangle \end{aligned}$$

where $z^\star = (x^\star, R_1 x^\star, \dots, R_n x^\star)$. Since this hold for any $z^\star \in Z^\star$ and hence any $x^\star \in X^\star$, the results of theorems follows by minimizing the RHS over $x^\star \in X^\star$. Note, since $R_i x^\star$ constant for all $x^\star \in X^\star$, the objective is convex and, since $X^\star$ is closed and convex, the minimum is then attained. □

Proof of Theorem 3.8 Combining Lemma 3.5 and 3.6 yield

$$\begin{aligned}\mathbb{E} \quad & [\|z^{k+1} - z^\star\|^2_{H\otimes I} + 2\lambda(n-\theta)(F((K\otimes I)z^{k+1}) - F(x^\star))|\mathcal{F}_k] \\ \leq \quad & \|z^k - z^\star\|^2_{H\otimes I} + 2\lambda(n-\theta)(F((K\otimes I)z^k) - F(x^\star)) \\ - \quad & \|\mathbf{B}z^k - \mathbf{B}z^\star\|^2_{(2M - \mathbb{E}[U_{i^k}^T H U_{i^k}] + \lambda(n-\theta)S - \xi I)\otimes I} - \xi n L\langle \nabla F(x^k)x^k - x^\star\rangle\end{aligned}$$

which holds for all $k \in \mathbb{N}$, $\xi \in [0, \frac{2\lambda}{nL}]$, and $z^\star \in Z^\star$. Since $H \succ 0$ for $\lambda > 0$, see the proof of Theorem 3.7, the first term is non-negative while the second term is non-negative if $\theta \leq n$. From cocoercivity of ∇F, the last term is non-positive and we assume, for now, that there exists $\lambda > 0$ and $\frac{2\lambda}{nL} \geq \xi > 0$ such that $2M - \mathbb{E}[U_{i^k}^T H U_{i^k}] + \lambda(n-\theta)S - \xi I \succ 0$, making the third term non-positive.

Applying Proposition 2.1 gives the a.s. summability of

$$\|\mathbf{B}z^k - \mathbf{B}z^\star\|^2_{(2M - \mathbb{E}[U_{i^k}^T H U_{i^k}] + \lambda(n-\theta)S - \xi I)\otimes I} + \xi n L\langle \nabla F(x^k) - \nabla F(x^\star)x^k - x^\star\rangle.$$

Since both term are positive, both terms are a.s. summable. From the first term we have the a.s. convergence of $\mathbf{B}z^k \to \mathbf{B}z^\star$ and Lemma 3.4 then gives that all cluster points of $(z^k)_{k\in\mathbb{N}}$ are almost surely in $Z^\star$. For the second term we note that by convexity we have

$$\langle \nabla F(x^k) - \nabla F(x^\star)x^k - x^\star\rangle \geq F(x^k) - F(x^\star) \geq 0$$

and $F(x^k) - F(x^\star)$ then is summable a.s. since $\xi nL > 0$. Using smoothness of F, (6) and the notation from (8) gives

$$\begin{aligned}F(x^\star) \quad & \leq \quad F((K\otimes I)z^k) \\ & = \quad F(x^k + \frac{\lambda}{n}(D\otimes I)\mathbf{B}z^k) \\ & \leq \quad F(x^k) + \langle (G\otimes I)\mathbf{B}z^k \frac{\lambda}{n}(D\otimes I)\mathbf{B}z^k\rangle + \frac{L}{2}\|\frac{\lambda}{n}(D\otimes I)\mathbf{B}z^k\|^2 \\ & \leq \quad F(x^k) + \|(G\otimes I)\mathbf{B}z^k\|\|\frac{\lambda}{n}(D\otimes I)\mathbf{B}z^k\| + \frac{L}{2}\|\frac{\lambda}{n}(D\otimes I)\mathbf{B}z^k\|^2 \\ & \to \quad F(x^\star) \text{ a.s.}\end{aligned}$$

since $(G\otimes I)\mathbf{B}z^k \to (G\otimes I)\mathbf{B}z^\star = 0$ and $(D\otimes I)\mathbf{B}z^k \to (D\otimes I)\mathbf{B}z^\star = 0$ almost surely. Therefore we have the a.s. convergence of $F((K\otimes I)z^k) - F(x^\star) \to 0$.

