IOP Publishing Inverse Problems
Inverse Problems **38** (2022) 025009 (34pp) https://doi.org/10.1088/1361-6420/ac4428

An analysis of stochastic variance reduced gradient for linear inverse problems*

Bangti Jin[1], Zehui Zhou[2] and Jun Zou[2,]**

[1] Department of Computer Science, University College London, Gower Street, London WC1E 6BT, United Kingdom
[2] Department of Mathematics, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong Special Administrative Region of China, People's Republic of China

E-mail: b.jin@ucl.ac.uk, zhzhou@math.cuhk.edu.hk and zou@math.cuhk.edu.hk

Received 25 August 2021, revised 7 December 2021
Accepted for publication 17 December 2021
Published 4 January 2022

Abstract
Stochastic variance reduced gradient (SVRG) is a popular variance reduction technique for accelerating stochastic gradient descent (SGD). We provide a first analysis of the method for solving a class of linear inverse problems in the lens of the classical regularization theory. We prove that for a suitable constant step size schedule, the method can achieve an optimal convergence rate in terms of the noise level (under suitable regularity condition) and the variance of the SVRG iterate error is smaller than that by SGD. These theoretical findings are corroborated by a set of numerical experiments.

Keywords: stochastic variance reduced gradient, regularizing property, convergence rate, saturation, inverse problems

(Some figures may appear in colour only in the online journal)

1. Introduction

In this paper, we consider the numerical solution of the following finite-dimensional linear inverse problem:

$$Ax = y^{\dagger}, \tag{1.1}$$

where $A \in \mathbb{R}^{n\times m}$ is the system matrix representing the data formation mechanism, and $x \in \mathbb{R}^m$ is the unknown signal of interest. In practice, we only have access to a noisy version y^{δ} of the

*The work of BJ is supported by UK EPSRC Grant EP/T000864/1. The work of JZ was substantially supported by Hong Kong RGC General Research Fund (Projects 14306718 and 14306719).

**Author to whom any correspondence should be addressed.

exact data $y^\dagger = Ax^\dagger$ (with $x^\dagger$ being the minimum norm solution relative to the initial guess x_0, cf (2.1)), i.e.,

$$y^\delta = y^\dagger + \xi,$$

where $\xi \in \mathbb{R}^n$ denotes the noise in the data with a noise level $\delta = \|\xi\|$, with $\|\cdot\|$ being the Euclidean norm of a vector (and also the spectral norm of a matrix). We denote the ith row of the matrix A by a column vector $a_i \in \mathbb{R}^m$, i.e., $A = [a_i^{\mathrm{t}}]_{i=1}^n$ (with the superscript t denoting the matrix/vector transpose), and the ith entry of the vector $y^\delta \in \mathbb{R}^n$ by y_i^δ. Linear inverse problems of the form (1.1) arise in a broad range of practical applications, e.g., computed tomography and optical imaging.

Over the last few years, stochastic iterative algorithms have received much interest in the inverse problems community. The most prominent example is stochastic gradient descent (SGD) due to Robbins and Monro [30]. The starting point is the following optimization problem:

$$J(x) = \frac{1}{2n}\|Ax - y^\delta\|^2 = \frac{1}{n}\sum_{i=1}^n f_i(x), \quad \text{with } f_i(x) = \frac{1}{2}\big((a_i, x) - y_i^\delta\big)^2, \tag{1.2}$$

where $(\cdot,\cdot)$ denotes the Euclidean inner product on $\mathbb{R}^m$. Then SGD reads as follows. Given an initial guess $\hat{x}_0^\delta \equiv x_0$, the iterate $\hat{x}_k^\delta$ is constructed as

$$\hat{x}_{k+1}^\delta = \hat{x}_k^\delta - \eta_k f_{i_k}'(\hat{x}_k^\delta),$$

where $\eta_k > 0$ is the step size at the $(k+1)$th step, and the index i_k is sampled uniformly from the index set $\{1, \ldots, n\}$. One attractive feature of the method is that the computational complexity per iteration does not depend on the data size n, and thus it is directly scalable to large data volume, which is especially attractive in the era of big data. SGD type methods have found applications in several inverse problems, e.g., randomized Kaczmarz method [12, 32] in computed tomography, ordered subset expectation maximization [13, 21] for positron emission tomography, and more recently also some nonlinear inverse problems, e.g., optical tomography [4] and phonon transmission coefficient recovery [8].

However, the relevant mathematical theory for inverse problems in the lens of regularization theory [7, 14, 20] is still not fully understood. Existing works [15–18] focus on the standard SGD for inverse problems, proving that SGD is a regularization method when equipped with a suitable stopping criterion, and the SGD iterates converge at a certain rate. However, the presence of stochastic gradient noise generally prevents SGD from converging to the solution when a constant step size is used and leads to a slow, sublinear rate of convergence when a diminishing step size schedule is employed. Among various acceleration strategies, variance reduction (VR) represents one prominent idea that has achieved great success, including SAG [24], SAGA [5], stochastic variance reduced gradient (SVRG) [19, 36] and SARAH [27] etc; these methods take advantage of the finite-sum structure prevalent in machine learning problems, and exhibit improved convergence behavior over SGD; see the work [9] for a recent overview of VR techniques in machine learning.

SVRG combines SGD with predictive VR and is very popular in stochastic optimization. It was proposed independently by two groups of researchers, i.e., Johnson and Zhang [19] and Zhang *et al* [36], for accelerating SGD for minimizing smooth and strongly convex objective functions. When applied to problem (1.2), the basic version of SVRG reads as follows. Given an initial guess $x_0^\delta \equiv x_0 \in \mathbb{R}^m$, SVRG updates the iterate x_k^δ recursively by

$$x_{k+1}^\delta = x_k^\delta - \eta_k \left(f_{i_k}'(x_k^\delta) - f_{i_k}'(x_{k_M}^\delta) + J'(x_{k_M}^\delta)\right), \quad k = 0, 1, \ldots, \tag{1.3}$$

Algorithm 1. SGD for problem (1.1).

Set initial guess $\hat{x}_0^\delta = x_0$ and step size schedule η_k
for $k = 0, 1, \cdots$ **do**
  draw i_k i.i.d. uniformly from $\{1, \cdots, n\}$
  update $\hat{x}_{k+1}^\delta = \hat{x}_k^\delta - \eta_k((a_{i_k}, \hat{x}_k^\delta) - y_{i_k}^\delta)a_{i_k}$
  check the stopping criterion
end

Algorithm 2. SVRG for problem (1.1).

Set initial guess $x_0^\delta = x_0$, frequency M and step size schedule η_k
for $K = 0, 1, \cdots$ **do**
  compute $J'(x_{KM}^\delta)$
  for $t = 0, 1, \cdots, M-1$ **do**
    draw i_{KM+t} i.i.d. uniformly from $\{1, \cdots, n\}$
    update $x_{KM+t+1}^\delta = x_{KM+t}^\delta - \eta_{KM+t}\big((a_{i_{KM+t}}, x_{KM+t}^\delta - x_{KM}^\delta)a_{i_{KM+t}} + J'(x_{KM}^\delta)\big)$
  end
  check the stopping criterion.
end

where the row index i_k is drawn uniformly from the index set $\{1, \ldots, n\}$, $\eta_k > 0$ is the step size at the kth iteration, M is the frequency of computing the full gradient, and $k_M = [\frac{k}{M}]M$, ($[\cdot]$ takes the integral part of a real number). The choice of the frequency M can affect the practical performance of the algorithm, and it was suggested to be $2n$ and $5n$ for convex and nonconvex optimization, respectively [19]. In this study, we show that SVRG can achieve optimal convergence rates when M is chosen such that $M \geqslant O(n^{\frac{1}{2}})$. When compared with SGD, SVRG employs the anchor/snapshot point $x_{k_M}^\delta$ to reduce the variance of the gradient estimate: it computes the full gradient $J'(x_{k_M}^\delta)$ of J at the anchor point $x_{k_M}^\delta$ for every M iterates, and then combines $J'(x_{k_M}^\delta)$ with the gradient gap $f_{i_k}'(x_k^\delta) - f_{i_k}'(x_{k_M}^\delta)$ to obtain a new gradient estimate for updating the SVRG iterate x_{k+1}^δ. In contrast, SGD employs the stochastic gradient $f_{i_k}'(\hat{x}_k^\delta)$ only, and the classical Landweber method (LM) uses only the gradient $J'(x)$. Thus, SVRG can be viewed as a hybridization between the LM and SGD. A detailed comparison between SGD and SVRG are given in algorithms 1 and 2, where SVRG is stated in the form of double loop. In practice, there are several variants of SVRG, dependent on the choice of the anchor point, e.g., last iterate, iterate average, random choice and weighted iterate average (within the inner loop). In this work, we study only the version given in algorithm 2.

It is known that VR enables speeding up the convergence of the algorithm in the sense of optimization [3, 9]. Since its first introduction, SVRG has received a lot of attention within the optimization community, and several convergence results of SVRG and its variants have been obtained [1, 2, 11, 23, 29, 31, 34]. Note that here the precise meaning of convergence depends crucially on the property of the objective function $J(x)$: (i) the distance of the SVRG iterate x_k^δ to a global minimizer for a strictly convex $J(x)$, (ii) the optimality gap (i.e., $J(x_k^\delta) - \min_x J(x)$) for a convex $J(x)$ and (iii) the norm of the gradient $\|J'(x_k^\delta)\|$ for a nonconvex $J(x)$, in terms of the iterate number k. For example, Allen-Zhu and Hazan [1] proved that SVRG (with a different choice of the anchor point) converges at an $O(n^{\frac{2}{3}}\epsilon^{-1})$ rate to an approximate stationary point x^* (i.e., $\|J'(x^*)\|^2 \leqslant \epsilon$) for a nonconvex but smooth $J(x)$. Reddi *et al* [29] proved a nonasymptotic

rate of convergence of SVRG for nonconvex optimization and identified a subclass of nonconvex problems (satisfied by gradient dominated functions) for which a variant of SVRG attains linear convergence.

These important breakthroughs in the optimization literature naturally motivate the following question: *does the desirable convergence property of SVRG carry over to inverse problems in the sense of regularization theory?* The answer to this question is not self-evident, since accelerated iterative schemes do not necessarily retain the optimal convergence in the sense of regularization (see [22, 26] for studies on Nesterov's accelerated scheme). For linear inverse problems in (1.1), the objective $J(x)$ in (1.2) is convex but not strictly so. Further, it is ill-posed in the sense that a global minimizer often does not exist, and even if it does exists, it is unstable with respect to the inevitable perturbation of the data y^δ and is probably physically irrelevant. Instead, we construct an approximate minimizer that converges to the exact solution $x^\dagger$ as the noise level δ tends to 0^+ by stopping the iteration properly, a procedure commonly known as iterative regularization (by early stopping) [20], and the accuracy of the approximation is measured in terms of the noise level δ. To the best of our knowledge, the theoretical properties of SVRG and other VR techniques have not been studied so far in the lens of regularization theory.

In this work, we contribute to the theoretical analysis of SVRG for a class of linear inverse problems from the perspective of classical regularization theory [7, 14, 20]. Under the constant step size schedule and the canonical source condition, we prove that the epochwise SVRG iterate x_{KM}^δ converges to the minimum norm solution $x^\dagger$ at an optimal rate (in terms of δ) when combined with *a priori* stopping rule, and that due to the built-in VR mechanism, for the same iterate number, the variance of SVRG iterate is indeed smaller than that of SGD, showing the beneficial effect of VR; see theorems 2.1 and 2.2. In particular, SVRG allows using larger step sizes than that for SGD while still overcoming the undesirable saturation phenomenon (cf remark 2.1). See section 2 for precise statements of the theoretical findings and related discussions in the context of inverse problems. These theoretical results are complemented by extensive numerical results in section 6.

The rest of the paper is organized as follows. In section 2 we present and discuss the main results of the work. In section 3, we recall preliminary results, especially a careful decomposition of the error of the epoch SVRG/SGD iterate into the bias and variance components. In section 4 we give the convergence rate analysis, and prove an optimal convergence rate, and in section 5 we present a comparative study of SVRG versus SGD, and show that variance component of the SVRG error is smaller than that of the SGD error. Finally, in section 6, we present several numerical experiments to complement the theoretical analysis. For better readability, the lengthy and technical proofs of several auxiliary results are deferred to the appendix. Throughout, the notation c with suitable subscripts denotes a generic constant.

2. Main results and discussions

In this section, we state the main results of the work. First we state the standing assumption. We denote by $\mathcal{F}_k$ the filtration generated by the random indices $\{i_0, i_1, \ldots, i_{k-1}\}$. Let $\mathcal{F} = \bigvee_{k=1}^\infty \mathcal{F}_k$, $\mathcal{F}_k^c = \mathcal{F}\backslash\mathcal{F}_k$, $(\Omega, \mathcal{F}, \mathbb{P})$ being the associated probability space, and $\mathbb{E}[\cdot]$ denotes taking the expectation with respect to the filtration $\mathcal{F}$ and $\mathbb{E}_j[\cdot] := \mathbb{E}[\cdot|\mathcal{F}_{j+1}^c \cup \mathcal{F}_j]$. The SVRG iterate x_k^δ is random, and measurable with respect to $\mathcal{F}_k$. Let $e_k^\delta = x_k^\delta - x^\dagger$ be the error of the SVRG iterate x_k^δ with respect to the unique minimum-norm solution $x^\dagger$, defined by

$$x^\dagger = \arg \min_{x\in\mathbb{R}^m : Ax = y^\dagger} \|x - x_0\|. \tag{2.1}$$

Let $B = \mathbb{E}[a_i a_i^t] = n^{-1}A^tA \in \mathbb{R}^{m\times m}$. Throughout we assume that $\|B\| \leqslant 1$, which can easily be achieved by scaling. In this work we consider a constant step size schedule, which is commonly employed by SVRG. Assumption 2.1(b) is commonly known as the source condition in the inverse problems literature [7], which implicitly assumes a certain regularity on the initial error. This condition is central for deriving convergence rates. It is well known that in the absence of source type conditions, the convergence for a regularization method can be arbitrarily slow [7]. Assumption 2.1(c) enables an important commuting property (cf lemma 3.2), which greatly facilitates the analysis. Numerically this property does not affect the performance of SVRG, and thus it seems largely due to the limitation of the analysis technique.

Assumption 2.1. The following assumptions hold.

(a) The step size $\eta_j = c_0$, $j = 0, 1, \ldots$, with $c_0 \leqslant \left(\max(\max_i \|a_i\|^2, \|B\|^2)\right)^{-1}$.
(b) There exist some $\nu > 0$ and $w \in \mathbb{R}^m$ such that the exact solution $x^\dagger$ satisfies $x^\dagger - x_0 = B^\nu w$.
(c) The matrix $A = \Sigma V^t$ with Σ being diagonal and nonnegative and V column orthonormal.

The next result represents the main theoretical contribution of the work. It implies that SVRG can achieve the optimal convergence rate for linear inverse problems under the given assumption on the step size. The step size restriction originates from the fact that SVRG still employs a randomized gradient estimate for the iterate update, albeit with reduced variance, when compared with the LM. Nonetheless, the restriction on the step size is more benign than that for SGD: it allows achieving optimal convergence rate under larger step size than that in SGD.

