SIAM/ASA J. UNCERTAINTY QUANTIFICATION
Vol. 9, No. 4, pp. 1553–1588



# On the Saturation Phenomenon of Stochastic Gradient Descent for Linear Inverse Problems*

Bangti Jin†, Zehui Zhou‡, and Jun Zou‡

**Abstract.** Stochastic gradient descent (SGD) is a promising method for solving large-scale inverse problems due to its excellent scalability with respect to data size. The current mathematical theory in the lens of regularization theory predicts that SGD with a polynomially decaying stepsize schedule may suffer from an undesirable saturation phenomenon; i.e., the convergence rate does not further improve with the solution regularity index when it is beyond a certain range. In this work, we present a refined convergence rate analysis of SGD and prove that saturation actually does not occur if the initial stepsize of the schedule is sufficiently small. Several numerical experiments are provided to complement the analysis.

**Key words.** stochastic gradient descent, regularizing property, convergence rate, saturation, inverse problems

**AMS subject classifications.** 65J20, 65J22, 90C90

**DOI.** 10.1137/20M1374456

## 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 the context of inverse problems, the matrix $A$ is commonly ill-conditioned. When the matrix $A$ is rank-deficient, (1.1) may have infinitely many solutions. The reference solution $x^\dagger$ is taken to be the minimum norm solution relative to the initial guess $x_1$, i.e.,

$$x^\dagger = \arg\min_{x\in\mathbb{R}^m}\{\|x - x_1\| \quad \text{s.t.} \quad Ax = y^\dagger\},$$

with $\|\cdot\|$ being the Euclidean norm of a vector (and also the spectral norm of a matrix). In practice, we only have access to a noisy version $y^\delta$ of the exact data $y^\dagger = Ax^\dagger$, i.e.,

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

*Received by the editors October 21, 2020; accepted for publication (in revised form) August 6, 2021; published electronically November 2, 2021.
https://doi.org/10.1137/20M1374456
**Funding:** The work of the first author was supported by the UK EPSRC grant EP/T000864/1. The work of the third author was substantially supported by the Hong Kong RGC General Research Fund (projects 14306718 and 14304517).

†Department of Computer Science, University College London, Gower Street, London WC1E 6BT, UK (b.jin@ucl.ac.uk, bangti.jin@gmail.com).
‡Department of Mathematics, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong (zhzhou@math.cuhk.edu.hk, zou@math.cuhk.edu.hk).

where $\xi \in \mathbb{R}^n$ denotes the noise in the data with a noise level $\delta := \|\xi\|$. We denote the $i$th row of the matrix $A$ by a column vector $a_i \in \mathbb{R}^m$, i.e., $A = [a_i^t]_{i=1}^n$ (with the superscript $t$ denoting the matrix/vector transpose), and the $i$th entry of the vector $y^\delta \in \mathbb{R}^n$ by $y_i^\delta$. Linear inverse problems of the form (1.1) arise in a broad range of applications, e.g., initial condition/source identification and optical imaging. A large number of numerical methods have been developed, prominently variational regularization [7, 14] and iterative regularization [21].

Stochastic gradient descent (SGD) is one very promising numerical method for solving problem (1.1). In its simplest form, it reads as follows: Given an initial guess $x_1^\delta \equiv x_1 \in \mathbb{R}^m$, we update the iterate $x_{k+1}^\delta$ recursively by

$$x_{k+1}^\delta = x_k^\delta - \eta_k((a_{i_k}, x_k^\delta) - y_{i_k}^\delta)a_{i_k}, \quad k = 1, 2, \ldots, \tag{1.2}$$