From Proposition 2.1 we can also conclude that $\|z^k - z^\star\|^2_{H\otimes I} + 2\lambda(n-\theta)(F((K\otimes I)z^k) - F(x^\star))$ a.s. converge to a non-negative random variable. Since $F((K\otimes I)z^k) - F(x^\star) \to 0$ a.s. we have that $\|z^k - z^\star\|^2_{H\otimes I}$ also must a.s. converge to a non-negative random variable. Proposition 2.2 then give the almost sure convergence of $(z^k)_{k\in\mathbb{N}}$ to $Z^\star$.

We now show that there exists $\lambda > 0$ and $\xi > 0$ such that

$$
\begin{aligned}
&2M - \mathbb{E}[U_{i^k}^T H U_{i^k}] + \lambda(n-\theta)S - \xi I \\
&\quad = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \otimes \frac{\lambda}{nL} I - \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \otimes \frac{\lambda^2}{n} I + \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \otimes \frac{\lambda^2}{n}(I - \frac{1}{n}\mathbf{1}\mathbf{1}^T) + \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \otimes \frac{\lambda^2}{n^2}\mathbf{1}\mathbf{1}^T \\
&\qquad + \begin{bmatrix} 2 & -1 \\ -1 & 0 \end{bmatrix} \otimes (n-\theta)(\theta-1)\frac{\lambda^2}{n^3}\mathbf{1}\mathbf{1}^T - \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \otimes (n-\theta)(\theta-1)^2 \frac{L\lambda^3}{n^3} I \\
&\qquad - \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \otimes \xi I \succ 0.
\end{aligned}
$$

We show positive definiteness by ensuring that the smallest eigenvalue is positive. The smallest eigenvalue $\sigma_{\min}$ is greater than the sum of the smallest eigenvalues of each term,

$$
\begin{aligned}
\sigma_{\min} \geq\ & \frac{\lambda}{nL} - \frac{\lambda^2}{n} - \frac{\lambda^2}{n} + 0 + (1 - \frac{\theta-1}{|\theta-1|}\sqrt{2})(n-\theta)(\theta-1)\frac{\lambda^2}{n^2} \\
& -2(n-\theta)(\theta-1)^2\frac{L\lambda^3}{n^3} - \xi.
\end{aligned}
$$

Assuming $\lambda \leq \frac{1}{2L}$ yields the following lower bound on the smallest eigenvalue

$$
\begin{aligned}
\sigma_{\min} &\geq \frac{\lambda}{nL} - \frac{2\lambda^2}{n} + (1 - \frac{\theta-1}{|\theta-1|}\sqrt{2})(n-\theta)(\theta-1)\frac{\lambda^2}{n^2} - (n-\theta)(\theta-1)^2\frac{\lambda^2}{n^3} - \xi \\
&= \frac{\lambda}{n}\big(L^{-1} - \lambda\big(2 + (n-\theta)\frac{\theta-1}{n}\big(\frac{\theta-1}{n} - 1 + \frac{\theta-1}{|\theta-1|}\sqrt{2}\big)\big)\big) - \xi.
\end{aligned}
$$

Selecting

$$
\xi = \frac{\lambda}{2n}\big(L^{-1} - \lambda\big(2 + (n-\theta)\frac{\theta-1}{n}\big(\frac{\theta-1}{n} - 1 + \frac{\theta-1}{|\theta-1|}\sqrt{2}\big)\big)\big),
$$

which satisfy the assumption $\frac{2\lambda}{nL} \geq \xi > 0$, yield $\sigma_{\min} \geq \xi$. Since $\lambda > 0$ by assumption, if

$$
\frac{1}{L}\frac{1}{2 + (n-\theta)\frac{\theta-1}{n}(\frac{\theta-1}{n} - 1 + \frac{\theta-1}{|\theta-1|}\sqrt{2})} > \lambda
$$

we have that $\sigma_{\min} \geq \xi > 0$ and hence that the examined matrix is positive definite. Furthermore, if λ satisfies the above inequality it also satisfies the assumption $\lambda \leq \frac{1}{2L}$.