Theorem 2.1. *Let assumption 2.1 hold, and $c_* > 1$ satisfy*

$$(4 + 2(Mc_0\|B\|)^2)nM^{-2}c_B c_{B,M} \leqslant 1 - c_*^{-1} \tag{2.2}$$

$$\text{with } c_{B,M} = \sum_{i=1}^{M-1}(1 - (1 - c_0\|B\|)^i)^2 \quad \text{and} \quad c_B = (1 - c_0\|B\|)^{-M}.$$

Then with constants $c_\nu = \nu^\nu(Mc_0)^{-\nu}$ and $c_{**} = (3 + 2(Mc_0\|B\|)^2)nMc_Bc_0^2\|B\|$, there holds

$$\mathbb{E}[\|e_{KM}^\delta\|^2] \leqslant \left(2 + 2^{2\nu}\|B\|c_{**}c_*\right)c_\nu^2 K^{-2\nu}\|w\|^2 + (2Mc_0 + c_{**}c_*)K\bar{\delta}^2.$$

Remark 2.1. Let $c = c_0\|B\|M$, which implies $c_B = (1 - cM^{-1})^{-M}$ and $c_{B,M} = \sum_{i=1}^{M-1}(1 - (1 - cM^{-1})^i)^2$, the condition (2.2) is satisfied whenever

$$nM^{-2} \leqslant (1 - c_*^{-1})(4 + 2c^2)^{-1}c_B^{-1}c_{B,M}^{-1},$$

which holds for $M = \mathcal{O}(n^{\frac{1}{2}})$ and sufficiently small $c = \mathcal{O}(1)$. It is instructive to compare the conditions ensuring an optimal convergence rate of SVRG and SGD: SGD requires the condition $c_0 = \mathcal{O}(n^{-1})$ [18], whereas SVRG requires only $M = \mathcal{O}(n^{\frac{1}{2}})$ and $c = c_0\|B\|M = \mathcal{O}(1)$. The latter implies $c_0 = \mathcal{O}(n^{-\frac{1}{2}})$ for SVRG. Since $\mathcal{O}(n^{-\frac{1}{2}})$ is much larger than $\mathcal{O}(n^{-1})$ when the data size n is large, SVRG should perform better for truly large-scale problems.

It is known that SGD with an inadvertent choice of the step size schedule can lead to the undesirable saturation phenomenon, i.e., the convergence rate does not improve with the regularity index ν in assumption 2.1(b), whenever ν exceeds the critical value $1/2$ [15, 18]. This is attributed to the inherent variance of the stochastic gradient estimate used by SGD, and one important issue is to overcome the saturation phenomenon. The next result sheds further insight

into this phenomenon by comparing the mean squared error of the (epochwise) SVRG iterate with that of the corresponding SGD iterate: it gives a refined comparison between the variance components of SVRG and SGD iterates, in view of the bias-variance decomposition. In particular, it shows that the built-in VR mechanism of SVRG does reduce the variance component of the error, which represents a distinct feature of SVRG over SGD, especially alleviating the step size restriction for achieving the optimal convergence.

Theorem 2.2. *Let assumption* 2.1(*a*) *and* (*c*) *be fulfilled and the constants* c_0, n *and* M *satisfy, with the constant* $c_B' = (1 - c_0\|B\|)^{-2(M-1)}$,

$$(M-1)^2 c_0^2 \|B\|^2 \leqslant (2c_B')^{-1} \quad \text{and} \quad (M+1)^2 \leqslant (2c_B')^{-1}(n-1). \tag{2.3}$$

For any $K \geqslant 0$, *let* R_1 *and* R_2 *be measurable with respect to* $\mathcal{F}_{KM}^c$ *and* R_1 *is combination of* M_0 *and* H_k (*cf* (3.1) *for the definition*). *Then for* ζ *defined in section* 3.1, *there holds*

$$\mathbb{E}[\|R_1(e_{KM}^\delta - B^{-1}\zeta) + R_2\|^2] \leqslant \mathbb{E}[\|R_1(\hat{e}_{KM}^\delta - B^{-1}\zeta) + R_2\|^2].$$

Remark 2.2. Let $c := c_0\|B\|(M-1)$, which implies $c_B' = (1 - c(M-1)^{-1})^{-2(M-1)}$. Then condition (2.3) can be rewritten as

$$c^2 \leqslant 2^{-1}(c_B')^{-1} \quad \text{and} \quad (M+1)^2 \leqslant 2^{-1}(c_B')^{-1} n.$$

The first essentially requires $c < \frac{1}{2}$. For any $M \geqslant 2$, $c_B' \leqslant 2e^{2c}$, the condition can be satisfied by $2ce^c \leqslant 1$ and $M + 1 \leqslant 2^{-1}e^{-c}n^{\frac{1}{2}}$.

Last we briefly comment on the overall analysis strategy for proving theorems 2.1 and 2.2. The overall strategy is to derive the recursion of the epochwise SVRG iterate x_{KM}^δ (and also the SGD iterate $\hat{x}_{KM}^\delta$), for any $K = 0, 1, \ldots$, i.e., at the anchor points only, and then bound the error $e_{KM}^\delta := x_{KM}^\delta - x^\dagger$ by bias-variance decomposition

$$\mathbb{E}[\|x_{KM}^\delta - x^\dagger\|^2] = \|\mathbb{E}[x_{KM}^\delta] - x^\dagger\|^2 + \mathbb{E}[\|x_{KM}^\delta - \mathbb{E}[x_{KM}^\delta]\|^2].$$

The two terms on the right-hand side represent respectively the bias of the error due to early stopping and data noise and the computational variance of error due to randomness of the gradient estimate. These are analyzed in proposition 3.1 and lemma 4.1, respectively, and allow proving the convergence rate in theorem 2.1. The analysis of the variance component relies on a novel refined decomposition into terms that are more tractable to estimate for both SVRG and SGD. This decomposition is also crucial for the comparative study between SVRG and SGD, where a careful componentwise comparison of the decomposition allows establishing theorem 2.2. Note that the decomposition relies heavily on the constant step size schedule, and thus the overall analysis differs greatly from existing analysis of the SGD in the lens of regularization theory [16–18] or the analysis of SGD in statistical learning theory [6, 25, 28, 33, 35]. The extension of the analysis to a general step size schedule represents an interesting future research problem.

3. Error decomposition

In this part, we present several preliminary results, especially error decompositions for SVRG and SGD iterates. The decompositions play a central role in the convergence rates analysis and comparative analysis in sections 4 and 5, respectively.

3.1. Notation and preliminary estimates

First we introduce several shorthand notation. Below, we denote the SVRG iterates for the exact data $y^\dagger$ and noisy data y^δ by x_k and x_k^δ, respectively, and that for SGD by $\hat{x}_k$ and $\hat{x}_k^\delta$, respectively. We use extensively the following shorthand notation for any $k = 0, 1, \ldots$:

$$\begin{aligned}
e_k &= x_k - x^\dagger,\\
e_k^\delta &= x_k^\delta - x^\dagger,\\
\hat{e}_k &= \hat{x}_k - x^\dagger,\\
\hat{e}_k^\delta &= \hat{x}_k^\delta - x^\dagger,\\
\bar{A} &= n^{-\frac{1}{2}} A,\\
\bar{\xi} &= n^{-\frac{1}{2}} \xi,\\
\bar{\delta} &= n^{-\frac{1}{2}} \delta,\\
M_0 &= I - c_0 B,\\
\zeta &= \bar{A}^t \bar{\xi},\\
P_k &= I - c_0 a_{i_k} a_{i_k}^t,\\
N_k &= B - a_{i_k} a_{i_k}^t,\\
\zeta_k &= a_{i_k} \xi_{i_k}.
\end{aligned}$$

Note that P_k is the random update operator for the iteration, and we have the identity $P_k = M_0 + c_0 N_k$ trivially. For all $k \in \mathbb{N}$, let

$$H_k = G_{k+1} N_k, \quad \text{with } G_k = \begin{cases} \prod_{i=k}^{k_M+M-1} P_i, & k \neq KM, \\ I, & k = KM. \end{cases} \tag{3.1}$$

Clearly, $H_{KM-1} = N_{KM-1}$. By definition, we have the following identity

$$\begin{aligned}
G_{KM+j} &= G_{KM+j+1} P_{KM+j} = G_{KM+j+1}(M_0 + c_0 N_{KM+j})\\
&= G_{KM+j+1} M_0 + c_0 H_{KM+j}, \quad j = 1, \ldots, M-1.
\end{aligned} \tag{3.2}$$

These notations are useful for representing the (epochwise) SVRG iterates x_{KM}^δ, cf proposition 3.1. The following simple identity will be used extensively.

Lemma 3.1. *The following identity holds*

$$G_{KM+i} = M_0^{M-i} + c_0 \sum_{j=0}^{M-i-1} H_{KM+i+j} M_0^j, \quad i = 1, \ldots, M-1. \tag{3.3}$$

Proof. It follows directly from the definition of G_k and H_k and the identity (3.2) that

$$\begin{aligned} G_{KM+i} &= G_{KM+i+1}M_0 + c_0\sum_{j=0}^{0} H_{KM+i}M_0^j \\ &= G_{KM+i+2}M_0^2 + c_0\sum_{j=0}^{1} H_{KM+i+j}M_0^j \\ &= \cdots = M_0^{M-i} + c_0\sum_{j=0}^{M-i-1} H_{KM+i+j}M_0^j. \end{aligned}$$

This shows the desired identity. □

We use extensively the following direct consequence of assumption 2.1(c).

Lemma 3.2. *Under assumption 2.1(c), the matrices M_0, B, P_{i_j} and $N_{i_{j'}}$ are commutative for any j and j'.*

Proof. Note that, for any j and j', we have

$$\begin{aligned} B &= n^{-1}\sum_{i=1}^{n} a_i a_i^t, \\ M_0 &= I - c_0 B = I - c_0 n^{-1}\sum_{i=1}^{n} a_i a_i^t, \\ P_{i_j} &= I - c_0 a_{i_j} a_{i_j}^t, \\ N_{i_{j'}} &= B - a_{i_{j'}} a_{i_{j'}}^t = n^{-1}\sum_{i=1}^{n} a_i a_i^t - a_{i_{j'}} a_{i_{j'}}^t. \end{aligned}$$

It suffices to show the claim that $a_i a_i^t$ and $a_j a_j^t$ are commutative for any $i, j = 1, \ldots, n$. This claim is trivial when $i = j$. If $i \neq j$, by assumption 2.1(c), there holds $a_i^t a_j = 0 = a_j^t a_i$. □

We also state an identity which is crucial for the proofs of theorems 2.1 and 2.2.

Lemma 3.3. *Let assumption 2.1(c) be fulfilled. Then for any diagonal matrix $D \in \mathbb{R}^{m\times m}$ and any vector $v \in \mathbb{R}^m$, which are independent of i_j, the following identities hold*

$$\begin{aligned} \mathbb{E}[\|VDV^t N_j v\|^2] &= (n-1)\mathbb{E}[\|VDV^t B v\|^2], \\ \mathbb{E}[\|VDV^t(\zeta_j - \zeta)\|^2] &= (n-1)\mathbb{E}[\|VDV^t \zeta\|^2]. \end{aligned}$$

Proof. Recall the standard bias-variance decomposition: for any matrix R and filtration $\mathcal{F}_a$,

$$\mathbb{E}[\|R - \mathbb{E}[R|\mathcal{F}_a]\|^2|\mathcal{F}_a] = \mathbb{E}[\|R\|^2|\mathcal{F}_a] - \|\mathbb{E}[R|\mathcal{F}_a]\|^2.$$

Then the identity $N_j = B - a_{i_j} a_{i_j}^t = \mathbb{E}_j[a_{i_j} a_{i_j}^t] - a_{i_j} a_{i_j}^t$ gives

$$\begin{aligned} \mathbb{E}_j[\|VDV^t N_j v\|^2] &= \mathbb{E}_j[\|VDV^t a_{i_j} a_{i_j}^t v\|^2] - \|VDV^t B v\|^2 \\ &= n^{-1}\sum_{i=1}^{n} \|VDV^t a_i a_i^t v\|^2 - \|VDV^t B v\|^2, \end{aligned}$$

where $a_i a_i^t v = A^t(a_i^t v)b_i$ with $b_i = (0, \ldots, 0, 1, 0, \ldots, 0)^t \in \mathbb{R}^n$ being the ith canonical Cartesian basis vector. By assumption 2.1(c), $DV^tA^t = D\Sigma$ is diagonal, and hence

$$n^{-1}\sum_{i=1}^{n} \|VDV^t a_i a_i^t v\|^2 = n^{-1}\sum_{i=1}^{n} \|VDV^t A^t(a_i^t v)b_i\|^2$$
$$= n^{-1}\|VDV^t A^t \sum_{i=1}^{n}(a_i^t v)b_i\|^2 = n\|VDV^t Bv\|^2.$$

This shows the first identity. Similarly, since $\mathbb{E}_j[\zeta_j] = \zeta$, by rewriting ζ_j as $\zeta_j = a_{t_j}\xi_j = A^t\xi_j b_{t_j}$, we obtain the second identity. This completes the proof of the lemma. □

Next we recall two technical estimates; see the appendix for the proof.

Lemma 3.4. *Let assumption 2.1(a) be fulfilled. For any $s \geqslant 0$, $t \in [0, 1]$ and $K \in \mathbb{N}$, there hold*

$$\|B^{-t}(I - M_0^{KM})\| \leqslant (Mc_0)^t K^t,$$
$$\|B^s M_0^{KM}\| \leqslant s^s (Mc_0)^{-s} K^{-s} := c_s K^{-s}.$$

3.2. Error decomposition

Now we derive error decompositions for the (epochwise) SVRG error $e_{KM}^\delta \equiv x_{KM}^\delta - x^\dagger$ and the SGD error $\hat{e}_{KM}^\delta \equiv \hat{x}_{KM}^\delta - x^\dagger$ into the bias and variance components. These representations follow from direct but lengthy computation using the definitions the SVRG and SGD iterates, and the detailed proof is deferred to the appendix.

Proposition 3.1. *Under assumption 2.1(a), for any $K \geqslant 1$, there hold*

$$\mathbb{E}[e_{(K+1)M}^\delta] = M_0^{(K+1)M} e_0^\delta + (I - M_0^{(K+1)M})B^{-1}\zeta$$
$$e_{(K+1)M}^\delta - \mathbb{E}[e_{(K+1)M}^\delta] = \sum_{j=0}^{K} M_0^{(K-j)M} L_j(\zeta - Be_{jM}^\delta),$$

with the random matrices L_j defined by

$$L_j = c_0 \sum_{i=1}^{M-1} H_{jM+i}(I - M_0^i)B^{-1}.$$

The next result gives an analogous bias-variance decomposition for the SGD iterate $\hat{x}_{KM}^\delta$. Note that when compared with proposition 3.1, the expressions of $\mathbb{E}[x_{KM}^\delta]$ and $\mathbb{E}[\hat{x}_{KM}^\delta]$ are actually identical, since both methods use an unbiased estimate for the gradient. Their difference lies in the variance component, which will be the main focus of the analysis below.