where the random row index $i_k$ is drawn independent and identically distributed (i.i.d.) uniformly from the index set $\{1, \ldots, n\}$, $\eta_k > 0$ is the stepsize at the $k$th iteration, and $(\cdot, \cdot)$ denotes the Euclidean inner product. We denote by $\mathcal{F}_k$ the filtration generated by the random indices $\{i_1, \ldots, i_{k-1}\}$, define $\mathcal{F}$ by $\mathcal{F} = \bigvee_{k\in\mathbb{N}} \mathcal{F}_k$, and let $(\Omega, \mathcal{F}, \mathbb{P})$ be the associated probability space. The notation $\mathbb{E}[\cdot]$ denotes taking expectation with respect to the filtration $\mathcal{F}$. The SGD iterate $x_k^\delta$ is random, and it is measurable with respect to $\mathcal{F}_k$. SGD is a randomized version of the classical Landweber method [23]

$$x_{k+1}^\delta = x_k^\delta - \eta_k n^{-1} A^t(Ax_k^\delta - y^\delta), \quad k = 1, 2, \ldots, \tag{1.3}$$

which is identical to the gradient descent applied to the following objective functional:

$$J(x) = (2n)^{-1}\|Ax - y^\delta\|^2. \tag{1.4}$$

When compared with the Landweber method in (1.3), SGD (1.2) employs only one data pair $(a_{i_k}, y_{i_k}^\delta)$ instead of all data pairs, and thus it enjoys excellent scalability with respect to the data size. It is worth noting that due to the ill-conditioning of $A$ and the presence of noise in the data $y^\delta$, the exact minimizer of $J(x)$ is not of interest.

Since its first proposal by Robbins and Monro [29] for statistical inference, SGD has received a lot of attention in many diverse research areas (see the monograph [22] for various asymptotic results). Due to its excellent scalability, the interest in SGD and its variants has grown explosively in recent years in machine learning, and its accelerated variants, e.g., ADAM, have been established as the workhorse in many challenging deep learning training tasks [2, 3]. It has also achieved great success in inverse problems, e.g., in computed tomography (known as algebraic reconstruction techniques or the randomized Kaczmarz method [13, 27, 31, 17]) and optical tomography [4].

The theoretical analysis of SGD for solving inverse problems is still in its infancy. Let $e_k^\delta = x_k^\delta - x^\dagger$ be its error with respect to the minimum-norm solution $x^\dagger$. Only very recently, the regularizing property was proved in [18]: when equipped with a priori stopping rules, the mean squared error $\mathbb{E}[\|e_k^\delta\|^2]^{\frac{1}{2}}$ of the SGD iterate $x_k^\delta$ converges to zero as $\delta$ tends to zero, and furthermore, under the canonical power type source condition (see (1.5) in Assumption 1.1 below), it converges to zero at a certain rate. However, the result predicts that SGD can suffer from an undesirable saturation phenomenon for smooth solutions (i.e., with $\nu > \frac{1}{2}$):

$\mathbb{E}[\|e_k^\delta\|^2]^{\frac{1}{2}}$ converges at most at a rate $O(\delta^{\frac{1}{2}})$, which is slower than that achieved by the Landweber method [7, Chapter 6]; see also [15] for a posteriori stopping using the discrepancy principle and numerical illustration on the saturation phenomenon for SGD. Thus, SGD is suboptimal for "smooth" inverse solutions with $\nu > \frac{1}{2}$. This phenomenon is attributed to the inherent computational variance of the SGD approximation $x_k^\delta$, which arises from the use of a random gradient estimate in place of the true gradient. To the best of our knowledge, it remains unclear whether the saturation phenomenon is intrinsic to SGD.

In this work, we revisit the convergence rate analysis of SGD with a polynomially decaying stepsize schedule for small initial stepsize $c_0$, and aim at addressing the saturation phenomenon, under the standard source condition. First we state the standing assumptions for the analysis of SGD. The choice in (i) is commonly known as a polynomially decaying stepsize schedule. Part (ii) is the classical source condition, which represents a type of smoothness of the initial error $x^\dagger - x_1$ (with respect to the matrix $B$), and the condition on $B$ is easily achieved by rescaling the problem. Source type conditions are needed in order to derive the convergence rate, without which the convergence can be arbitrarily slow [7]. Loosely speaking, it restricts $x^\dagger - x_1$ to a suitable subspace which enables quantitatively bounding the approximation error. Note that the condition generally is insufficient to ensure a contractive map for the Landweber method. Below we shall focus on the case $\nu > \frac{1}{2}$, for which the current analysis [18] exhibits the saturation phenomenon, as mentioned above. Part (iii) assumes that the forward map $A$ takes a special form. Alternatively, it can be viewed as SGD applied to a preconditioned version of problem (1.1). To validate this condition, we present some numerical results for typical inverse problems in subsection 4.2, which indicates that this structure is irrelevant to the performance of SGD in the sense that it performs nearly identically on the problems with or without this structure. Thus this restriction is due to the limitation of the proof technique; see Remark 2.7 for the obstruction in the proof in the general case.