Rates are gotten in the same way as for Theorem 3.7, the total expectation is taken of the Lyapunov inequality at the beginning of the proof and the inequalities are summed from $k = 0$ to $k = t$.

$$\begin{aligned}
&\|z^0 - z^\star\|^2_{H\otimes I} + 2\lambda(n-\theta)(F((K\otimes I)z^0) - F(x^\star)) \\
&\quad \geq \textstyle\sum_{k=0}^{t}\big(\sigma_{\min}\mathbb{E}[\|\mathbf{B}z^k - \mathbf{B}z^\star\|^2] + \mathbb{E}[\sigma_{\min} nL\langle \nabla F(x^k)x^k - x^\star\rangle]\big) \\
&\quad \geq \sigma_{\min}(t+1)\min_{k\in\{1,\dots,t\}}\big(\mathbb{E}[\|\mathbf{B}z^k - \mathbf{B}z^\star\|^2] + \mathbb{E}[nL\langle \nabla F(x^k)x^k - x^\star\rangle]\big) \\
&\quad \geq \sigma_{\min}(t+1)\min_{k\in\{1,\dots,t\}}\big(\mathbb{E}[\|\mathbf{B}z^k - \mathbf{B}z^\star\|^2] + nL\mathbb{E}[F(x^k) - F(x^\star)]\big).
\end{aligned}$$

Inserting the lower bound on $\sigma_{\min}$, rearranging and minimizing over $x^\star \in X^\star$ yield the results of the theorem. □

4 Numerical experiments

A number of experiments, outlined below, were performed to verify the tightness of the theory in the cocoercive operator case and examine the effect of bias in the cocoercive gradient case. The experiments were implemented in `Julia` [3] and, together with several other VR-SG methods, can be found at https://github.com/mvmorin/VarianceReducedSG.jl.

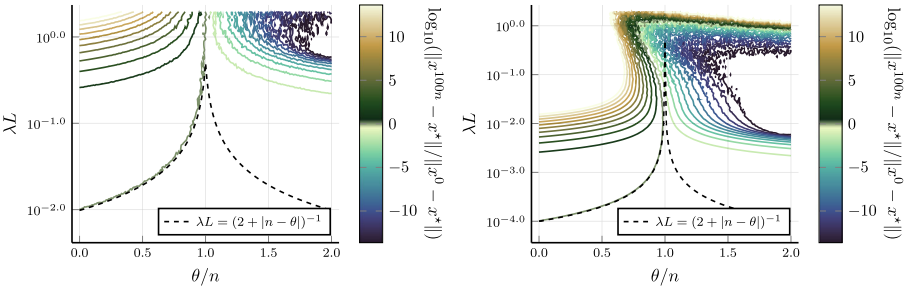
4.1 Cocoercive operators case

In order for the difference between cocoercive operators and cocoercive gradients to not be an artifact of our analysis, the results in the operator case can not be overly conservative. We therefore construct a cocoercive operator problem for which the results appear to be tight, thereby verifying the difference. Consider problem (2) where the operator $R_i : \mathbb{R}^2 \to \mathbb{R}^2$ is an averaged rotation

$$R_i = \frac{1}{2}\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix} + \frac{1}{2}\begin{bmatrix}\cos\tau & -\sin\tau\\ \sin\tau & \cos\tau\end{bmatrix}$$

for all $i \in \{1,\dots,n\}$ and some $\tau \in [0, 2\pi)$. The operators are 1-cocoercive and the zero vector is the only solution to (2) if $\tau \neq \pi$. The step-size condition from Theorem 3.7 appears to be tight for $\theta \in [0, n]$ when the angle of rotation τ approaches π. We therefore let $\tau = \frac{179}{180}\pi$ and solve the problem with different configurations of step-size λ and innovation weight θ.