Proposition 3.2. *Under assumption 2.1(a), for any $K \geqslant 0$, $\hat{e}^{\delta}_{(K+1)M}$ satisfies*

$$\mathbb{E}[\hat{e}^{\delta}_{(K+1)M}] = M_0^{(K+1)M}\hat{e}^{\delta}_0 + (I - M_0^{(K+1)M})B^{-1}\zeta,$$

$$\begin{aligned}\hat{e}^{\delta}_{(K+1)M} - \mathbb{E}[\hat{e}^{\delta}_{(K+1)M}] = {} & c_0^2\sum_{j=0}^{K}\sum_{i=0}^{M-2}\sum_{t=0}^{M-i-2} M_0^{(K-j)M}H_{jM+i+t+1}M_0^t(\zeta_{jM+i} - \zeta)\\ & + c_0\sum_{j=0}^{K}\sum_{i=0}^{M-1} M_0^{(K-j)M}\left(H_{jM+i}\left(M_0^i\hat{e}^{\delta}_{jM} + (I - M_0^i)B^{-1}\zeta\right)\right.\\ & \left. + M_0^{M-i-1}(\zeta_{jM+i} - \zeta)\right).\end{aligned}$$

Remark 3.1. Equation (A.4) in the proof (in the appendix) indicates that at the snapshot point x^{δ}_{KM}, SVRG performs a gradient descent step, and in-between the snapshot points, the update direction is a linear combination between gradient and gradient offset (between the current iterate and the anchor point). Thus in this sense, SVRG is actually a hybridization of the Landweber method and SGD. Note that since $J'(x^{\delta}_{k_M})$ is independent of the random index i_k and the gap $f'_{i_k}(x^{\delta}_k) - f'_{i_k}(x^{\delta}_{k_M})$ is independent of the noise ξ_k for linear inverse problems, the SVRG iterate x^{δ}_k does not actually depend on ξ_k. This property contributes to the variance reduction, and constitutes one major difference between SVRG and SGD in terms of the noise influence.

4. Proof of theorem 2.1

Now we prove the convergence rate for SVRG in theorem 2.1. We begin with bounding the mean squared residual $\mathbb{E}[\|R_1(e^{\delta}_{KM} - B^{-1}\zeta) + R_2\|^2]$ and weighted variance $\mathbb{E}[\|R_1(e^{\delta}_{KM} - \mathbb{E}[e^{\delta}_{KM}])\|^2]$, where the quantities R_1 and R_2 are measurable with respect to the filtration $\mathcal{F}^c_{KM}$ and commutative with B, M_0, $\{P_k\}$ and $\{N_k\}$ for any $k \geqslant 0$. The specific forms of R_1 and R_2 arise from the refined decompositions of SVRG errors in lemma 4.1 and SGD errors in lemma 5.1, in order to carry out the componentwise comparison between them; see the proof of theorem 2.2 in section 5 for further details.

Lemma 4.1. *Under assumption 2.1(a) and (c), for any $K \geqslant 0$, let R_1 and R_2 be measurable with respect to $\mathcal{F}^c_{(K+1)M}$ and commutative with B, M_0, $\{P_k\}$ and $\{N_k\}$, for any $k \geqslant 0$. Then there hold*

$$\mathbb{E}[\|R_1(e^{\delta}_{(K+1)M} - B^{-1}\zeta) + R_2\|^2] = \mathrm{I}_0 + \sum_{j=0}^{K}\mathrm{I}_{1,j},$$

$$\mathbb{E}[\|R_1(e^{\delta}_{(K+1)M} - \mathbb{E}[e^{\delta}_{(K+1)M}])\|^2] = \sum_{j=0}^{K}\mathrm{I}_{1,j},$$

with the terms I_0 and $\mathrm{I}_{1,j}$ given by

$$\mathrm{I}_0 = \mathbb{E}[\|R_1M_0^{(K+1)M}(e^{\delta}_0 - B^{-1}\zeta) + R_2\|^2], \tag{4.1}$$

$$\mathrm{I}_{1,j} = c_0^2\sum_{i=1}^{M-1}\mathbb{E}[\|R_1M_0^{(K-j)M}H_{jM+i}(I - M_0^i)(e^{\delta}_{jM} - B^{-1}\zeta)\|^2]. \tag{4.2}$$

Now we bound the mean squared (generalized) residual $\mathbb{E}[\|R_1(e_{KM}^\delta - B^{-1}\zeta)\|^2]$ of the epochwise SVRG iterate x_{KM}^δ. This bound is useful in the proof of theorem 2.1 below. The proof relies on mathematical induction, and the decomposition in lemma 4.1.

Theorem 4.1. *Let assumption 2.1(a) and (c) be fulfilled, R_1 be a combination of M_0 and B, and $c_* > 1$ be chosen such that (2.2) holds. Then for any $K \geqslant 0$, there holds*

$$\mathbb{E}[\|R_1(e_{KM}^\delta - B^{-1}\zeta)\|^2] \leqslant c_*\|R_1 M_0^{\frac{KM}{2}}(e_0^\delta - B^{-1}\zeta)\|^2.$$

Proof. We prove the theorem by mathematical induction. The case $K = 0$ holds true trivially. Now assume that the assertion holds up to some $K \geqslant 0$, i.e.,

$$\mathbb{E}[\|R_1(e_{jM}^\delta - B^{-1}\zeta)\|^2] \leqslant c_*\|R_1 M_0^{\frac{jM}{2}}(e_0^\delta - B^{-1}\zeta)\|^2, \quad j = 0, 1, \ldots, K, \tag{4.3}$$

and we prove it for the case $K + 1$. Lemma 4.1 with $R_2 = 0$ gives

$$\mathbb{E}[\|R_1(e_{(K+1)M}^\delta - B^{-1}\zeta)\|^2] = \mathrm{I}_0 + \sum_{j=0}^{K} \mathrm{I}_{1,j},$$

with the terms I_0 and $\mathrm{I}_{1,j}$ given by (4.1) (with $R_2 = 0$) and (4.2). Note that $V^t R_1 M_0^{(K-j)M} V$ is diagonal, then direct computation with lemmas 3.2 and 3.3, the inequalities $\|G_{jM+i+1}\| \leqslant 1$ and $\|I - M_0^i\| = 1 - (1 - c_0\|B\|)^i$ and the definition of the constant $c_{B,M}$ in theorem 2.1 gives

$$\begin{aligned}
\mathrm{I}_0 &\leqslant \mathbb{E}\left[\|R_1 M_0^{\frac{(K+1)M}{2}}(e_0^\delta - B^{-1}\zeta)\|^2\right],\\
\mathrm{I}_{1,j} &\leqslant c_0^2 \sum_{i=1}^{M-1} \|I - M_0^i\|^2 \|G_{jM+i+1}\|^2 \mathbb{E}[\|R_1 M_0^{(K-j)M} N_{jM+i}(e_{jM}^\delta - B^{-1}\zeta)\|^2]\\
&\leqslant n c_0^2 c_{B,M} \mathbb{E}[\|R_1 M_0^{(K-j)M} B(e_{jM}^\delta - B^{-1}\zeta)\|^2]\\
&\leqslant n c_0^2 c_{B,M} \|M_0^{-\frac{M}{2}}\|^2 \|M_0^{\frac{(K-j)M}{2}} B\|^2 \mathbb{E}\left[\|R_1 M_0^{\frac{(K-j+1)M}{2}}(e_{jM}^\delta - B^{-1}\zeta)\|^2\right].
\end{aligned}$$

This, the induction hypothesis (4.3), and the identity

$$\|M_0^{-\frac{M}{2}}\|^2 = (1 - c_0\|B\|)^{-M} := c_B, \tag{4.4}$$

give

$$\sum_{j=0}^{K} \mathrm{I}_{1,j} \leqslant n c_0^2 c_B c_{B,M} c_* \sum_{j=0}^{K} \|M_0^{\frac{(K-j)M}{2}} B\|^2 \|R_1 M_0^{\frac{(K+1)M}{2}}(e_0^\delta - B^{-1}\zeta)\|^2.$$

By lemma 3.4,

$$\|M_0^{\frac{(K-j)M}{2}} B\| \leqslant 2((K-j)Mc_0)^{-1}, \quad j = 0, \ldots, K-2,$$

and consequently,

$$\sum_{j=0}^{K} \|M_0^{\frac{(K-j)M}{2}} B\|^2 \leqslant 2\|B\|^2 + 4c_0^{-2}M^{-2}\sum_{j=0}^{K-2}(K-j)^{-2}$$
$$\leqslant (4 + 2(Mc_0\|B\|)^2)c_0^{-2}M^{-2}. \tag{4.5}$$

The preceding estimates together imply

$$\mathbb{E}[\|R_1(e^{\delta}_{(K+1)M} - B^{-1}\zeta)\|^2] \leqslant \left(1 + (4 + 2(Mc_0\|B\|)^2)nM^{-2}c_B c_{B,M} c_*\right)$$
$$\times \|R_1 M_0^{\frac{(K+1)M}{2}}(e_0^{\delta} - B^{-1}\zeta)\|^2.$$

The condition on c_* from (2.2) shows the induction step, and this completes the proof of the theorem. □

Setting $R_1 = n^{\frac{1}{2}}B^{\frac{1}{2}}$ in theorem 4.1 gives an upper bound on the mean squared residual $\mathbb{E}[\|Ax^{\delta}_{KM} - y^{\delta}\|^2]$ of the (epochwise) SVRG iterate x^{δ}_{KM}. Note that the mean squared residual consists of one decaying term related to the source condition in assumption 2.1(b) and one constant term related to the noise level. In particular, it is essentially bounded, independent of the iteration index. This behavior is similar to that for the standard LM.

Corollary 4.1. *Under assumption* 2.1 *and condition* (2.2)*, there holds*

$$\mathbb{E}[\|Ax^{\delta}_{KM} - y^{\delta}\|^2] \leqslant 2^{2\nu+2}c^2_{\nu+\frac{1}{2}}nc_*K^{-2\nu-1}\|w\|^2 + 2nc_*\bar{\delta}^2.$$

Proof. Theorem 4.1 and the triangle inequality imply (noting $e_0^{\delta} = e_0$)

$$\mathbb{E}[\|Ax^{\delta}_{KM} - y^{\delta}\|^2] = \mathbb{E}\left[\|n^{\frac{1}{2}}B^{\frac{1}{2}}(e^{\delta}_{KM} - B^{-1}\zeta)\|^2\right]$$
$$\leqslant nc_*\|B^{\frac{1}{2}}M_0^{\frac{KM}{2}}(e_0^{\delta} - B^{-1}\zeta)\|^2,$$
$$\leqslant 2nc_*\|M_0^{\frac{KM}{2}}B^{\frac{1}{2}}e_0^{\delta}\|^2 + 2nc_*\|M_0^{\frac{KM}{2}}B^{-\frac{1}{2}}\zeta\|^2.$$

Meanwhile, it follows from lemma 3.4 and the source condition in assumption 2.1(b) that

$$\|M_0^{\frac{KM}{2}}B^{\frac{1}{2}}e_0\| \leqslant 2^{\nu+\frac{1}{2}}c_{\nu+\frac{1}{2}}K^{-\nu-\frac{1}{2}}\|w\|,$$
$$\|M_0^{\frac{KM}{2}}B^{-\frac{1}{2}}\zeta\|^2 \leqslant \|M_0^{\frac{KM}{2}}B^{-\frac{1}{2}}\bar{A}^t\|^2\|\bar{\xi}\|^2 \leqslant \bar{\delta}^2.$$

Combining the preceding estimates gives the desired assertion. □

Now we can present the proof of theorem 2.1. The proof employs the representation in theorem 4.1, and follows by directly bounding the involved terms using lemma 3.4 (under assumption 2.1(b)) and theorem 4.1.

Proof. By lemma 4.1, setting $R_1 = I$ and $R_2 = B^{-1}\zeta$ gives

$$\mathbb{E}[\|e^{\delta}_{(K+1)M}\|^2] \leqslant \mathrm{I}_0 + \sum_{j=0}^{K}\mathrm{I}_{1,j},$$

with the terms I_0 and $\mathrm{I}_{1,j}$ given by (4.1) and (4.2), respectively. Now we bound them separately. By the triangle inequality, assumption 2.1(b) and lemma 3.4, we deduce

$$\begin{aligned}\mathrm{I}_0 &= \|M_0^{(K+1)M}e_0 + (I - M_0^{(K+1)M})B^{-1}\zeta\|^2\\ &\leqslant 2\|M_0^{(K+1)M}e_0\|^2 + 2\|(I - M_0^{(K+1)M})B^{-1}\bar{A}^t\bar{\xi}\|^2\\ &\leqslant 2c_\nu^2(K+1)^{-2\nu}\|w\|^2 + 2Mc_0(K+1)\bar{\delta}^2.\end{aligned}$$

Meanwhile, (4.2) with $R_1 = I$ gives

$$\mathrm{I}_{1,j} = c_0^2\sum_{i=1}^{M-1}\mathbb{E}[\|M_0^{(K-j)M}(I - M_0^i)H_{jM+i}(e_{jM}^\delta - B^{-1}\zeta)\|^2].$$

Note that by lemma 3.2, the matrices $I - M_0^i$ and H_{jM+i} are commuting, and $H_{jM+i} = G_{jM+i+1}N_{jM+i}$. Thus by lemma 3.3 (with $V^tM_0^{(K-j)M}(I - M_0^i)G_{jM+i+1}V$ being diagonal) and $\|G_{jM+i+1}\| \leqslant 1$, we obtain

$$\begin{aligned}\mathrm{I}_{1,j} &= (n-1)c_0^2\sum_{i=1}^{M-1}\|M_0^{(K-j)M}(I - M_0^i)G_{jM+i+1}B(e_{jM}^\delta - B^{-1}\zeta)\|^2\\ &\leqslant nc_0^2\sum_{i=1}^{M-1}\mathbb{E}[\|M_0^{(K-j)M}B(I - M_0^i)(e_{jM}^\delta - B^{-1}\zeta)\|^2].\end{aligned}$$

Next by the identity

$$c_0\sum_{i=0}^{j-1}M_0^i = (I - M_0^j)B^{-1},$$

the trivial inequality $(\sum_{t=0}^{i-1}a_t)^2 \leqslant i\sum_{t=0}^{i-1}a_t^2$, and $\|M_0\| \leqslant 1$, we have

$$\begin{aligned}\mathrm{I}_{1,j} &\leqslant nc_0^4\sum_{i=1}^{M-1}\mathbb{E}\left[\|M_0^{(K-j)M}B^2\sum_{t=0}^{i-1}M_0^t(e_{jM}^\delta - B^{-1}\zeta)\|^2\right]\\ &\leqslant nc_0^4\sum_{i=1}^{M-1}i\sum_{t=0}^{i-1}\mathbb{E}[\|M_0^{(K-j)M}BM_0^t(Be_{jM}^\delta - \zeta)\|^2]\\ &\leqslant nc_0^4\sum_{i=1}^{M-1}i^2\mathbb{E}[\|M_0^{(K-j)M}B(Be_{jM}^\delta - \zeta)\|^2].\end{aligned}$$

Since $\sum_{i=1}^{M-1}i^2 \leqslant 3^{-1}M^3$, it follows from theorem 4.1 and (4.4) that

$$\begin{aligned}\mathrm{I}_{1,j} &\leqslant 3^{-1}nM^3c_0^4\mathbb{E}[\|M_0^{(K-j)M}B(Be_{jM}^\delta - \zeta)\|^2].\\ &\leqslant 3^{-1}nM^3c_0^4\|M_0^{-\frac{M}{2}}\|^2\|M_0^{\frac{(K-j)M}{2}}B\|^2\mathbb{E}\left[\|M_0^{\frac{(K-j+1)M}{2}}(Be_{jM}^\delta - \zeta)\|^2\right]\\ &\leqslant 3^{-1}nc_BM^3c_0^4c_*\|M_0^{\frac{(K-j)M}{2}}B\|^2\|M_0^{\frac{(K+1)M}{2}}(Be_0^\delta - \zeta)\|^2.\end{aligned}$$

This and the inequality (4.5) imply

$$\sum_{j=0}^{K} \mathrm{I}_{1,j} \leqslant 3^{-1}(4+2(Mc_0\|B\|)^2)nMc_Bc_0^2c_*\|M_0^{\frac{(K+1)M}{2}}(Be_0^\delta-\zeta)\|^2$$
$$\leqslant (3+2(Mc_0\|B\|)^2)nMc_Bc_0^2c_*\left(2^{2\nu}\|B\|^2c_\nu^2(K+1)^{-2\nu}\|w\|^2+\|B\|\bar{\delta}^2\right).$$

The last two estimates together yield

$$\mathbb{E}[\|e_{(K+1)M}^\delta\|^2] \leqslant \left(2+2^{2\nu}(3+2(Mc_0\|B\|)^2)nMc_Bc_0^2\|B\|^2c_*\right)$$
$$\times c_\nu^2(K+1)^{-2\nu}\|w\|^2+(2Mc_0$$
$$+(3+2(Mc_0\|B\|)^2)nMc_Bc_0^2\|B\|c_*\big)(K+1)\bar{\delta}^2.$$

This completes the proof of the theorem. □

5. Proof of theorem 2.2

This section is devoted to the proof of theorem 2.2, and presents a comparative study on the variance $\mathbb{E}[\|e_{KM}^\delta-\mathbb{E}[e_{KM}^\delta]\|^2]$ of SVRG iterates with $\mathbb{E}[\|\hat{e}_{KM}^\delta-\mathbb{E}[\hat{e}_{KM}^\delta]\|^2]$ of SGD iterates. First we give a bias-variance decomposition of the SGD iterate $\hat{x}_{KM}^\delta$, in analogy with lemma 4.1. The representations in lemmas 4.1 and 5.1 facilitate the comparison between the variance components directly, which, under certain conditions, enables comparing the variance of SVRG and SGD iterates.