*Assumption* 1.1. Let $B = n^{-1}A^tA$ with $\|B\| \le 1$. The following assumptions hold.

(i) The stepsize $\eta_j = c_0 j^{-\alpha}$, $j = 1, \ldots, \alpha \in [0, 1)$, with

$$c_0 \le \min\left((\max_i \|a_i\|^2)^{-1}, 1\right) \quad \text{and} \quad c_0\|B\| \le (2e)^{-1}.$$

(ii) There exist $w \in \mathbb{R}^m$ and $\nu > \frac{1}{2}$ such that the exact solution $x^\dagger$ satisfies

$$x^\dagger - x_1 = B^\nu w. \tag{1.5}$$

(iii) The matrix $A = \Sigma V^t$ with $\Sigma$ being diagonal and nonnegative and $V$ column orthonormal.

Now we can state the main result of this work. By choosing the stopping index $k(\delta)$ in accordance with the (unknown) regularity index $\nu$ as $k = O(\|w\|\delta^{-1})^{\frac{2}{(1+2\nu)(1-\alpha)}}$, the result implies a convergence rate

$$\mathbb{E}[\|e_{k(\delta)}^\delta\|^2]^{\frac{1}{2}} \le c\|w\|^{\frac{1}{1+2\nu}}\delta^{\frac{2\nu}{1+2\nu}},$$

which is identical to that for the Landweber method [7, Chapter 6]. Thus, under the given condition, the aforementioned saturation phenomenon does not occur for SGD. This result partly settles the saturation phenomenon, and it complements existing analysis [18, 19].

Theorem 1.2. *Let Assumption 1.1 hold and $c_0$ be sufficiently small. Then there exist constants $c_*$ and $c_{**}$, which depend on $\nu$, $n$, $c_0$, and $\alpha$, such that*

$$\mathbb{E}[\|e_k^\delta\|^2] \le c_* k^{-2\nu(1-\alpha)}\|w\|^2 + c_{**}\delta^2 k^{1-\alpha}.$$

The condition "$c_0$ being sufficiently small" can be made more precise as $c_0 = O(n^{-1})$. Note that the stepsize choice $O(n^{-1})$ has been extensively used in the convergence analysis of SGD with random shuffling [33, 11, 30]. In Theorem 1.2, the constant condition on $c_0$ is not given explicitly. When $\alpha = 0$, the following condition is sufficient:

$$2(1+\phi(2\epsilon))nc_0^{2-2\epsilon} \le 1 \quad \text{for some } \epsilon \in (\tfrac{1}{2}, 1), \tag{1.6}$$

and the function $\phi$ is defined in Lemma 2.2 below; see Theorem 3.11. The numerical experiments in section 4 indicate that with a small initial stepsize $c_0$, SGD can indeed deliver reconstructions with accuracy comparable to that of the Landweber method for a range of regularity index and noise levels, and in the absence of the smallness condition on $c_0$, the results obtained by SGD are indeed suboptimal. These numerical results indicate the necessity and sufficiency of a small stepsize for achieving the optimal convergence rate.