Figure 1 displays the relative distance to the solution after $100n$ iterations of SVAG together with the upper bound on the step-size. When $\theta \in [0, n]$ and λ exceeds the upper bound, the distance to the solution increases for both $n = 100$ and $n = 10000$, i.e., the method does not converge. Hence, for $\theta \in [0, n]$, the step-size bound in Theorem 3.7 appears to be tight. However, it is noteworthy that for this particular problem it seems beneficial to exceed the step-size bound when $\theta > n$.

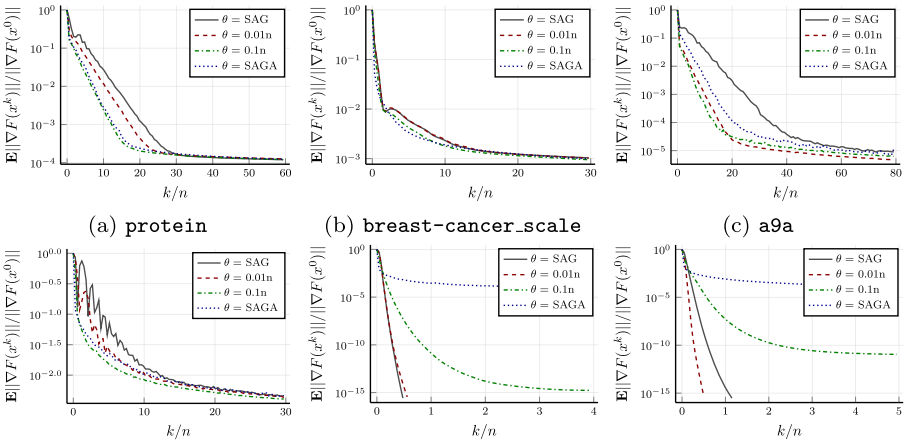


(a) $n = 100$ (b) $n = 10000$

Fig. 1 Root-finding of averaged rotations: Relative distance to the solution after $100n$ iterations of SVAG together the step-size upper bound, $\lambda L < (2 + |n - \theta|)^{-1}$. Note how well the 0th level, i.e., the boundary between convergence and divergence, follow the upper bound on the step-size

4.2 Cocoercive gradients case

Since, as we stated in Section 3.2, we do not believe that the theoretical rates are particularly tight in the optimization case, we examine the effects of the bias numerically. These experiments can of course not be exhaustive and we choose to focus on only the bias parameter θ and therefore perform all experiments with the same step-size. This also demonstrate why we believe the analysis to be conservative since the chosen step-size in some cases are a 1000 times larger than upper bound from Theorem 3.8. Convergence with this large of a step-size have also been seen elsewhere with both [40] and [17] disregarding their own the theoretical step-size conditions.



(a) `protein` (b) `breast-cancer_scale` (c) `a9a`

(d) `gisette_scale` (e) `mushrooms` (f) `mnist.scale`

Fig. 2 Logistic regression: Expected gradient norm for each iteration. The expected value is estimated with the sample average of 100 runs. A step-size of $\lambda = \frac{1}{2L}$ was used in all cases

The experiments are done by performing a rough parameter sweep over the innovation weight θ on two different binary classification problems and we will look for patterns in how the convergence is affected. '
The first problem is logistic regression,

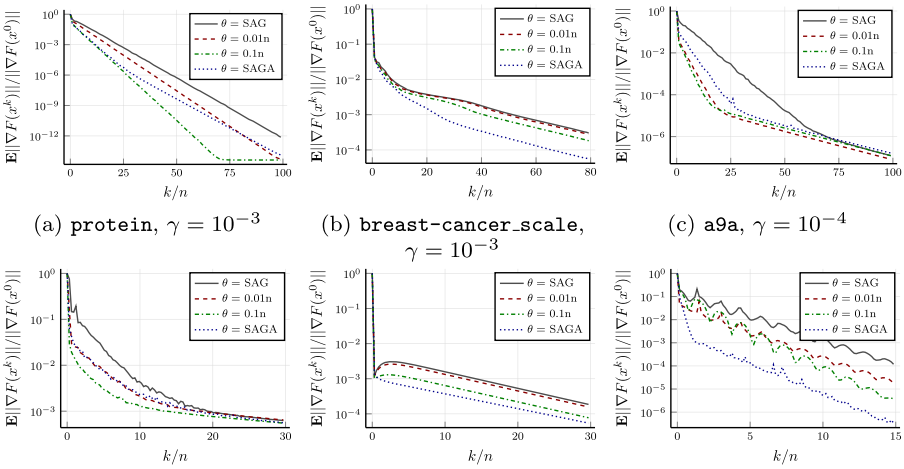
$$\min_{x\in\mathbb{R}^N} \frac{1}{n}\sum_{i=1}^{n} \log(1+e^{-y_i a_i^T x}).$$