Lemma 5.1. *Under assumption 2.1(a) and (c), for any $K\geqslant 0$, let R_1 and R_2 be measurable with respect to $\mathcal{F}_{(K+1)M}^c$ and commutative with B, M_0, $\{P_k\}$ and $\{N_k\}$, for any $k\geqslant 0$. Then there hold*

$$\mathbb{E}[\|R_1(\hat{e}_{(K+1)M}^\delta-B^{-1}\zeta)+R_2\|^2]=\mathrm{I}_0+\sum_{j=0}^{K}(\mathrm{I}_{2,j}+\mathrm{I}_{3,j}),$$

$$\mathbb{E}[\|R_1(\hat{e}_{(K+1)M}^\delta-\mathbb{E}[\hat{e}_{(K+1)M}^\delta])\|^2]=\sum_{j=0}^{K}(\mathrm{I}_{2,j}+\mathrm{I}_{3,j}),$$

with I_0 given by (4.1) and $\mathrm{I}_{2,j}$ and $\mathrm{I}_{3,j}$ given by

$$\mathrm{I}_{2,j}=c_0^2\sum_{i=0}^{M-1}\mathbb{E}\left[\|R_1M_0^{(K-j)M}\left(H_{jM+i}M_0^i(\hat{e}_{jM}^\delta-B^{-1}\zeta)\right.\right.$$
$$\left.\left.+H_{jM+i}B^{-1}\zeta+M_0^{M-i-1}(\zeta_{jM+i}-\zeta)\right)\|^2\right], \tag{5.1}$$

$$\mathrm{I}_{3,j}=c_0^4\sum_{i=1}^{M-1}\sum_{t=0}^{i-1}\mathbb{E}[\|R_1M_0^{(K-j)M}H_{jM+i}M_0^t(\zeta_{jM+i-1-t}-\zeta)\|^2]. \tag{5.2}$$

Now, we can prove theorem 2.2. This result states that the variance component of the SVRG iterate x_{KM}^δ is indeed smaller than that of the SGD iterate $\hat{x}_{KM}^\delta$, as one may expect from the con-

struction of VR, and thus the VR step does reduce the variance of the iterate, thereby alleviating the deleterious effect of the stochastic iteration noise on the convergence of the SVRG iterates. The proof relies heavily on the explicit representations of the variances for the iterates x_{KM}^{δ} and $\hat{x}_{KM}^{\delta}$ derived in lemmas 4.1 and 5.1, and employs mathematical induction, certain independence relations (cf (5.5)–(5.7)) as well as lengthy computation.

Proof. Recall that the assumption on R_1 implies that it is commutative with B, M_0, $\{P_k\}$ and $\{N_k\}$ for any $k \geqslant 0$, and that in the inequality, R_1 and R_2 are measurable with respect to $\mathcal{F}_{jM}^{c}$ (when considering e_{jM}^{δ}). These facts will be used extensively without explicit mentioning below. The proof proceeds by mathematical induction. The case $K = 0$ is trivial since $\hat{e}_0^{\delta} = e_0^{\delta}$. Now suppose that the assertion holds up to some K, i.e.,

$$\mathbb{E}[\|R_1(e_{jM}^{\delta} - B^{-1}\zeta) + R_2\|^2] \leqslant \mathbb{E}[\|R_1(\hat{e}_{jM}^{\delta} - B^{-1}\zeta) + R_2\|^2], \quad j = 0, 1, \ldots, K, \quad (5.3)$$

and we prove it for $j = K + 1$. By lemmas 4.1 and 5.1, we deduce

$$\mathbb{E}[\|R_1(e_{(K+1)M}^{\delta} - B^{-1}\zeta) + R_2\|^2] = \mathrm{I}_0 + \sum_{j=0}^{K} \mathrm{I}_{1,j},$$

$$\mathbb{E}[\|R_1(\hat{e}_{(K+1)M}^{\delta} - B^{-1}\zeta) + R_2\|^2] = \mathrm{I}_0 + \sum_{j=0}^{K} (\mathrm{I}_{2,j} + \mathrm{I}_{3,j}),$$

with the terms $\mathrm{I}_{1,j}$, $\mathrm{I}_{2,j}$ and $\mathrm{I}_{3,j}$ are given by (4.2), (5.1) and (5.2), respectively. Thus, it suffices to show

$$\mathrm{I}_{1,j} \leqslant \mathrm{I}_{2,j} + \mathrm{I}_{3,j}, \quad j = 0, 1, \ldots, K. \quad (5.4)$$

By the inequality $(\sum_{t=1}^{i} a_i)^2 \leqslant i\sum_{t=1}^{i} a_i^2$, (A.5) and the identity $\|M_0^{-1}\| = (1 - c_0\|B\|)^{-1}$, we have

$$\begin{aligned}
\mathrm{I}_{1,j} &= c_0^4 \sum_{i=1}^{M-1} \mathbb{E}\left[\|R_1 M_0^{(K-j)M} H_{jM+i} B \sum_{t=0}^{i-1} M_0^t (e_{jM}^{\delta} - B^{-1}\zeta)\|^2\right] \\
&\leqslant c_0^4 \sum_{i=1}^{M-1} i \sum_{t=0}^{i-1} \|M_0^{-i}\|^2 \mathbb{E}\left[\|M_0^i R_1 M_0^{(K-j)M} H_{jM+i} B M_0^t (e_{jM}^{\delta} - B^{-1}\zeta)\|^2\right] \\
&\leqslant c_0^4 \sum_{i=1}^{M-1} i(1 - c_0\|B\|)^{-2i} \sum_{t=0}^{i-1} \mathbb{E}\left[\|M_0^i R_1 M_0^{(K-j)M} H_{jM+i} M_0^t B (e_{jM}^{\delta} - B^{-1}\zeta)\|^2\right] \\
&\leqslant c_0^4 \sum_{i=1}^{M-1} i(1 - c_0\|B\|)^{-2i} \sum_{t=0}^{i-1} \mathbb{E}\left[\|M_0^i R_1 M_0^{(K-j)M} H_{jM+i} M_0^t B (\hat{e}_{jM}^{\delta} - B^{-1}\zeta)\|^2\right],
\end{aligned}$$

where the last step is due to the induction hypothesis (5.3). Then by lemma 3.2, adding and subtracting suitable terms, and the triangle inequality, since $\|M_0\| \leqslant 1$, we deduce (with shorthand notation $c_B' = (1 - c_0\|B\|)^{-2(M-1)}$)

$$
\begin{aligned}
\mathrm{I}_{1,j} \leqslant c_0^4 \sum_{i=1}^{M-1} i(1-c_0\|B\|)^{-2i} \sum_{t=0}^{i-1} \mathbb{E}\Big[\|R_1 M_0^{(K-j)M+t} B \\
\times \left(H_{jM+i} M_0^i (\hat{e}_{jM}^{\delta} - B^{-1}\zeta) + H_{jM+i} B^{-1}\zeta + M_0^{M-i-1}(\zeta_{jM+i} - \zeta)\right) \\
- R_1 M_0^{(K-j)M+t} \left(H_{jM+i}\zeta + M_0^{M-i-1} B(\zeta_{jM+i} - \zeta)\right) \|^2 \Big] \\
\leqslant 2(M-1)^2 \|B\|^2 c_B' c_0^2 \mathrm{I}_{2,j} + c_0^4 \sum_{i=1}^{M-1} i(1-c_0\|B\|)^{-2i} \\
\times \sum_{t=0}^{i-1} \Big(4\mathbb{E}[\|R_1 M_0^{(K-j)M+t} H_{jM+i}\zeta\|^2] \\
+ 4\mathbb{E}[\|R_1 M_0^{(K-j+1)M+t-i-1} B(\zeta_{jM+i} - \zeta)\|^2] \Big).
\end{aligned}
$$

Now assumption 2.1(c) and the condition on R_1 imply that $V^t R_1 M_0^{s_1} G_{k+1} N_k^{s_3} B^{s_2} V$ is diagonal for any $s_1, s_2 \geqslant 0$, $s_3 = 0, 1$ and $k \in \mathbb{N}$. Thus, by lemma 3.3, we obtain

$$
\mathbb{E}[\|R_1 M_0^{(K-j)M+t} H_{jM+i}\zeta\|^2] = (n-1)\mathbb{E}[\|R_1 M_0^{(K-j)M+t} G_{jM+i+1} B\zeta\|^2], \tag{5.5}
$$

$$
\begin{aligned}
&\mathbb{E}[\|R_1 M_0^{(K-j+1)M+t-i-1} B(\zeta_{jM+i} - \zeta)\|^2] \\
&\quad = (n-1)\mathbb{E}[\|R_1 M_0^{(K-j+1)M+t-i-1} B\zeta\|^2],
\end{aligned} \tag{5.6}
$$

$$
\begin{aligned}
&\mathbb{E}[\|R_1 M_0^{(K-j)M+t} H_{jM+i}(\zeta_{jM+i-1-t} - \zeta)\|^2] \\
&\quad = (n-1)\mathbb{E}[\|R_1 M_0^{(K-j)M+t} H_{jM+i}\zeta\|^2] \\
&\quad = (n-1)^2 \mathbb{E}[\|R_1 M_0^{(K-j)M+t} G_{jM+i+1} B\zeta\|^2].
\end{aligned} \tag{5.7}
$$

Using the relation $H_{jM+M-1} = N_{jM+M-1}$ and (5.7) leads to

$$
\begin{aligned}
\mathrm{I}_{3,j} &= c_0^4 \sum_{i=1}^{M-2} \sum_{t=0}^{i-1} \mathbb{E}[\|R_1 M_0^{(K-j)M+t} H_{jM+i}(\zeta_{jM+i-1-t} - \zeta)\|^2] \\
&\quad + c_0^4 \sum_{t=0}^{M-2} \mathbb{E}[\|R_1 M_0^{(K-j)M+t} N_{jM+M-1}(\zeta_{jM+M-2-t} - \zeta)\|^2] \\
&= (n-1) c_0^4 \sum_{i=1}^{M-2} \sum_{t=0}^{i-1} \mathbb{E}[\|R_1 M_0^{(K-j)M+t} H_{jM+i}\zeta\|^2] \\
&\quad + (n-1)^2 c_0^4 \sum_{t=0}^{M-2} \mathbb{E}[\|R_1 M_0^{(K-j)M+t} B\zeta\|^2].
\end{aligned}
$$

Let $\mathrm{II}_{j,i,t} = \mathbb{E}[\|R_1 M_0^{(K-j)M+t} H_{jM+i}\zeta\|^2]$, and $\mathrm{II}_{j,0,t} = \mathbb{E}[\|R_1 M_0^{(K-j)M+t} B\zeta\|^2]$. Similarly, with the identities (5.5) and (5.6), we deduce

$$\begin{aligned}
\mathrm{I}_{1,j} &\leqslant 2(M-1)^2\|B\|^2 c_B' c_0^2 \mathrm{I}_{2,j} + 4c_0^4 \sum_{i=1}^{M-2} i(1-c_0\|B\|)^{-2i} \\
&\quad \times \sum_{t=0}^{i-1} \mathbb{E}[\|R_1 M_0^{(K-j)M+t} H_{jM+i}\zeta\|^2] \\
&\quad + 4c_B' c_0^4 (M-1) \sum_{t=0}^{M-2} \mathbb{E}[\|R_1 M_0^{(K-j)M+t} N_{jM+M-1}\zeta\|^2] \\
&\quad + 4c_0^4 \sum_{i=1}^{M-1} i(1-c_0\|B\|)^{-2i} \\
&\quad \times \sum_{t=0}^{i-1} \mathbb{E}[\|R_1 M_0^{(K-j+1)M+t-i-1} B(\zeta_{jM+i} - \zeta)\|^2] \\
&\leqslant 2(M-1)^2\|B\|^2 c_B' c_0^2 \mathrm{I}_{2,j} + 4(M-2)c_B c_0^4 \sum_{i=1}^{M-2}\sum_{t=0}^{i-1} \mathrm{II}_{j,i,t} \\
&\quad + 4(n-1)(M-1)c_B' c_0^4 \sum_{t=0}^{M-2} \mathbb{E}[\|R_1 M_0^{(K-j)M+t} B\zeta\|^2] \\
&\quad + 4(n-1)c_B' c_0^4 \sum_{i=1}^{M-1} i \sum_{t=0}^{i-1} \mathbb{E}[\|R_1 M_0^{(K-j+1)M+t-i-1} B\zeta\|^2].
\end{aligned}$$

Note that $\|M_0^{M-i-1}\|^2 \leqslant 1$ for any $1 \leqslant i \leqslant M-1$. The last two terms on the right-hand side of the inequality, denoted by II, can be bounded by

$$\begin{aligned}
\mathrm{II} &\leqslant 4(n-1)c_B' c_0^4 \left((M-1)\sum_{t=0}^{M-2} + \sum_{i=1}^{M-1} i \sum_{t=0}^{i-1} \right) \mathbb{E}[\|R_1 M_0^{(K-j)M+t} B\zeta\|^2] \\
&= 4(n-1)c_B' c_0^4 \sum_{t=0}^{M-2} \left(M - 1 + \sum_{i=t+1}^{M-1} i \right) \mathrm{II}_{j,0,t} \\
&\leqslant 2(n-1)(M+1)^2 c_B' c_0^4 \sum_{t=0}^{M-2} \mathrm{II}_{j,0,t},
\end{aligned}$$

since $M - 1 + \sum_{i=t+1}^{M-1} i \leqslant \frac{1}{2}(M+1)^2$, for $0 \leqslant t \leqslant M-2$. Consequently,

$$\begin{aligned}
\mathrm{I}_{1,j} &\leqslant 2(M-1)^2\|B\|^2 c_B' c_0^2 \mathrm{I}_{2,j} + 4(M-2)c_B' c_0^4 \sum_{i=1}^{M-2}\sum_{t=0}^{i-1} \mathrm{II}_{j,i,t} \\
&\quad + 2(n-1)(M+1)^2 c_B' c_0^4 \sum_{t=0}^{M-2} \mathrm{II}_{j,0,t}.
\end{aligned}$$