The general strategy of proof is to decompose the error $e_k^\delta := x_k^\delta - x^\dagger$ into three components (with $x_k$ being the SGD iterate for exact data $y^\dagger$)

$$x_k^\delta - x^\dagger = (\mathbb{E}[x_k] - x^\dagger) + (\mathbb{E}[x_k^\delta] - \mathbb{E}[x_k]) + (x_k^\delta - \mathbb{E}[x_k^\delta]),$$

which represent, respectively, approximation error due to early stopping, propagation error due to data noise, and stochastic error due to randomness of gradient estimate, and then to bound the terms by bias-variance decomposition and the triangle inequality as

$$\mathbb{E}[\|x_k^\delta - x^\dagger\|^2] \le 2\|\mathbb{E}[x_k] - x^\dagger\|^2 + 2\|\mathbb{E}[x_k^\delta] - \mathbb{E}[x_k]\|^2 + \mathbb{E}[\|x_k^\delta - \mathbb{E}[x_k^\delta]\|^2].$$

In our analysis, we refine this decomposition by repeatedly expanding the random iterate noise within the third term and applying the bias-variance decomposition up to the $\ell$th fold; see Theorem 2.5 for the details. In the decomposition, Assumption 1.1(iii) is used in an essential manner to arrive at a simple recursion. It improves the existing analysis [18, 19] for SGD in the sense that the stochastic component is further decomposed. Then the analysis proceeds by bounding the first two components separately, and the third component by recursion, all of which in turn involves lengthy computation of certain summations. It is noteworthy that for the case of a constant stepsize, the convergence analysis can be greatly simplified; see section 3.4 for the details.

Finally, we situate the current work within a large body of literature on SGD. The convergence issue of SGD has been extensively studied in different senses, and two main lines of research that are related to this work are optimization and statistical learning, in addition to the aforementioned results for inverse problems. In the context of optimization, when the objective function is strictly convex, many results on the convergence of the iterates to the global minimizer are available; see, e.g., [16] for matching lower and upper bounds, and the references therein for further results. Note that $J(x)$ in (1.4) is not strictly convex. In general, the convergence of SGD is often measured by the optimality gap (i.e., the expected objective

function value to the optimal one) or the magnitude of the gradient. See the survey [3] for a recent overview on this line of research, including advances on nonconvex problems. Very recently the work [8] proved the local convergence of SGD with rates to minima of the objective function while avoiding convexity or contractivity assumptions. It is noteworthy that these results cannot be directly compared with the convergence rates given in Theorem 1.2, since the global minimizer to the objective function $J(x)$ is not of practical interest due to the ill-conditioning of $A$. This represents one essential difference between the results from optimization and those from regularization theory. The second line of research is the generalization error in reproducing kernel Hilbert spaces in statistical learning theory [34, 32, 5, 26, 28, 25]. These works aim at establishing upper bounds on the generalization error for SGD or its variants (often combined with a suitable averaging scheme), which differs from the error bound on the iterate itself. Nonetheless, the high level idea of analysis is similar: both use the bias-variance decomposition to bound relevant quantities, which often depend on source type conditions given in Assumption 1.1(ii). One major technical novelty of this work is to develop a recursive version of the bias-variance decomposition for the mean squared error.

The rest of the paper is organized as follows. In section 2, we derive a novel error decomposition, and then in section 3, we give the convergence rate analysis by bounding the three error components of the SGD iterate $x_k^\delta$. Finally, in section 4, we provide some illustrative numerical experiments to complement the theoretical analysis. Throughout, the notation $c$, with or without a subscript, denotes a generic constant, which may differ at each occurrence, but it is always independent of the iteration number $k$ (and the random index $i_k$) and the noise level $\delta$.

## 2. Error decomposition.

In this section, we present several preliminary estimates and a refined error decomposition.