The second is SVM with a square hinge loss,

$$\min_{x\in\mathbb{R}^N} \frac{1}{n}\sum_{i=1}^{n} \left(\max(0,1-y_i a_i^T x)^2 + \frac{\gamma}{2}\|x\|^2\right)$$

where $\gamma > 0$ is a regularization parameter. In both problems are $y_i \in \{-1,1\}$ the label and $a_i \in \mathbb{R}^N$ the features of the ith training data point. Note, although not initially obvious, $\max(0,\cdot)^2$ is convex and differentiable with Lipschitz continuous derivative and the second problem is therefore indeed smooth. The logistic regression problem does not necessarily have a unique solution and the distance to the solution set is therefore hard to estimate. For this reason, we examine the convergence of $\|\nabla F(x^k)\| \to 0$ instead of the distance to the solution set.

The datasets for both these classification problems are taken from the `LibSVM` [6] collection of datasets. The number of examples in the datasets varies between $n = 683$ and $n = 60,000$ while the number of features is between $N = 10$ and $N = 5,000$. Two of the datasets, `mnist.scale` and `protein`, consist of more than 2 classes. These are converted to binary classification problems by grouping the different classes into two groups. For the digit classification dataset `mnist.scale`, the digits are divided into the groups 0–4 and 5–9. For the `protein` data set, the



(a) `protein`, $\gamma = 10^{-3}$ (b) `breast-cancer_scale`, $\gamma = 10^{-3}$ (c) `a9a`, $\gamma = 10^{-4}$

(d) `gisette_scale`, $\gamma = 10^{-1}$ (e) `mushrooms`, $\gamma = 10^{-3}$ (f) `mnist.scale`, $\gamma = 10^{-1}$

Fig. 3 Square hinge loss SVM: Expected gradient norm for each iteration. The expected value is estimated with the sample average of 100 runs. A step-size of $\lambda = \frac{1}{2L}$ was used in all cases

classes are grouped as 0 and 1–2. The results of solving the classification problems above can be found in Figs. 2 and 3.

From Figs. 2 and 3 it appears like the biggest difference between the innovation weights are in the early stages of the convergence. Most innovation weight choices appear to eventually converge with the same rate. In the cases where this does not happen, the fastest converging choice of innovation weight actually reaches machine precision. It is therefore not possible to say whether these cases would eventually reach the same rate as well. Since none of the choices of θ appears to consistently be at a significant disadvantage, even though the step-size used exceeds the upper bound in Theorem 3.8 when $\theta = 0.1n$ and $\theta = 0.01n$, we conjecture that the asymptotic rates for a given step-size is independent of θ.

The initial phase can clearly have a large impact on the convergence and it can therefore still be beneficial to tuning the bias. However, comparing the different choices of innovation weight yields no clear conclusion since no single choice of innovation weight consistently outperforms another. In most cases do the lower bias choices—$\theta = n$ (SAGA) or $\theta = 0.1n$—seem perform best but, when they do not, the high bias choices—$\theta = 1$ (SAG) and $\theta = 0.01n$—perform significantly better. Another observation is that lowering θ increases any oscillations. We speculate that it is due to the increased inertia and we also believe that this inertia is what allows the lower innovation weights to sometimes perform better.