Now the condition (2.3) implies (5.4), which shows the induction step and completes the proof of the theorem. □

Remark 5.1. For exact data, i.e., $\delta = 0, \zeta = 0, \zeta_i = 0$ for any $i \geqslant 0$, the comparative analysis can be greatly simplified. Indeed, setting $R_1 = I$ and $R_2 = 0$ in the analysis leads to

$$\mathbb{E}[\|e_{(K+1)M}\|^2] \leqslant \mathrm{I}_0 + \sum_{j=0}^{K} \mathrm{I}_{1,j},$$

with

$$\mathrm{I}_0 = \|M_0^{(K+1)M} e_0\|^2 \quad \text{and} \quad \mathrm{I}_{1,j} = c_0^2 \sum_{i=1}^{M-1} \mathbb{E}[\|M_0^{(K-j)M} H_{jM+i}(I - M_0^i) e_{jM}\|^2].$$

Straightforward computation with lemma 3.3 gives

$$\begin{aligned}\mathrm{I}_{1,j} &\leqslant (n-1)c_0^4 \sum_{i=1}^{M-1} i^2 \mathbb{E}[\|M_0^{(K-j)M} G_{jM+i+1} B^2 e_{jM}\|^2] \\ &\leqslant (n-1)(M-1)^2 c_B' c_0^4 \|B\|^2 \sum_{i=1}^{M-1} \mathbb{E}[\|M_0^{(K-j)M+i} G_{jM+i+1} B e_{jM}\|^2].\end{aligned}$$

Similarly, lemma 5.1 with $R_1 = I$ and $R_2 = 0$ implies

$$\mathbb{E}[\|\hat{e}_{(K+1)M}\|^2] = \mathrm{I}_0 + \sum_{j=0}^{K} \mathrm{I}_{2,j},$$

with

$$\begin{aligned}\mathrm{I}_{2,j} &= c_0^2 \sum_{i=0}^{M-1} \mathbb{E}[\|M_0^{(K-j)M} H_{jM+i} M_0^i \hat{e}_{jM}\|^2] \\ &= (n-1)c_0^2 \sum_{i=0}^{M-1} \mathbb{E}[\|M_0^{(K-j)M+i} G_{jM+i+1} B \hat{e}_{jM}\|^2].\end{aligned}$$

When $c_0\|B\|(M-1) \leqslant (1 - c_0\|B\|)^{(M-1)}$, the conditions for the optimal convergence rate of SVRG is weaker than that of SGD. With $c = c_0\|B\|(M-1)$ and $c_1 = (1 - c(M-1)^{-1})^{(M-1)}$, the conditions can be satisfied if $c \leqslant c_1$. This short analysis clearly shows the beneficial effect of VR on the variance of the iterates x_k^δ, and hence SVRG allows larger step size while maintaining the optimal convergence.

6. Numerical experiments and discussions

In this section, we provide numerical experiments to complement the theoretical findings in section 2. The experimental setting is identical with that in [18]. Specifically, we employ three academic examples, i.e., `s-phillips` (mildly ill-posed), `s-gravity` (severely ill-posed) and `s-shaw` (severely ill-posed), generated from `phillips`, `gravity` and `shaw`, taken from the `MATLAB` package Regutools [10] (available at http://people.compute.dtu.dk/pcha/Regutools/, last accessed on 20 August 2020), all of size $n = m = 1000$. To explicitly control

the regularity index ν in assumption 2.1(b), we generate $x^\dagger$ by $x^\dagger = \|(A^tA)^\nu x_e\|_{\ell^\infty}^{-1}(A^tA)^\nu x_e$, where x_e is the exact solution given by the package, and $\|\cdot\|_{\ell^\infty}$ denotes the Euclidean maximum norm. The index ν in assumption 2.1(b) is slightly larger than the one defined above. The corresponding exact data $y^\dagger$ is given by $y^\dagger = Ax^\dagger$ and the noise data y^δ generated by

$$y_i^\delta := y_i^\dagger + \epsilon\|y^\dagger\|_{\ell^\infty}\xi_i, \quad i = 1, \ldots, n,$$

where ξ_is follow the standard Gaussian distribution, and $\epsilon > 0$ is the relative noise level. The maximum number of epochs is fixed at 9×10^5, where one epoch refers to $\frac{nM}{n+M}$ SVRG iterations or n SGD iterations so that the computational complexity of each method is comparable. All statistical quantities are computed from 100 runs. We present also numerical results for the LM [7, chapter 6] (with a step size $\|A\|^{-2}$), since it enjoys order optimality. All methods are initialized with $x_0 = 0$.

The accuracy of the reconstructions is measured by the mean squared errors $e_{\text{svrg}} = \mathbb{E}[\|x_{k_*}^\delta - x^\dagger\|^2]$, $e_{\text{sgd}} = \mathbb{E}[\|\hat{x}_{k_*}^\delta - x^\dagger\|^2]$ for SVRG and SGD, respectively, and the squared error $e_{\text{lm}} = \|x_{k_*}^\delta - x^\dagger\|^2$ for LM. The stopping index k_* (measured in epoch count) is taken such that the error is smallest along the respective iteration trajectory, due to a lack of rigorous *a posteriori* stopping rules for SVRG and SGD (the discrepancy principle is indeed convergent for SGD, without a rate [15]). The constant c in the step size c_0 is $c = (\max_i(\|a_i\|^2))^{-1}$, so that $c_0 = \mathcal{O}(cM^{-1})$ for SVRG and $c_0 = \mathcal{O}(cn^{-1})$ for SGD.

6.1. Numerical results for general A

The numerical results for the three examples with different regularity index ν and different noise levels are shown in tables 1–3, where the employed constant step size is determined in order to achieve optimal convergence (while maintaining good computational efficiency). For each fixed regularity index ν, all the errors e_{svrg}, e_{sgd} and e_{lm} decrease to zero as the (relative) noise level ϵ tends to zero with a certain rate, and the precise convergence rate depends on the index ν roughly as the theoretical prediction $\mathcal{O}(\delta^{\frac{4\nu}{2\nu+1}})$ (cf theorem 2.1 for SVRG, and remark 2.1 for SGD). Generally a larger ν leads to a faster convergence with respect to δ as the theory indicates, but the required number of iterations to reach the optimal error may not necessarily decrease, due to the use of smaller step sizes. The latter contrasts sharply with that for LM, for which a smoother exact solution $x^\dagger$ requires fewer iterations to reach optimal accuracy (when δ is fixed). Note that for both SVRG and SGD, optimal convergence holds only for a sufficiently small step size, and otherwise they suffer from the undesirable saturation phenomenon, i.e., the error decay may saturate when the index ν exceeds a certain value, which also concurs with the observation for SGD in [15, 18].

Now we examine more closely the convergence behavior of the SVRG iterates, and compare it with that of SGD and LM. For all these three examples and all ν values, both SVRG and SGD can achieve an accuracy comparable with that by LM, thereby achieving the order optimality of these methods, when the step size c_0 for SVRG and SGD is taken to be of order $\mathcal{O}(M^{-1})$ and $\mathcal{O}(n^{-1})$, respectively. This observation agrees well with the analysis in theorem 2.1. Generally, the larger the index ν is, the smaller the value c_0 should be taken in order to achieve the optimal rate. This can also be seen partly from the constant $2^{2\nu}c_\nu$ in the error bound in theorem 2.1. Next we discuss the computational complexity. For all three examples, SVRG takes fewer epochs to reach the optimal error than SGD for a large index ν, and LM requires fewest iterations among the three methods. For small ν, SVRG stops earlier than LM, and can be faster than SGD for suitably chosen c_0 (see, e.g., the case $\nu = 0$ in table 1). These empirical observations agree with the fact that SVRG hybridizes SGD and LM. Since in practice the index ν is rarely known, SVRG is an excellent choice, due to its low sensitivity with respect to ν.

Table 1. Comparison between SVRG (with $M = 100$), SGD and LM for `s-phillips`.

Method		SVRG			SGD			LM	
ν	ϵ	c_0	e_{svrg}	k_{svrg}	c_0	e_{sgd}	k_{sgd}	e_{lm}	k_{lm}
0	1×10^{-3}	$5c/M$	1.67×10^{-2}	4134.35	$4c/n$	1.66×10^{-2}	4691.28	1.65×10^{-2}	5851
	1×10^{-2}	$5c/M$	1.31×10^{-1}	180.95	$4c/n$	1.29×10^{-1}	204.90	1.28×10^{-1}	249
	5×10^{-2}	$5c/M$	5.42×10^{-1}	96.25	$4c/n$	5.42×10^{-1}	108.90	5.34×10^{-1}	136
1	1×10^{-3}	$1.5c/M$	3.31×10^{-4}	430.65	c/n	3.48×10^{-4}	539.19	2.28×10^{-4}	157
	1×10^{-2}	$1.5c/M$	5.96×10^{-3}	41.25	c/n	6.64×10^{-3}	57.81	5.12×10^{-3}	16
	5×10^{-2}	$1.5c/M$	3.22×10^{-2}	21.45	c/n	3.52×10^{-2}	29.40	3.16×10^{-2}	8
2	1×10^{-3}	$c/(2M)$	7.16×10^{-5}	155.10	$c/(30n)$	7.02×10^{-5}	2115.54	3.22×10^{-5}	19
	1×10^{-2}	$c/(2M)$	1.07×10^{-3}	68.75	$c/(30n)$	1.09×10^{-3}	938.70	9.82×10^{-4}	8
	5×10^{-2}	$c/(2M)$	2.90×10^{-2}	46.75	$c/(30n)$	2.92×10^{-2}	636.51	1.57×10^{-2}	5
4	1×10^{-3}	$c/(5M)$	3.05×10^{-5}	202.95	$c/(30n)$	9.77×10^{-5}	1966.38	1.30×10^{-5}	8
	1×10^{-2}	$c/(5M)$	2.41×10^{-3}	142.45	$c/(30n)$	2.56×10^{-3}	785.94	1.42×10^{-3}	5
	5×10^{-2}	$c/(5M)$	5.20×10^{-2}	110.00	$c/(30n)$	5.23×10^{-2}	596.73	2.49×10^{-2}	3

Table 2. Comparison between SVRG (with $M = 100$), SGD and LM for `s-gravity`.

Method		SVRG			SGD			LM	
ν	ϵ	c_0	e_{svrg}	k_{svrg}	c_0	e_{sgd}	k_{sgd}	e_{lm}	k_{lm}
0	1×10^{-3}	$c/10$	9.50×10^{-2}	5495.05	$c/20$	9.37×10^{-2}	1000.50	9.39×10^{-2}	27201
	1×10^{-2}	$c/10$	5.98×10^{-1}	217.80	$c/20$	5.81×10^{-1}	34.11	5.73×10^{-1}	793
	5×10^{-2}	$c/10$	2.16×10^{0}	35.75	$c/20$	2.23×10^{0}	5.61	2.07×10^{0}	149
1	1×10^{-3}	$c/(5M)$	5.78×10^{-4}	1019.15	$c/(30n)$	5.90×10^{-4}	5604.80	5.68×10^{-4}	99
	1×10^{-2}	$c/(5M)$	1.14×10^{-2}	246.40	$c/(30n)$	1.15×10^{-2}	1356.87	1.12×10^{-2}	24
	5×10^{-2}	$c/(5M)$	6.47×10^{-2}	112.20	$c/(30n)$	6.48×10^{-2}	613.41	6.19×10^{-2}	11
2	1×10^{-3}	$c/(10M)$	7.57×10^{-5}	474.10	$c/(50n)$	1.32×10^{-4}	2441.85	6.82×10^{-5}	23
	1×10^{-2}	$c/(10M)$	1.80×10^{-3}	229.90	$c/(50n)$	1.92×10^{-3}	1047.03	1.47×10^{-3}	10
	5×10^{-2}	$c/(10M)$	2.32×10^{-2}	156.75	$c/(50n)$	2.35×10^{-2}	708.72	1.61×10^{-2}	6
4	1×10^{-3}	$c/(10M)$	2.51×10^{-5}	250.80	$c/(60n)$	1.03×10^{-4}	2212.26	1.30×10^{-5}	10
	1×10^{-2}	$c/(10M)$	1.14×10^{-3}	170.50	$c/(60n)$	1.29×10^{-3}	941.19	6.42×10^{-4}	6
	5×10^{-2}	$c/(10M)$	2.23×10^{-2}	138.05	$c/(60n)$	2.25×10^{-2}	746.67	8.58×10^{-3}	3

Table 3. Comparison between SVRG (with $M = 100$), SGD and LM for `s-shaw`.

Method		SVRG			SGD			LM	
ν	ϵ	c_0	e_{svrg}	k_{svrg}	c_0	e_{sgd}	k_{sgd}	e_{lm}	k_{lm}
0	1×10^{-3}	c	2.81×10^{-1}	30246.15	c	2.81×10^{-1}	2704.92	2.81×10^{-1}	760 983
	1×10^{-2}	c	6.92×10^{-1}	503.25	c	7.08×10^{-1}	42.42	6.67×10^{-1}	12 385
	5×10^{-2}	c	3.01×10^{0}	139.15	c	3.91×10^{0}	10.59	2.91×10^{0}	3392
1	1×10^{-3}	c/M	6.80×10^{-5}	579.15	$c/(2n)$	7.05×10^{-5}	1047.60	5.95×10^{-5}	144
	1×10^{-2}	c/M	5.35×10^{-3}	222.75	$c/(2n)$	5.42×10^{-3}	394.00	5.21×10^{-3}	54
	5×10^{-2}	c/M	1.50×10^{-1}	148.50	$c/(2n)$	1.50×10^{-1}	271.00	1.47×10^{-1}	36
2	1×10^{-3}	$c/(2M)$	6.94×10^{-5}	434.50	$c/(20n)$	7.08×10^{-5}	4147.00	6.36×10^{-5}	50
	1×10^{-2}	$c/(2M)$	5.80×10^{-3}	246.95	$c/(20n)$	5.80×10^{-3}	2242.50	5.71×10^{-3}	30
	5×10^{-2}	$c/(2M)$	7.84×10^{-2}	52.80	$c/(20n)$	7.79×10^{-2}	480.80	7.08×10^{-2}	5
4	1×10^{-3}	$c/(4M)$	3.83×10^{-5}	184.25	$c/(30n)$	5.79×10^{-5}	1966.38	3.13×10^{-5}	9
	1×10^{-2}	$c/(4M)$	1.96×10^{-3}	121.55	$c/(30n)$	1.99×10^{-3}	828.45	1.01×10^{-3}	4
	5×10^{-2}	$c/(4M)$	3.61×10^{-2}	95.15	$c/(30n)$	3.61×10^{-2}	645.75	6.45×10^{-3}	1