### 2.1. Notation and preliminary estimates.

We will employ the following index sets extensively. For any $k_1 \le k_2$ and $1 \le i \le k_2 - k_1 + 1$, let

$$\begin{aligned}\mathcal{J}_{[k_1,k_2],i} &= \{\{j_\ell\}_{\ell=1}^i : k_1 \le j_i < j_{i-1} < \cdots < j_2 < j_1 \le k_2\},\\ J_i &= \{j_1, j_2, \ldots, j_i\}.\end{aligned}$$

Note that the set $\mathcal{J}_{[k_1,k_2],i}$ consists of (strictly monotone) multi-indices of length $i$, which arises naturally in the proof of Theorem 2.5 below. For $i = 0$, we adopt the convention $\mathcal{J}_{[k_1,k_2],0} = \{\emptyset\}$ and $J_0 = \emptyset$. For all $J_i = \{j_1, \ldots, j_i\} \in \mathcal{J}_{[k_1,k_2],i}$, with $0 \le i \le k_2 - k_1 + 1$, we define

$$J^c_{[k_1,k_2],i} = \{k_1, \ldots, k_2\} \setminus J_i,$$

where we omit the dependency on $J_i$ for notational simplicity. In particular,

$$\mathcal{J}_{[k_1,k_2],1} = \{\{k_1\}, \ldots, \{k_2\}\} \quad \text{and} \quad J^c_{[k_1,k_2],0} = \{k_1, \ldots, k_2\}.$$

For $i > k_2 - k_1 + 1$, we adopt the convention $\mathcal{J}_{[k_1,k_2],i} = \{\emptyset\}$, $J_i = \emptyset$, $J^c_{[k_1,k_2],i} = \emptyset$.

The next lemma collects useful identities on the summation over the indices $\mathcal{J}_{[1,k],i+1}$.

Lemma 2.1. *The following identities hold:*

$$\sum_{J_{i+1}\in\mathcal{J}_{[1,k],i+1}} = \sum_{J_i\in\mathcal{J}_{[2,k],i}} \sum_{j_{i+1}=1}^{j_i-1} = \sum_{j_{i+1}=1}^{k-i} \sum_{J_i\in\mathcal{J}_{[j_{i+1}+1,k],i}}. \tag{2.1}$$

*Proof.* The identities are direct from the definition:

$$\sum_{J_{i+1}\in\mathcal{J}_{[1,k],i+1}} = \sum_{j_1=i+1}^{k} \cdots \sum_{j_i=2}^{j_{i-1}-1} \sum_{j_{i+1}=1}^{j_i-1} = \sum_{J_i\in\mathcal{J}_{[2,k],i}} \sum_{j_{i+1}=1}^{j_i-1},$$

$$\sum_{J_{i+1}\in\mathcal{J}_{[1,k],i+1}} = \sum_{j_{i+1}=1}^{k-i} \sum_{j_i=j_{i+1}+1}^{k-i+1} \cdots \sum_{j_1=j_2+1}^{k} = \sum_{j_{i+1}=1}^{k-i} \sum_{J_i\in\mathcal{J}_{[j_{i+1}+1,k],i}}.$$

This directly shows the assertion. ∎

We use the following elementary inequality extensively.

Lemma 2.2. *For any $k\in\mathbb{N}$ and $s\in\mathbb{R}$, there holds*

$$\sum_{j=1}^{k} j^{-s} \le \begin{cases} 2^{1-s}(1-s)^{-1}k^{1-s}, & s<0,\\ (1-s)^{-1}k^{1-s}, & s\in[0,1),\\ 2\max(\ln k,1), & s=1,\\ s(s-1)^{-1}, & s>1. \end{cases} \tag{2.2}$$

Throughout, we denote the constant and power on the right-hand side of the inequality (2.2) by $\phi(s)$ and $k^{\max(1-s,0)}$, respectively, with the shorthand $k^{\max(0,0)}=\max(\ln k,1)$.