5 Conclusion

We presented SVAG, a variance-reduced stochastic gradient method with adjustable bias and with SAG and SAGA as special cases. It was analyzed in two scenarios, one being the minimization of a finite sum of functions with cocoercive gradients and the other being finding a root of a finite sum of cocoercive operators. The analysis improves on the previously best known analyses in both settings and, more significantly, the two different scenarios gave different convergence conditions for the step-size. In the cocoercive operator setting a much more restrictive condition was found and it was verified numerically. This difference is not present in ordinary gradient descent and can therefore easily be overlooked, however, these results suggest that is inadvisable in the variance-reduced stochastic gradient setting.

The theoretical results in the minimization case was further examined with numerical experiments. Several choices of bias were examined but we did not find the same dependence on the bias that the theory suggests. In fact, the asymptotic convergence behavior was similar for the different choices of bias, indicating that further improvements of the theory is still needed. The bias mainly impacted the early stages of the convergence and in a couple of cases this impact was significant. There might therefore still be benefits to tuning the bias to the particular problem but further work is needed to efficiently do so.

Funding Open access funding provided by Lund University. This work is funded by the Swedish Research Council via grant number 2016-04646.

Availability of data and material The datasets used can be found at [6].

Code availability Implementations of the algorithms used can be found at https://github.com/mvmorin/VarianceReducedSG.jl.

Declarations

Conflict of interest The authors declare no competing interests.

References

1. Allen-Zhu, Z.: Katyusha: the first direct acceleration of stochastic gradient methods. In: Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2017, pp. 1200–1205. ACM, New York, NY, USA (2017). https://doi.org/10.1145/3055399.3055448
2. Bauschke, H.H., Combettes, P.L.: Convex analysis and monotone operator theory in Hilbert spaces, second edn. CMS Books in Mathematics. Springer International Publishing. http://www.springer.com/gp/book/9783319483108 (2017)
3. Bezanson, J., Edelman, A., Karpinski, S., Shah, V.B.: Julia: a fresh approach to numerical computing. SIAM Rev. **59**(1), 65–98 (2017). https://doi.org/10.1137/141000671
4. Briceño-Arias, L.M., Davis, D.: Forward-backward-half forward algorithm for solving monotone inclusions. SIAM J. Optim. **28**(4), 2839–2871 (2018). https://doi.org/10.1137/17M1120099
5. Carmon, Y., Jin, Y., Sidford, A., Tian, K.: Variance reduction for matrix games. Adv. Neural Inf. Process Syst **32**, 11381–11392 (2019). https://proceedings.neurips.cc/paper/2019/hash/6c442e0e996fa84f344a14927703a8c1-Abstract.html
6. Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. ACM Trans. Intell. Syst. Technol. (TIST) **2**(3), 27:1–27:27 (2011). https://doi.org/10.1145/1961189.1961199. Software available at http://www.csie.ntu.edu.tw/cjlin/libsvm
7. Chavdarova, T., Gidel, G., Fleuret, F., Lacoste-Julien, S.: Reducing noise in GAN training with variance reduced extragradient. Adv. Neural Inf. Process. Syst. **32**, 393–403 (2019). https://proceedings.neurips.cc/paper/2019/hash/58a2fc6ed39fd083f55d4182bf88826d-Abstract.html
8. Combettes, P.L.: Solving monotone inclusions via compositions of nonexpansive averaged operators. Optimization **53**(5-6), 475–504 (2004). https://doi.org/10.1080/02331930412331327157
9. Combettes, P.L., Eckstein, J.: Asynchronous block-iterative primal-dual decomposition methods for monotone inclusions. Math. Program. **168**(1), 645–672 (2018). https://doi.org/10.1007/s10107-016-1044-0
10. Combettes, P.L., Glaudin, L.E.: Solving composite fixed point problems with block updates. Adv. Nonlinear Anal. **10**(1), 1154–1177 (2021). https://doi.org/10.1515/anona-2020-0173
11. Combettes, P.L., Pesquet, J.C.: Primal-dual splitting algorithm for solving inclusions with mixtures of composite, lipschitzian, and Parallel-Sum type monotone operators. Set-Valued Var. Anal. **20**(2), 307–330 (2012). https://doi.org/10.1007/s11228-011-0191-y
12. Combettes, P.L., Pesquet, J.C.: Stochastic quasi-fejér block-coordinate fixed point iterations with random sweeping. SIAM J. Optim. **25**(2), 1221–1248 (2015). https://doi.org/10.1137/140971233