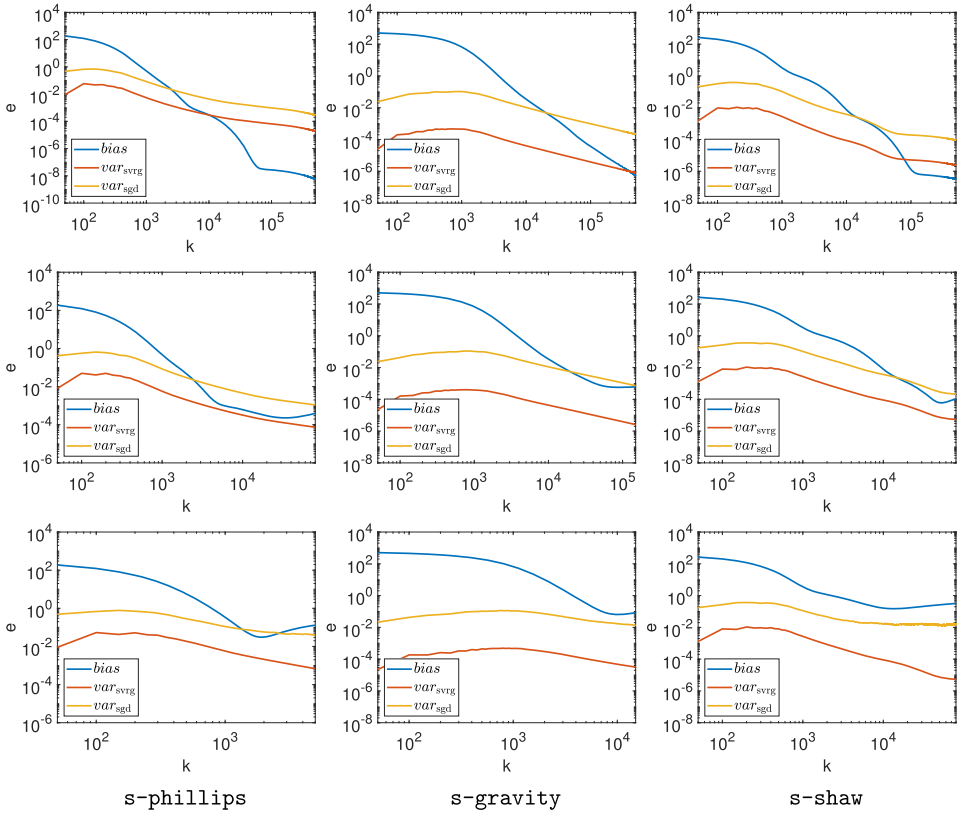

Figure 1. The convergence of the bias or variance with generic term e versus iteration number for the examples with $\nu = 1$. The rows from top to bottom rows are for $\epsilon = 0$, $\epsilon = 1 \times 10^{-3}$ and $\epsilon = 5 \times 10^{-2}$, respectively.

To verify the analysis in section 5, we examine the bias $\text{bias} = \|\mathbb{E}[x_k^\delta] - x^\dagger\|^2 = \|\mathbb{E}[\hat{x}_k^\delta] - x^\dagger\|^2$, and the variances $\text{var}_{\text{svrg}} = \mathbb{E}[\|x_k^\delta - \mathbb{E}[x_k^\delta]\|^2]$ and $\text{var}_{\text{sgd}} = \mathbb{E}[\|\hat{x}_k^\delta - \mathbb{E}[\hat{x}_k^\delta]\|^2]$. The numerical results are shown in figure 1, for the examples with $\nu = 1$, with the step size c_0 for SVRG used for both methods. Although not presented, we note that any other suitable c_0 under condition (2.3) leads to nearly identical observations. Note that the iteration index k in the figures refers to the exact number of iterations (not counted in epoch), to facilitate the comparison of the convergence behavior. For both exact and noisy data, when the iteration number k is fixed, the SVRG variance var_{svrg} is always orders of magnitude smaller than the SGD variance var_{sgd}, which is fully in line with theorem 2.2. This shows clearly the role of the VR effect, which in particular allows using larger step size. Note that the frequency $M = 100$ is selected by the condition (2.2) for optimal accuracy, but actually does not satisfy condition (2.3). Nonetheless, we still observe the assertion in theorem 2.2.

Further, in the experiments, bias (which is equal to the error e_{lm} of LM) is always much larger than the SVRG variance var_{svrg} (of similar magnitude during a few iterations before stopping), and thus the variance has little influence on the optimal accuracy, especially for noisy data. In contrast, the SGD variance var_{sgd} dominates the error sometimes and causes the undesirable saturation phenomenon. These observations also agree with theorem 2.1, which states that the saturation of SVRG does not exist by choosing suitable frequency M and initial

Table 4. SVRG with different M for `s-phillips`.

		$\nu = 0$			$\nu = 2$		
M	ϵ	c_0	e	k	c_0	e	k
$0.1n$	1×10^{-3}	$5c/M$	1.67×10^{-2}	4134.35	$c/(2M)$	7.16×10^{-5}	155.10
	1×10^{-2}	$5c/M$	1.31×10^{-1}	180.95	$c/(2M)$	1.07×10^{-3}	68.75
	5×10^{-2}	$5c/M$	5.42×10^{-1}	96.80	$c/(2M)$	2.90×10^{-2}	46.75
$0.5n$	1×10^{-3}	$5c/M$	1.66×10^{-2}	5650.35	$c/(2M)$	4.18×10^{-5}	204.30
	1×10^{-2}	$5c/M$	1.31×10^{-1}	125.70	$c/(2M)$	9.90×10^{-4}	93.30
	5×10^{-2}	$5c/M$	5.40×10^{-1}	66.15	$c/(2M)$	2.90×10^{-2}	63.75
n	1×10^{-3}	$10c/M$	1.67×10^{-2}	3757.40	c/M	5.83×10^{-5}	139.50
	1×10^{-2}	$10c/M$	1.29×10^{-1}	163.80	c/M	1.04×10^{-3}	62.20
	5×10^{-2}	$10c/M$	5.38×10^{-1}	87.40	c/M	2.92×10^{-2}	42.50
$2n$	1×10^{-3}	$15c/M$	1.67×10^{-2}	3781.35	$1.5c/M$	7.63×10^{-5}	144.38
	1×10^{-2}	$15c/M$	1.30×10^{-1}	164.70	$1.5c/M$	1.08×10^{-3}	62.25
	5×10^{-2}	$15c/M$	5.39×10^{-1}	87.08	$1.5c/M$	2.93×10^{-2}	42.53
$5n$	1×10^{-3}	$25c/M$	1.66×10^{-2}	4519.86	$2c/M$	7.33×10^{-5}	214.32
	1×10^{-2}	$25c/M$	1.29×10^{-1}	197.28	$2c/M$	1.05×10^{-3}	93.60
	5×10^{-2}	$25c/M$	5.40×10^{-1}	104.64	$2c/M$	2.90×10^{-2}	63.84

step size c_0. They also confirm the theoretical prediction in remark 5.1, i.e., the condition for the optimality of SVRG is weaker than that of SGD, partly concurring with theorem 2.2. These empirical observations show clearly the beneficial effect of incorporating VR into stochastic iterative methods from the perspective of regularization theory.

6.2. Influence of M

SVRG involves one free parameter, the frequency M of evaluating the full gradient. Clearly, the parameter M will influence the overall computational efficiency of SVRG: ideally one would like to make it as large as possible, but a too large M would bring too little VR into SGD iteration. The theoretical analysis in this work indicates that SVRG can achieve optimal convergence rates when $M \geqslant \mathcal{O}(n^{\frac{1}{2}})$ (cf remark 2.1), and that $M \leqslant \mathcal{O}(n^{\frac{1}{2}})$ is sufficient for ensuring the SVRG variance smaller than SGD variance (cf remark 2.2). Nonetheless, a complete theoretical analysis of the influence of the frequency M on the performance of SVRG is still unknown. To gain insight, we present the numerical results for `s-phillips` with noisy data by SVRG with different M ranging from $0.1n$ to $5n$ in table 4. Note that the choices $2n$ and $5n$ were recommended for convex and nonconvex optimization problems, respectively [19]. The numerical results indicate that SVRG with all these frequencies can actually achieve an accuracy comparable with that by the LM when the constant step size is chosen suitably. In general, a larger M requires smaller step sizes in order to maintain the optimal convergence rate, agreeing well with the theoretical analysis in section 4. Interestingly, the overall computational complexity for these different M does not vary too much. Thus, the choice of M within a certain range actually has little impact on the performance of SVRG. Although not presented, the same observations can be drawn from the numerical results for the examples `s-shaw` and `s-gravity`.

Table 5. Comparison between SVRG (with $M = 100$) for `s-phillips` with A and $\tilde{A}$.

Method			SVRG with A		SVRG with $\tilde{A}$	
ν	ϵ	c_0	e	k	e	k
0	1×10^{-3}	$5c/M$	1.67×10^{-2}	4134.35	1.65×10^{-2}	4129.40
	1×10^{-2}	$5c/M$	1.31×10^{-1}	180.95	1.28×10^{-1}	176.55
	5×10^{-2}	$5c/M$	5.42×10^{-1}	96.80	5.36×10^{-1}	96.25
1	1×10^{-3}	$1.5c/M$	3.31×10^{-4}	430.65	2.29×10^{-4}	372.35
	1×10^{-2}	$1.5c/M$	5.96×10^{-3}	41.25	5.32×10^{-3}	40.70
	5×10^{-2}	$1.5c/M$	3.22×10^{-2}	21.45	3.17×10^{-2}	20.90
2	1×10^{-3}	$c/(2M)$	7.16×10^{-5}	155.10	3.49×10^{-5}	148.50
	1×10^{-2}	$c/(2M)$	1.07×10^{-3}	68.75	9.77×10^{-4}	68.75
	5×10^{-2}	$c/(2M)$	2.90×10^{-2}	46.75	2.89×10^{-2}	46.75
4	1×10^{-3}	$c/(5M)$	3.05×10^{-5}	202.95	2.46×10^{-5}	201.30
	1×10^{-2}	$c/(5M)$	2.41×10^{-3}	142.45	2.41×10^{-3}	142.45
	5×10^{-2}	$c/(5M)$	5.20×10^{-2}	110.00	5.21×10^{-2}	110.00

6.3. On assumption 2.1(c)

Assumption 2.1(c) is crucial to the analysis in sections 4 and 5. It is natural to ask whether the assumption is actually necessary. We examine the issue numerically as follows. Let $A = U\Sigma V^t$ be the SVD of A, and $\tilde{A}$ by $\tilde{A} = U^t A$, and then replace A in (1.1) by $\tilde{A}$ and y^δ by $\tilde{y}^\delta = U^t y^\delta$. Then preconditioned system $\tilde{A}x = \tilde{y}^\delta$ satisfies assumption 2.1(c). The numerical results for `s-phillips` are shown in table 5, and the trajectories of e_k^δ for the examples with $\nu = 1$ in figure 2. It is observed that for noisy data, the SVRG results for A and $\tilde{A}$ are nearly identical with each other in terms of the accuracy, stopping index, and convergence trajectory. For exact data (cf the top row of figure 2), the trajectories overlap up to a certain point around 1×10^{-3} for `s-phillips` and 1×10^{-5} for `s-gravity` and `s-shaw`, which can be further decreased by choosing smaller c_0. These observations resemble closely the empirical observations for SGD, see, especially figure 4.3 of [18]. Thus, assumption 2.1(c) is probably due to a limitation of the proof technique, and there might be alternative proof strategies that circumvent the restriction.

Data availability statement

All data that support the findings of this study are included within the article (and any supplementary files).

Appendix A. Technical proofs

In this appendix, we collect the proofs of several technical estimates.

A.1. Proof of lemma 3.4

The proof relies on spectral decomposition. Let $\mathrm{Sp}(B)$ be the spectrum of B. Then by direct computation, we have

$$c_0^s \|B^s M_0^{KM}\| = c_0^s \sup_{\lambda \in \mathrm{Sp}(B)} |\lambda^s (1 - c_0\lambda)^{KM}| \leqslant \sup_{a \in [0,1]} a^s (1-a)^{KM}.$$

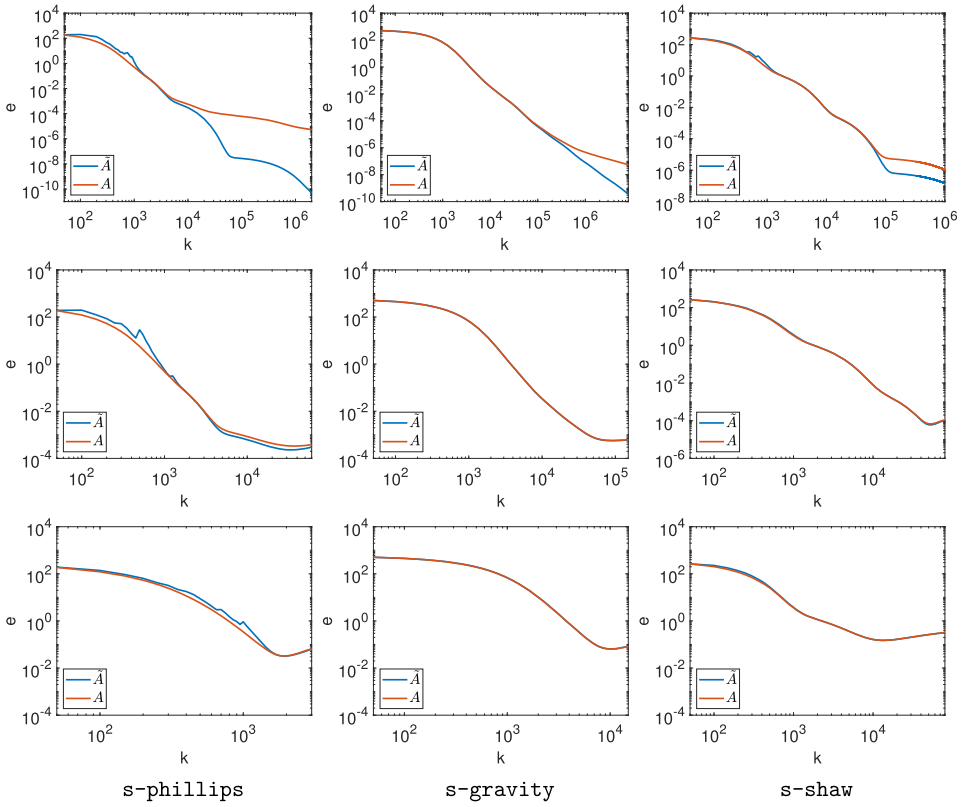

Figure 2. The convergence of the error e versus iteration number for the examples with $\nu = 1$, computed using A and $\tilde{A}$. The rows from top to bottom rows are for $\epsilon = 0$, $\epsilon = 1 \times 10^{-3}$ and $\epsilon = 5 \times 10^{-2}$, respectively.