The next result bounds the spectral norm of the matrix product $\Pi_J(B)B^s$, which, for each index set $J$, is defined by (with the convention $\Pi_\emptyset(B)=I$)

$$\Pi_J(B)=\prod_{j\in J}(I-\eta_j B).$$

Lemma 2.3. *Under Assumption 1.1(i), for any $s>0$ and $J_\ell\in\mathcal{J}_{[k',k],\ell}$ with $k'\le k$, $0\le\ell<k+1-k'$,*

$$\|\Pi_{J^c_{[k',k],\ell}}(B)B^s\| \le s^s(ec_0)^{-s}(k+1-k'-\ell)^{-s}k^{\alpha s}.$$

*Proof.* For any $s>0$ and $J_\ell\in\mathcal{J}_{[k',k],\ell}$ with $k'\le k$, $0\le\ell<k+1-k'$, there holds

$$\|\Pi_{J^c_{[k',k],\ell}}(B)B^s\| \le \sup_{\lambda\in\mathrm{Sp}(B)}|\lambda^s\Pi_{J^c_{[k',k],\ell}}(\lambda)| = \sup_{\lambda\in\mathrm{Sp}(B)}\lambda^s\prod_{i\in J^c_{[k',k],\ell}}(1-\eta_i\lambda),$$

where $\mathrm{Sp}(B)$ denotes the spectrum of $B$. For any $x\in\mathbb{R}$, there holds the inequality $1-x\le e^{-x}$, and thus

$$\lambda^s\prod_{i\in J^c_{[k',k],\ell}}(1-\eta_i\lambda) \le \lambda^s\prod_{i\in J^c_{[k',k],\ell}}e^{-\eta_i\lambda} = \lambda^s e^{-\lambda\sum_{i\in J^c_{[k',k],\ell}}\eta_i}.$$

For the function $g(\lambda) = \lambda^s e^{-\lambda a}$, with $a > 0$, the maximum is attained at $\lambda^* = sa^{-1}$, with a maximum value $s^s(ea)^{-s}$. Then setting $a = \sum_{i \in J^c_{[k',k],\ell}} \eta_i$ and applying the inequality $a \geq c_0(k+1-k'-\ell)k^{-\alpha}$ complete the proof of the lemma. ∎

The last lemma gives two useful bounds on the summations over the set $\mathcal{J}_{[1,k],i}$.

Lemma 2.4. *The following estimates hold.*

(i) *For any $k \geq 2$, $\alpha \in [0,1)$, and $2 \leq i \leq k$, there holds*

$$\sum_{J_i \in \mathcal{J}_{[1,k],i}} \prod_{t=1}^{i} j_t^{-2\alpha} \leq \phi(2\alpha)^i (k^{\max(1-2\alpha,0)})^i.$$

(ii) *For any $j = 0, \ldots, k-1$ and $i = 1, \ldots, k-j$, we have*

$$\sum_{J_i \in \mathcal{J}_{[j+1,k],i}} 1 \leq \frac{(k-j)^i}{i!}.$$

*Proof.* Assertion (i) follows from (2.2) as

$$\begin{aligned}\sum_{J_i \in \mathcal{J}_{[1,k],i}} \prod_{t=1}^{i} j_t^{-2\alpha} &= \sum_{j_1=i}^{k} \sum_{j_2=i-1}^{j_1-1} \cdots \sum_{j_i=1}^{j_{i-1}-1} j_i^{-2\alpha} \prod_{t=1}^{i-1} j_t^{-2\alpha} \\ &\leq \prod_{t=1}^{i} \Big( \sum_{j_t=1}^{k} j_t^{-2\alpha} \Big) \leq \big(\phi(2\alpha) k^{\max(1-2\alpha,0)}\big)^i.\end{aligned}$$

By the definition of the index set $\mathcal{J}_{[j+1,k],i}$, we have the identity

$$\begin{aligned}\sum_{J_i \in \mathcal{J}_{[j+1,k],i}} 1 &= \sum_{j_i=j+1}^{k-i+1} \cdots \sum_{j_2=j_3+1}^{k-1} \sum_{j_1=j_2+1}^{k} 1 \\ \leq \sum_{j_i=j+1}^{k-1} \cdots \sum_{j_2=j_3+1}^{k-1} \sum_{j_1=j_2+1}^{k} 1 &= \sum_{j_i=j+1}^{k-1} \cdots \sum_{j_2=j_3+1}^{k-1} (k-j_2).\end{aligned}$$