13. Combettes, P.L., Woodstock, Z.C.: A fixed point framework for recovering signals from nonlinear transformations. In: 2020 28Th European Signal Processing Conference (EUSIPCO), pp. 2120–2124 (2021). https://doi.org/10.23919/Eusipco47968.2020.9287736
14. Davis, D., Yin, W.: A three-operator splitting scheme and its optimization applications. Set-Valued Var. Anal. **25**(4), 829–858 (2017). https://doi.org/10.1007/s11228-017-0421-z
15. Defazio, A., Bach, F., Lacoste-Julien, S.: SAGA: a fast incremental gradient method with support for non-strongly convex composite objectives. In: Advances in Neural Information Processing Systems 27, pp. 1646–1654. Inc, Curran Associates (2014)
16. Defazio, A., Domke, J.: Caetano: Finito: A faster, permutable incremental gradient method for big data problems. In: International Conference on Machine Learning, pp. 1125–1133 (2014)
17. Driggs, D., Liang, J., Schönlieb, C.B.: On biased stochastic gradient estimation. arXiv:1906.01133v2 [math] (2020)
18. Giselsson, P.: Nonlinear forward-backward splitting with projection correction. SIAM J. Optim., pp 2199–2226. https://doi.org/10.1137/20M1345062 (2021)
19. Goldstein, A.A.: Convex programming in Hilbert space. Bull. Am. Math. Soc. **70**(5), 709–711 (1964). https://doi.org/10.1090/S0002-9904-1964-11178-2
20. Gower, R.M., Richtárik, P., Bach, F.: Stochastic quasi-gradient methods: variance reduction via Jacobian sketching. Math. Program. **188**(1), 135–192 (2021). https://doi.org/10.1007/s10107-020-01506-0
21. Hanzely, F., Mishchenko, K., Richtárik, P.: SEGA: Variance reduction via gradient sketching. In: Advances in Neural Information Processing Systems 31, pp. 2082–2093. Curran Associates, Inc. (2018)
22. Hofmann, T., Lucchi, A., Lacoste-Julien, S., McWilliams, B.: Variance reduced stochastic gradient descent with neighbors. In: Advances in Neural Information Processing Systems 28, pp. 2305–2313. Inc, Curran Associates (2015)
23. Johnson, R., Zhang, T.: Accelerating stochastic gradient descent using predictive variance reduction. In: Advances in Neural Information Processing Systems 26, pp. 315–323. Curran Associates, Inc. (2013)
24. Konečný, J., Richtárik, P.: Semi-stochastic gradient descent methods. Frontiers in applied mathematics and statistics, 3. https://doi.org/10.3389/fams.2017.00009 (2017)
25. Kovalev, D., Horváth, S., Richtárik, P.: Don't jump through hoops and remove those loops: SVRG and Katyusha are better without the outer loop. In: Proceedings of the 31st International Conference on Algorithmic Learning Theory, pp. 451–467. PMLR. https://proceedings.mlr.press/v117/kovalev20a.html (2020)
26. Latafat, P., Patrinos, P.: Primal-dual proximal algorithms for structured convex optimization: A unifying framework. In: Large-Scale and Distributed Optimization, Lecture Notes in Mathematics, pp. 97–120. Springer International Publishing (2018). https://doi.org/10.1007/978-3-319-97478-1_5
27. Le Roux, N., Schmidt, M., Bach, F.: A stochastic gradient method with an exponential convergence rate for finite training sets. In: Advances in Neural Information Processing Systems 25, pp. 2663–2671. Inc, Curran Associates (2012)
28. Levitin, E.S., Polyak, B.T.: Constrained minimization methods. USSR Comput. math. math. phys. **6**(5), 1–50 (1966)
29. Lions, P.L., Mercier, B.: Splitting algorithms for the sum of two nonlinear operators. SIAM J. Numer. Anal. **16**(6), 964–979 (1979). https://doi.org/10.1137/0716071
30. Mairal, J.: Optimization with first-order surrogate functions. In: Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28, ICML'13, pp. III–783–III–791. JMLR.org, Atlanta, GA, USA (2013)
31. Morin, M., Giselsson, P.: Sampling and update frequencies in proximal variance reduced stochastic gradient methods. arXiv:2002.05545 [cs, math] (2020)
32. Nesterov, Y.: Introductory lectures on convex optimization: A basic course. Applied Optimization. Springer US. http://www.springer.com/us/book/9781402075537 (2004)
33. Nguyen, L.M., Liu, J., Scheinberg, K., Takáč, M.: SARAH: A novel method for machine learning problems using stochastic recursive gradient. In: Proceedings of the 34th International Conference on Machine Learning - Volume 70, ICML'17, pp. 2613–2621. JMLR.org, Sydney, NSW, Australia (2017)
34. Palaniappan, B., Bach, F.: Stochastic variance reduction methods for saddle-point problems. In: Advances in Neural Information Processing Systems 29, pp. 1416–1424. Curran Associates, Inc. (2016)