Let $g(a) = a^s(1-a)^{KM}$. Then $g'(a) = (s(1-a) - KMa)\, a^{s-1}(1-a)^{KM-1}$, so that $g(a)$ achieves its maximum over the interval $[0,1]$ at $a^* = s(s+KM)^{-1}$. Consequently,

$$c_0^s \|B^s M_0^{KM}\| \leqslant g(a^*) = \left(\frac{KM}{s+KM}\right)^{s+KM} s^s (KM)^{-s} \leqslant s^s M^{-s} K^{-s}.$$

This shows the second estimate. Similarly,

$$\begin{aligned} c_0^{-t}\|B^{-t}(I - M_0^{KM})\| &= \sup_{\lambda \in \mathrm{Sp}(B)} |(c_0\lambda)^{-t}(1-(1-c_0\lambda)^{KM})| \\ &\leqslant \sup_{a\in[0,1]} a^{-t}(1-(1-a)^{KM}). \end{aligned}$$

Note that for any $a \in [0,1]$, there holds $1-(1-a)^{KM} \leqslant 1$, and $\min_{t\in[0,1]}(aKM)^t = \min(aKM, 1)$, since $(aKM)^t$ is monotone with respect to t. Let $h(a) := aKM - (1-(1-a)^{KM})$ which is increasing over $[0,1]$, that implies $h(a) \geqslant h(0) = 0$. Thus

$$1-(1-a)^{KM} \leqslant \min(aKM, 1) \leqslant (aKM)^t.$$

This shows the first estimate and completes the proof of the lemma.

A.2. Proof of proposition 3.1

To prove proposition 3.1, we first give a representation of the (epochwise) SVRG iterate x_{KM}^{δ}.

Lemma A.1. *The following recursion holds for any* $K \geqslant 0$,

$$e_{(K+1)M}^{\delta} = (M_0^M - L_K B)e_{KM}^{\delta} + \left(c_0 \sum_{i=0}^{M-1} M_0^i + L_K\right)\zeta, \tag{A.1}$$

where the random matrix L_K *is given by*

$$L_K = c_0 \sum_{i=1}^{M-1} H_{KM+i}(I - M_0^i)B^{-1}. \tag{A.2}$$

Proof. Note that the SVRG iterate x_{k+1}^{δ}, $k = 0, 1, \ldots$, can be rewritten as

$$\begin{aligned} x_{k+1}^{\delta} &= x_k^{\delta} - c_0\left((a_{i_k}, e_k^{\delta} - e_{k_M}^{\delta})a_{i_k} + Be_{k_M}^{\delta} - \zeta\right) \\ &= x_k^{\delta} - c_0 a_{i_k} a_{i_k}^t (e_k^{\delta} - e_{k_M}^{\delta}) - c_0(Be_{k_M}^{\delta} - \zeta). \end{aligned}$$

Using the definitions of P_k and N_k, the error $e_k^{\delta} \equiv x_k^{\delta} - x^{\dagger}$ of the SVRG iterate x_k^{δ} satisfies

$$e_{k+1}^{\delta} = (I - c_0 a_{i_k} a_{i_k}^t)e_k^{\delta} + c_0(a_{i_k} a_{i_k}^t - B)e_{k_M}^{\delta} + c_0\zeta = P_k e_k^{\delta} - c_0 N_k e_{k_M}^{\delta} + c_0\zeta. \tag{A.3}$$

For any $K \geqslant 0$, it follows from (A.3) and direct computation that

$$e_{KM+1}^{\delta} = P_{KM}e_{KM}^{\delta} - c_0 N_{KM}e_{KM}^{\delta} + c_0\zeta = M_0 e_{KM}^{\delta} + c_0\zeta. \tag{A.4}$$

Meanwhile, setting $k = (K+1)M - 1$ in the recursion (A.3), then repeatedly applying the recursion (A.3) and using the definitions of the matrices G_k and H_k lead to

$$\begin{aligned} e_{(K+1)M}^{\delta} &= P_{(K+1)M-1}e_{(K+1)M-1}^{\delta} - c_0 N_{(K+1)M-1}e_{KM}^{\delta} + c_0\zeta \\ &= G_{(K+1)M-2}e_{(K+1)M-2}^{\delta} - c_0\left(P_{(K+1)M-1}N_{(K+1)M-2}\right. \\ &\quad \left. + N_{(K+1)M-1}\right)e_{KM}^{\delta} + c_0(P_{(K+1)M-1} + I)\zeta \\ &= \cdots = G_{KM+1}e_{KM+1}^{\delta} - c_0\sum_{i=1}^{M-1} H_{KM+i}e_{KM}^{\delta} + c_0\sum_{i=2}^{M} G_{KM+i}\zeta. \end{aligned}$$

This identity and (A.4) imply that for any $K \geqslant 0$,

$$e_{(K+1)M}^{\delta} = \left(G_{KM+1}M_0 - c_0\sum_{i=1}^{M-1} H_{KM+i}\right)e_{KM}^{\delta} + \left(c_0\sum_{i=1}^{M} G_{KM+i}\right)\zeta.$$

Next we simplify the two terms in the brackets using the identity (3.3). It follows directly from (3.3) that

$$G_{KM+1}M_0 - c_0\sum_{i=1}^{M-1} H_{KM+i} = M_0^M - c_0\sum_{i=1}^{M-1} H_{KM+i}(I - M_0^i).$$

Similarly, by the identity (3.3), we deduce

$$\begin{aligned} c_0\sum_{i=1}^{M} G_{KM+i} &= c_0 I + c_0\sum_{i=1}^{M-1}\left(M_0^{M-i} + c_0\sum_{j=0}^{M-i-1} H_{KM+i+j}M_0^j\right)\\ &= c_0\sum_{i=1}^{M} M_0^{M-i} + c_0^2\sum_{i=1}^{M-1}\sum_{j=0}^{M-i-1} H_{KM+i+j}M_0^j\\ &= c_0\sum_{i=0}^{M-1} M_0^i + c_0^2\sum_{i=1}^{M-1} H_{KM+i}\left(\sum_{j=0}^{i-1} M_0^j\right)\\ &= c_0\sum_{i=0}^{M-1} M_0^i + c_0\sum_{i=1}^{M-1} H_{KM+i}(I - M_0^i)B^{-1}, \end{aligned}$$

where the last line follows from the identity

$$c_0\sum_{i=0}^{j-1} M_0^i = (I - M_0^j)B^{-1}. \tag{A.5}$$

Combining the preceding identities completes the proof of the lemma. □

Now we can give the proof of proposition 3.1.

Proof. By the definitions of the matrices N_i and G_{i+1}, they are independent. Thus, there hold

$$\mathbb{E}[H_i] = \mathbb{E}[G_{i+1}]\mathbb{E}[N_i] = 0 \quad \text{and} \quad \mathbb{E}[L_j] = 0.$$

Then by lemma A.1, we have

$$\mathbb{E}[e^\delta_{(K+1)M}] = M_0^M\mathbb{E}[e^\delta_{KM}] + c_0\sum_{i=0}^{M-1} M_0^i\zeta.$$

Repeatedly applying this identity gives

$$\begin{aligned} \mathbb{E}[e^\delta_{(K+1)M}] &= M_0^M\left(M_0^M\mathbb{E}[e^\delta_{(K-1)M}] + c_0\sum_{i=0}^{M-1} M_0^i\zeta\right) + c_0\sum_{i=0}^{M-1} M_0^i\zeta\\ &= M_0^{2M}\mathbb{E}[e^\delta_{(K-1)M}] + c_0\sum_{i=0}^{2M-1} M_0^i\zeta = \cdots = M_0^{(K+1)M}e_0^\delta\\ &\quad + c_0\sum_{i=0}^{(K+1)M-1} M_0^i\zeta. \end{aligned}$$

This and the identity (A.5) show the expression for $\mathbb{E}[e^\delta_{KM}]$. Let $z_K := e^\delta_{KM} - \mathbb{E}[e^\delta_{KM}]$. Then for any $K \geqslant 0$, it follows from lemma A.1 that

$$z_{K+1} = M_0^M z_K + R_K, \quad \text{with } R_K := L_K(\zeta - Be^\delta_{KM}),$$

and $z_0 = 0$. Repeatedly applying the recursion directly gives

$$z_{K+1} = M_0^{(K+1)M} z_0 + \sum_{j=0}^{K} M_0^{jM} R_{K-j} = \sum_{j=0}^{K} M_0^{(K-j)M} R_j.$$

This completes the proof of the proposition. □

A.3. Proof of proposition 3.2

The following recursion is direct from the definition of SGD iteration in (1.3)

$$\hat{e}_{k+1}^{\delta} = (I - c_0 a_{i_k} a_{i_k}^t)\hat{e}_k^{\delta} + c_0 \xi_k a_{i_k} = P_k \hat{e}_k^{\delta} + c_0 \zeta_k.$$

Repeatedly applying the recursion and using the identity (3.3) (and its proof) yield that for any $K \geqslant 0$,

$$\begin{aligned}
\hat{e}_{(K+1)M}^{\delta} &= G_{KM+1} P_{KM} \hat{e}_{KM}^{\delta} + c_0 \sum_{i=0}^{M-1} G_{KM+i+1} \zeta_{KM+i} \\
&= \left(M_0^M + c_0 \sum_{i=0}^{M-1} H_{KM+i} M_0^i \right) \hat{e}_{KM}^{\delta} + c_0 \zeta_{(K+1)M-1} \\
&\quad + c_0 \sum_{i=1}^{M-1} \left(M_0^{M-i} + c_0 \sum_{t=0}^{M-i-1} H_{KM+i+t} M_0^t \right) \zeta_{KM+i-1}.
\end{aligned}$$

Since $\mathbb{E}[H_{KM+i}] = 0$, for $i = 0, \ldots, M-1$, and H_{KM+i+t}, $t \geqslant 0$, and ζ_{KM+i-1} are independent, by the identity (A.5),

$$\begin{aligned}
\mathbb{E}[\hat{e}_{(K+1)M}^{\delta}] &= M_0^M \mathbb{E}[\hat{e}_{KM}^{\delta}] + c_0 \sum_{i=0}^{M-1} M_0^i \zeta \\
&= M_0^{(K+1)M} \hat{e}_0^{\delta} + \left(I - M_0^{(K+1)M} \right) B^{-1} \zeta.
\end{aligned}$$

This gives the desired expression of $\mathbb{E}[\hat{x}_{KM}^{\delta}]$. Next, the variance component $\hat{e}_{(K+1)M}^{\delta} - \mathbb{E}[\hat{e}_{(K+1)M}^{\delta}]$ is given by

$$\begin{aligned}
\hat{e}_{(K+1)M}^{\delta} - \mathbb{E}[\hat{e}_{(K+1)M}^{\delta}] &= M_0^M (\hat{e}_{KM}^{\delta} - \mathbb{E}[\hat{e}_{KM}^{\delta}]) + c_0 \sum_{i=0}^{M-1} H_{KM+i} M_0^i \hat{e}_{KM}^{\delta} \\
&\quad + c_0 \sum_{i=1}^{M} M_0^{M-i} (\zeta_{KM+i-1} - \zeta) + c_0^2 \sum_{i=1}^{M-1} \sum_{t=0}^{M-i-1} H_{KM+i+t} M_0^t \zeta_{KM+i-1} \\
&= c_0 \sum_{j=0}^{K} \sum_{i=0}^{M-1} M_0^{(K-j)M} H_{jM+i} M_0^i \hat{e}_{jM}^{\delta} + c_0 \sum_{j=0}^{K} \sum_{i=1}^{M} M_0^{(K-j+1)M-i} (\zeta_{jM+i-1} - \zeta) \\
&\quad + c_0^2 \sum_{j=0}^{K} \sum_{i=1}^{M-1} \sum_{t=0}^{M-i-1} M_0^{(K-j)M} H_{jM+i+t} M_0^t (\zeta_{jM+i-1} - \zeta) \\
&\quad + c_0^2 \sum_{j=0}^{K} \sum_{i=1}^{M-1} \sum_{t=0}^{M-i-1} M_0^{(K-j)M} H_{jM+i+t} M_0^t \zeta.
\end{aligned}$$

Then it follows from the identity (A.5) that

$$c_0\sum_{i=1}^{M-1}\sum_{t=0}^{M-i-1} H_{jM+i+t}M_0^t = c_0\sum_{i=1}^{M-1} H_{jM+i}\left(\sum_{t=0}^{i-1} M_0^t\right)$$
$$= \sum_{i=1}^{M-1} H_{jM+i}(I - M_0^i)B^{-1}.$$

Finally we derive

$$\begin{aligned}
\hat{e}^\delta_{(K+1)M} - \mathbb{E}[\hat{e}^\delta_{(K+1)M}] &= c_0\sum_{j=0}^{K}\sum_{i=0}^{M-1} M_0^{(K-j)M}H_{jM+i}M_0^i\hat{e}^\delta_{jM} \\
&\quad + c_0\sum_{j=0}^{K}\sum_{i=0}^{M-1} M_0^{(K-j+1)M-i-1}(\zeta_{jM+i} - \zeta) \\
&\quad + c_0^2\sum_{j=0}^{K}\sum_{i=0}^{M-2}\sum_{t=0}^{M-i-2} M_0^{(K-j)M}H_{jM+i+t+1}M_0^t(\zeta_{jM+i} - \zeta) \\
&\quad + c_0\sum_{j=0}^{K}\sum_{i=1}^{M-1} M_0^{(K-j)M}H_{jM+i}(I - M_0^i)B^{-1}\zeta \\
&= c_0\sum_{j=0}^{K}\sum_{i=0}^{M-1} M_0^{(K-j)M}\left(H_{jM+i}\left(M_0^i\hat{e}^\delta_{jM}\right.\right. \\
&\quad \left.\left. + (I - M_0^i)B^{-1}\zeta\right) + M_0^{M-i-1}(\zeta_{jM+i} - \zeta)\right) \\
&\quad + c_0^2\sum_{j=0}^{K}\sum_{i=0}^{M-2}\sum_{t=0}^{M-i-2} M_0^{(K-j)M}H_{jM+i+t+1}M_0^t(\zeta_{jM+i} - \zeta).
\end{aligned}$$

This completes the proof of the proposition.