Then assertion (ii) follows by repeatedly applying the inequality

$$\sum_{t=1}^{T} t^s \leq (s+1)^{-1}(T+1)^{s+1} \quad \forall T \in \mathbb{N}, s \geq 0.$$

This completes the proof of the lemma. ∎

**2.2. Error decomposition.** Now we derive an important error decomposition. Below, we denote the SGD iterates for the exact data $y^\dagger$ and noisy data $y^\delta$ by $x_k$ and $x_k^\delta$, respectively, and also use the following shorthand notation:

$$\bar{A} = n^{-\frac{1}{2}} A, \quad \bar{\xi} = n^{-\frac{1}{2}} \xi, \quad \bar{\delta} = n^{-\frac{1}{2}} \delta, \quad \text{and} \quad e_k = x_k - x^\dagger.$$

The following result plays a central role in the convergence analysis.

Theorem 2.5. *Under Assumption 1.1*(iii)*, for any* $0 \le \ell < k$*, the following error decomposition holds:*

$$\mathbb{E}[\|e_{k+1}^\delta\|^2] \le \sum_{i=0}^{\ell} \mathrm{I}_{i,1}^\delta + \sum_{i=0}^{\ell} \mathrm{I}_{i,2}^\delta + (\mathrm{I}_\ell^\delta)^c, \tag{2.3}$$

*where the terms* $\mathrm{I}_{i,j}^\delta$, $i = 0, 1, \ldots, \ell$, $j = 1, 2$, *are defined by*

$$\begin{aligned}
\mathrm{I}_{0,1}^\delta &= 2\|\Pi_{J^c_{[1,k],0}}(B)e_1\|^2,\\
\mathrm{I}_{0,2}^\delta &= 2\bar\delta^2\Big(\Big\|\sum_{j=1}^{k}\eta_j\Pi_{J^c_{[j+1,k],0}}(B)B^{\frac12}\Big\|^2 + (n-1)\sum_{j=1}^{k}\eta_j^2\|\Pi_{J^c_{[j+1,k],0}}(B)B^{\frac12}\|^2\Big),\\
\mathrm{I}_{i,1}^\delta &= 2^{i+1}(n-1)^i\sum_{J_i\in\mathcal{J}_{[1,k],i}}\prod_{t=1}^{i}\eta_{j_t}^2\|\Pi_{J^c_{[1,k],i}}(B)B^ie_1\|^2 \quad \forall 1\le i\le \ell,\\
\mathrm{I}_{i,2}^\delta &= 2^{i+1}(n-1)^i\bar\delta^2\sum_{J_i\in\mathcal{J}_{[2,k],i}}\prod_{t=1}^{i}\eta_{j_t}^2\Big(\Big\|\sum_{j_{i+1}=1}^{j_i-1}\eta_{j_{i+1}}\Pi_{J^c_{[j_{i+1}+1,k],i}}(B)B^{i+\frac12}\Big\|^2\\
&\qquad + (n-1)\sum_{j_{i+1}=1}^{j_i-1}\eta_{j_{i+1}}^2\|\Pi_{J^c_{[j_{i+1}+1,k],i}}(B)B^{i+\frac12}\|^2\Big) \quad \forall 1\le i\le \ell,\\
(\mathrm{I}_\ell^\delta)^c &= 2^{\ell+1}(n-1)^{\ell+1}\sum_{J_{\ell+1}\in\mathcal{J}_{[1,k],\ell+1}}\prod_{t=1}^{\ell+1}\eta_{j_t}^2\mathbb{E}[\|\Pi_{J^c_{[j_{\ell+1}+1,k],\ell}}(B)B^{\ell+1}e_{j_{\ell+1}}^\delta\|^2].
\end{aligned}$$