35. Qian, X., Qu, Z., Richtárik, P.: SAGA with arbitrary sampling. In: Proceedings of the 36th International Conference on Machine Learning, pp. 5190–5199. PMLR. (2019). https://proceedings.mlr.press/v97/qian19a.html
36. Raguet, H., Fadili, J., Peyré, G.: A generalized forward-backward splitting. SIAM J. Imaging Sci. **6**(3), 1199–1226 (2013). https://doi.org/10.1137/120872802
37. Robbins, H., Siegmund, D.: A convergence theorem for non negative almost supermartingales and some applications. In: Optimizing Methods in Statistics, pp. 233–257. Academic Press. 10.1016/B978-0-12-604550-5.50015-8 (1971)
38. Rockafellar, R.T.: Monotone operators and the proximal point algorithm. SIAM J. Control. Optim. **14**(5), 877–898 (1976). https://doi.org/10.1137/0314056
39. Schmidt, M., Babanezhad, R., Ahmed, M., Defazio, A., Clifton, A., Sarkar, A.: Non-uniform stochastic average gradient method for training conditional random fields. In: Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics, Proceedings of Machine Learning Research, vol. 38, pp. 819–828. PMLR. (2015). http://proceedings.mlr.press/v38/schmidt15.html
40. Schmidt, M., Le Roux, N., Bach, F.: Minimizing finite sums with the stochastic average gradient. Math. Program. **162**(1), 83–112 (2017). https://doi.org/10.1007/s10107-016-1030-6
41. Shalev-Shwartz, S., Zhang, T.: Stochastic dual coordinate ascent methods for regularized loss minimization. Journal of Machine Learning Research 14(Feb), 567–599. http://www.jmlr.org/papers/v14/shalev-shwartz13a.html (2013)
42. Shi, Z., Zhang, X., Yu, Y.: Bregman divergence for stochastic variance reduction: Saddle-point and adversarial prediction. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17, pp. 6033–6043. Curran Associates Inc., Red Hook, NY, USA (2017)
43. Tang, M., Qiao, L., Huang, Z., Liu, X., Peng, Y., Liu, X.: Accelerating SGD using flexible variance reduction on large-scale datasets Neural Computing and Applications. https://doi.org/10.1007/s00521-019-04315-5 (2019)
44. Tseng, P.: A modified forward-backward splitting method for maximal monotone mappings. SIAM J. Control. Optim. **38**(2), 431–446 (2000). https://doi.org/10.1137/S0363012998338806
45. Xiao, L., Zhang, T.: A proximal stochastic gradient method with progressive variance reduction. SIAM J. Optim. **24**(4), 2057–2075 (2014). https://doi.org/10.1137/140961791
46. Zhang, X., Haskell, W.B., Ye, Z.: A unifying framework for variance reduction algorithms for finding zeroes of monotone operators. arXiv:1906.09437v2 [cs, stat] (2021)
47. Zhou, K., Ding, Q., Shang, F., Cheng, J., Li, D., Luo, Z.Q.: Direct acceleration of SAGA using sampled negative momentum. In: The 22Nd International Conference on Artificial Intelligence and Statistics, pp. 1602–1610 (2019)