A.4. Proof of lemma 4.1

The proof employs the standard bias-variance decomposition and certain independence. By proposition 3.1, the following identities hold

$$\begin{aligned}
&\mathbb{E}[R_1(e^\delta_{(K+1)M} - B^{-1}\zeta) + R_2|\mathcal{F}^c_{(K+1)M}] \\
&\quad = R_1(M_0^{(K+1)M}e_0^\delta - B^{-1}\zeta) + R_2, \\
&R_1(e^\delta_{(K+1)M} - B^{-1}\zeta) + R_2 - \mathbb{E}[R_1(e^\delta_{(K+1)M} - B^{-1}\zeta) + R_2|\mathcal{F}^c_{(K+1)M}] \\
&\quad = R_1(e^\delta_{(K+1)M} - \mathbb{E}[e^\delta_{(K+1)M}]) = R_1\sum_{j=0}^{K} M_0^{(K-j)M}L_j(\zeta - Be^\delta_{jM}),
\end{aligned}$$

where the random matrices L_j are defined in (A.2). Then we claim the following identity for any $i, i' = 0, \ldots, M-1$,

$$\mathbb{E}[\langle H_{jM+i}e^\delta_{jM}, H_{j'M+i'}e^\delta_{j'M}\rangle] = 0, \quad \text{if } i \neq i' \text{ or } j \neq j'. \tag{A.6}$$

Clearly, it suffices to analyze the two cases $0 \leqslant i < i' \leqslant M-1$, and $j < j'$ and $0 \leqslant i, i' \leqslant M-1$ separately. Indeed, for any $0 \leqslant i < i' \leqslant M-1$, the random matrix N_{jM+i} is independent of $G_{jM+i+1}e^{\delta}_{jM}$ and $N_{jM+i'}G_{jM+i'+1}e^{\delta}_{jM}$. Thus, using the identity $\mathbb{E}_{jM+i}[N_{jM+i}] = 0$, for any $i = 0, \ldots, M-1$, we obtain

$$\begin{aligned}
&\mathbb{E}_{jM+i}[\langle H_{jM+i}e^{\delta}_{jM}, H_{jM+i'}e^{\delta}_{jM}\rangle] \\
&\quad = \mathbb{E}_{jM+i}[\langle N_{jM+i}G_{jM+i+1}e^{\delta}_{jM}, N_{jM+i'}G_{jM+i'+1}e^{\delta}_{jM}\rangle] \\
&\quad = \langle \mathbb{E}_{jM+i}[N_{jM+i}]G_{jM+i+1}e^{\delta}_{jM}, N_{jM+i'}G_{jM+i'+1}e^{\delta}_{jM}\rangle = 0.
\end{aligned}$$

Similarly, for any $j < j'$ and $0 \leqslant i, i' \leqslant M-1$, the random matrix $N_{j'M+i'}$ is independent of $N_{jM+i}G_{jM+i+1}e^{\delta}_{jM}$ and $G_{j'M+i'+1}e^{\delta}_{j'M}$, and hence

$$\begin{aligned}
&\mathbb{E}_{j'M+i'}[\langle H_{jM+i}e^{\delta}_{jM}, H_{j'M+i'}e^{\delta}_{j'M}\rangle] \\
&\quad = \mathbb{E}_{j'M+i'}[\langle N_{jM+i}G_{jM+i+1}e^{\delta}_{jM}, N_{j'M+i'}G_{j'M+i'+1}e^{\delta}_{j'M}\rangle] \\
&\quad = \langle N_{jM+i}G_{jM+i+1}e^{\delta}_{jM}, \mathbb{E}_{j'M+i'}[N_{j'M+i'}]G_{j'M+i'+1}e^{\delta}_{j'M}\rangle = 0.
\end{aligned}$$

The desired claim (A.6) follows by taking full conditional of the last two identities. Note that by assumption, R_1 is independent of $e^{\delta}_{(K+1)M} - \mathbb{E}[e^{\delta}_{(K+1)M}]$. Then the bias-variance decomposition and the claim (A.6) imply

$$\begin{aligned}
&\mathbb{E}[\mathbb{E}[\|R_1(e^{\delta}_{(K+1)M} - B^{-1}\zeta) + R_2\|^2|\mathcal{F}^c_{(K+1)M}]] \\
&\quad = \mathrm{I}_0 + \mathbb{E}\left[\|R_1\sum_{j=0}^{K} M_0^{(K-j)M}L^1_j(\zeta - Be^{\delta}_{jM})\|^2\right] \\
&\quad = \mathrm{I}_0 + c_0^2\sum_{j=0}^{K}\sum_{i=1}^{M-1}\mathbb{E}[\|R_1M_0^{(K-j)M}H_{jM+i}(I - M_0^i)(e^{\delta}_{jM} - B^{-1}\zeta)\|^2].
\end{aligned}$$

This and the definitions of the terms I_0 and $\mathrm{I}_{1,j}$ complete the proof of the lemma.

A.5. Proof of lemma 5.1

The proof of the lemma is similar to lemma 4.1, and employs suitable independence relation crucially. By proposition 3.2 and the standard bias-variance decomposition, we have

$$\mathbb{E}[\|R_1(\hat{e}^{\delta}_{(K+1)M} - B^{-1}\zeta) + R_2\|^2] = \mathrm{I}_0 + \mathbb{E}[\|R_1(\hat{e}^{\delta}_{(K+1)M} - \mathbb{E}[\hat{e}^{\delta}_{(K+1)M}])\|^2],$$

with

$$\hat{e}^{\delta}_{(K+1)M} - \mathbb{E}[\hat{e}^{\delta}_{(K+1)M}] := \sum_{j=0}^{K}\sum_{i=0}^{M-1} d_{j,i},$$

where $d_{j,i}$, in view of proposition 3.2, are given by

$$\begin{aligned} d_{j,i} = \operatorname{sgn}(M-1-i)c_0^2 \sum_{t=0}^{M-i-2} M_0^{(K-j)M} H_{jM+i+t+1} M_0^t(\zeta_{jM+i}-\zeta) \\ + c_0 M_0^{(K-j)M}\left(H_{jM+i}\left(M_0^i \hat{e}_{jM}^{\delta} + (I-M_0^i)B^{-1}\zeta\right)\right. \\ \left. + M_0^{M-i-1}(\zeta_{jM+i}-\zeta)\right) := \sum_{t=0}^{M-i-2} d_{j,i,t} + d_{j,i,-1}, \end{aligned}$$

where the notation $\operatorname{sgn}(\cdot)$ denotes the sign function with the convention $\operatorname{sgn}(0)=0$. Next we repeat the argument for deriving (4.2), and claim that $\mathbb{E}_{jM+i}[R_1 d_{j,i}] = 0$ and $d_{j,i}|\mathcal{F}_{jM+i} \cup \mathcal{F}_{jM+i+1}^c$ is independent of $d_{j',i'}|\mathcal{F}_{jM+i} \cup \mathcal{F}_{jM+i+1}^c$ for any $j \neq j'$ or $i \neq i'$ where $0 \leqslant j' \leqslant j \leqslant K$, $0 \leqslant i, i' \leqslant M-1$. Indeed, the random variable $d_{j',i'}$ is measurable with respect to $\mathcal{F}_{jM+i} \cup \mathcal{F}_{jM+i+1}^c$. Then the direct computation using the identities $\mathbb{E}_{jM+i}[\zeta_{jM+i}-\zeta] = 0$ and $\mathbb{E}_{jM+i}[H_{jM+i}] = 0$ implies that for any $0 \leqslant j \leqslant K$ and $0 \leqslant i \leqslant M-1$, the following identity holds

$$\begin{aligned} \mathbb{E}_{jM+i}[R_1 d_{j,i}] = \operatorname{sgn}(M-1-i)c_0^2 \sum_{t=0}^{M-i-2} R_1 M_0^{(K-j)M} \\ \times H_{jM+i+t+1} M_0^t \mathbb{E}_{jM+i}[\zeta_{jM+i}-\zeta] + c_0 R_1 M_0^{(K-j)M} \\ \times \left(\mathbb{E}_{jM+i}[H_{jM+i}]\left(M_0^i \hat{e}_{jM}^{\delta} + (I-M_0^i)B^{-1}\zeta\right)\right. \\ \left. + M_0^{M-i-1}\mathbb{E}_{jM+i}[\zeta_{jM+i}-\zeta]\right) = 0. \end{aligned}$$

Thus we derive

$$\mathbb{E}[\|R_1(\hat{e}_{(K+1)M}^{\delta} - \mathbb{E}[\hat{e}_{(K+1)M}^{\delta}])\|^2] = \sum_{j=0}^{K}\sum_{i=0}^{M-1} \mathbb{E}[\|R_1 d_{j,i}\|^2].$$

Similarly, for fixed j, i and any $0 \leqslant t, t' \leqslant M-i-2$, $\mathbb{E}[d_{j,i,t}|\mathcal{F}_{jM+i+t+1}] = 0$ and $d_{j,i,t}|\mathcal{F}_{jM+i+t+1}$ is independent of $d_{j,i,t'}|\mathcal{F}_{jM+i+t+1}$ when $t > t'$. Consequently,

$$\mathbb{E}[\|R_1 d_{j,i}\|^2] = \sum_{t=-1}^{M-i-2} \mathbb{E}[\|R_1 d_{j,i,t}\|^2].$$

Thus, we obtain

$$\begin{aligned} &\mathbb{E}[\|R_1(\hat{e}_{(K+1)M}^{\delta} - \mathbb{E}[\hat{e}_{(K+1)M}^{\delta}])\|^2] \\ &= c_0^2 \sum_{j=0}^{K}\sum_{i=0}^{M-1} \mathbb{E}\left[\|R_1 M_0^{(K-j)M}\left(H_{jM+i}\left(M_0^i \hat{e}_{jM}^{\delta} + (I-M_0^i)B^{-1}\zeta\right)\right.\right. \\ &\quad \left.\left. + M_0^{M-i-1}(\zeta_{jM+i}-\zeta)\right)\|^2\right] \\ &\quad + c_0^4 \sum_{j=0}^{K}\sum_{i=0}^{M-2}\sum_{t=0}^{M-i-2} \mathbb{E}[\|R_1 M_0^{(K-j)M} H_{jM+i+t+1} M_0^t(\zeta_{jM+i}-\zeta)\|^2]. \end{aligned}$$

Reorganizing the last summation gives

$$\sum_{i=0}^{M-2}\sum_{t=0}^{M-i-2}\mathbb{E}[\|R_1M_0^{(K-j)M}H_{jM+i+t+1}M_0^t(\zeta_{jM+i}-\zeta)\|^2]$$
$$=\sum_{i=1}^{M-1}\sum_{t=0}^{i-1}\mathbb{E}[\|R_1M_0^{(K-j)M}H_{jM+i}M_0^t(\zeta_{jM+i-1-t}-\zeta)\|^2].$$

This completes the proof of the lemma.

ORCID iDs

Bangti Jin https://orcid.org/0000-0002-3775-9155
Jun Zou https://orcid.org/0000-0002-4809-7724

References

[1] Allen-Zhu Z and Hazan E 2016 Variance reduction for faster non-convex optimization *Proc. 33rd Int. Conf. Machine Learning, PMLR* vol 48 pp 699–707
[2] Allen-Zhu Z and Yuan Y 2016 Improved SVRG for non-strongly-convex or sum-of-non-convex objectives *Proc. 33rd Int. Conf. Machine Learning, PMLR* vol 48 pp 1080–9
[3] Bottou L, Curtis F E and Nocedal J 2018 Optimization methods for large-scale machine learning *SIAM Rev.* **60** 223–311
[4] Chen K, Li Q and Liu J-G 2018 Online learning in optical tomography: a stochastic approach *Inverse Problems* **34** 075010
[5] Defazio A, Bach F and Lacoste-Julien S 2014 SAGA: a fast incremental gradient method with support for non-strongly convex composite objectives *Adv. Neural Inf. Process. Syst.* vol 27 pp 1646–54
[6] Dieuleveut A and Bach F 2016 Nonparametric stochastic approximation with large step-sizes *Ann. Stat.* **44** 1363–99
[7] Engl H W, Hanke M and Neubauer A 1996 *Regularization of Inverse Problems* (Dordrecht: Kluwer)
[8] Gamba I M, Li Q and Nair A 2020 Reconstructing the thermal phonon transmission coefficient at solid interfaces in the phonon transport equation (arXiv:2011.13047)
[9] Gower R M, Schmidt M, Bach F and Richtarik P 2020 Variance-reduced methods for machine learning *Proc. IEEE* **108** 1968–83
[10] Hansen P C 2007 Regularization tools version 4.0 for Matlab 7.3 *Numer. Algorithms* **46** 189–94
[11] Harikandeh R B, Ahmed M O, Virani A, Schmidt M, Konevňý J and Sallinen S 2015 Stop wasting my gradients: practical SVRG *Advances in Neural Information Processing Systems (NIPS 2015)* vol 28 pp 2251–9
[12] Herman G T, Lent A and Lutz P H 1978 Relaxation methods for image reconstruction *Commun. ACM* **21** 152–8
[13] Hudson H M and Larkin R S 1994 Accelerated image reconstruction using ordered subsets of projection data *IEEE Trans. Med. Imaging* **13** 601–9
[14] Ito K and Jin B 2015 *Inverse Problems: Tikhonov Theory and Algorithms* (Singapore: World Scientific)
[15] Jahn T and Jin B 2020 On the discrepancy principle for stochastic gradient descent *Inverse Problems* **36** 095009
[16] Jin B and Lu X 2019 On the regularizing property of stochastic gradient descent *Inverse Problems* **35** 015004
[17] Jin B, Zhou Z and Zou J 2020 On the convergence of stochastic gradient descent for nonlinear ill-posed problems *SIAM J. Optim.* **30** 1421–50

[18] Jin B, Zhou Z and Zou J 2021 On the saturation phenomenon of stochastic gradient descent for linear inverse problems *SIAM/ASA J. Uncertain. Quantification* **9** 1553–88
[19] Johnson R and Zhang T 2013 Accelerating stochastic gradient descent using predictive variance reduction *NIPS'13* ed C J C Burges, L Bottou, M Welling, Z Ghahramani and K Q Weinberger (Lake Tahoe, Nevada) pp 315–23
[20] Kaltenbacher B, Neubauer A and Scherzer O 2008 *Iterative Regularization Methods for Nonlinear Ill-Posed Problems* (Berlin: de Gruyter & Co)
[21] Kereta Ž, Twyman R, Arridge S, Thielemans K and Jin B 2021 Stochastic EM methods with variance reduction for penalised PET reconstructions *Inverse Problems* **37** 115006
[22] Kindermann S 2021 Optimal-order convergence of Nesterov acceleration for linear ill-posed problems *Inverse Problems* **37** 065002
[23] Kovalev D, Horváth S and Richtárik P 2020 Don't jump through hoops and remove those loops: SVRG and Katyusha are better without the outer loop *Proc. 31st Int. Conf. Algorithmic Learning Theory, PMLR* vol 117 pp 451–67
[24] Le Roux N, Schmidt M and Bach F 2012 A stochastic gradient method with an exponential convergence rate for strongly-convex optimization with finite training sets *Adv. Neural Inf. Process. Syst.* vol 25 pp 2663–71
[25] Lin J and Rosasco L 2017 Optimal rates for multi-pass stochastic gradient methods *J. Mach. Learn. Res.* **18** 1–47
[26] Neubauer A 2017 On Nesterov acceleration for Landweber iteration of linear ill-posed problems *J. Inverse Ill-Posed Problems* **25** 381–90
[27] Nguyen L M, Liu J, Scheinberg K and Takáč M 2017 SARAH: a novel method for machine learning problems using stochastic recursive gradient *Proc. 34th Int. Conf.Machine Learning, PMLR* vol 70 pp 2613–21
[28] Pillaud-Vivien L, Rudi A and Bach F 2018 Statistical optimality of stochastic gradient descent on hard learning problems through multiple passes *Adv. Neural Inf. Process. Syst.* pp 8125–35
[29] Reddi S J, Hefny A, Sra S, Poczos B and Smola A 2016 Stochastic variance reduction for nonconvex optimization *Proc. 33rd Int. Conf. Machine Learning, PMLR* vol 48 pp 314–23
[30] Robbins H and Monro S 1951 A stochastic approximation method *Ann. Math. Stat.* **22** 400–7
[31] Shang F, Zhou K, Liu H, Cheng J, Tsang I W, Zhang L, Tao D and Jiao L 2020 VR-SGD: a simple stochastic variance reduction method for machine learning *IEEE Trans. Knowl. Data Eng.* **32** 188–202
[32] Strohmer T and Vershynin R 2009 A randomized Kaczmarz algorithm with exponential convergence *J. Fourier Anal. Appl.* **15** 262–78
[33] Tarrès P and Yao Y 2014 Online learning as stochastic approximation of regularization paths: optimality and almost-sure convergence *IEEE Trans. Inf. Theory* **60** 5716–35
[34] Xu Y, Lin Q and Yang T 2017 Adaptive SVRG methods under error bound conditions with unknown growth parameter *Advances in Neural Information Processing Systems* vol 31 pp 3279–89
[35] Ying Y and Pontil M 2008 Online gradient descent learning algorithms *Found. Comput. Math.* **8** 561–96
[36] Zhang L, Mahdavi M and Jin R 2013 Linear convergence with condition number independent access of full gradients *Advances in Neural Information Processing Systems* vol 26 pp 980–8