*Proof.* Recall that $J_i = \{j_1, \ldots, j_i\}$ for any $i \ge 1$ and $J_0 = \emptyset$. By the definition of the SGD iteration (1.2), we have

$$e_k^\delta = (I - \eta_{k-1}B)e_{k-1}^\delta + \eta_{k-1}H_{k-1} = \Pi_{J^c_{[1,k-1],0}}(B)e_1^\delta + \sum_{j=1}^{k-1}\big(\eta_j\Pi_{J^c_{[j+1,k-1],0}}(B)H_j\big), \tag{2.4}$$

where $H_j$ is defined by

$$H_j = Be_j^\delta - \big((a_{i_j}, x_j^\delta) - y_{i_j}^\delta\big)a_{i_j} = \big(B - a_{i_j}a_{i_j}^t\big)e_j^\delta + \xi_{i_j}a_{i_j}. \tag{2.5}$$

By bias-variance decomposition and the triangle inequality, we have

$$\begin{aligned}
\mathbb{E}[\|e_{k+1}^\delta\|^2] &= \|\mathbb{E}[e_{k+1}^\delta]\|^2 + \mathbb{E}[\|\mathbb{E}[x_{k+1}^\delta] - x_{k+1}^\delta\|^2]\\
&\le 2\|\mathbb{E}[e_{k+1}]\|^2 + 2\|\mathbb{E}[x_{k+1} - x_{k+1}^\delta]\|^2 + \mathbb{E}[\|\mathbb{E}[x_{k+1}^\delta] - x_{k+1}^\delta\|^2].
\end{aligned} \tag{2.6}$$

It is known that the following estimates hold [18]:

$$\|\mathbb{E}[e_{k+1}]\| = \|\Pi_{J^c_{[1,k],0}}(B)e_1\|,$$

$$\|\mathbb{E}[x_{k+1} - x^\delta_{k+1}]\| \le \bar{\delta}\|\sum_{j=1}^{k}\eta_j\Pi_{J^c_{[j+1,k],0}}(B)B^{\frac{1}{2}}\|,$$

$$\mathbb{E}[\|x^\delta_{k+1} - \mathbb{E}[x^\delta_{k+1}]\|^2] = \sum_{j=1}^{k}\eta_j^2\mathbb{E}[\|\Pi_{J^c_{[j+1,k],0}}(B)A^tN^\delta_j\|^2], \tag{2.7}$$

with the iteration noise $N^\delta_j$ (at the $j$th SGD iteration) given by

$$N^\delta_j = n^{-1}(Ax^\delta_j - y^\delta) - ((a_{i_j}, x^\delta_j) - y^\delta_{i_j})b_{i_j},$$

where $b_i = (0,\dots,0,1,0,\dots,0)^t \in \mathbb{R}^n$ denotes the $i$th canonical Cartesian basis vector. Let

$$\tilde{N}^\delta_j = ((a_{i_j}, x^\delta_j) - y^\delta_{i_j})b_{i_j} = ((a_{i_j}, e^\delta_j) - \xi_{i_j})b_{i_j}.$$

Then the iteration noise $N^\delta_j$ can be rewritten as

$$N^\delta_j = \mathbb{E}[\tilde{N}^\delta_j|\mathcal{F}_j] - \tilde{N}^\delta_j.$$

Next we claim that under Assumption 1.1(iii), there holds

$$\|\Pi_{J^c_{[j+1,k],0}}(B)A^t(Ae^\delta_j - \xi)\|^2 = \sum_{i=1}^{n}\|\Pi_{J^c_{[j+1,k],0}}(B)A^t\big((a_i, e^\delta_j) - \xi_i\big)b_i\|^2. \tag{2.8}$$

Actually, in view of Assumption 1.1(iii), for any $1 \le j \le k$, the following hold:

$$Ae^\delta_j - \xi = \sum_{i=1}^{n}\big((a_i, e^\delta_j) - \xi_i\big)b_i \quad \text{and} \quad \Pi_{J^c_{[j+1,k],0}}(B)A^t = V\Pi_{J^c_{[j+1,k],0}}(n^{-1}\Sigma^t\Sigma)\Sigma^t.$$

Then the claim (2.8) follows from these two identities and column orthonormality of $V$ as

$$\|\Pi_{J^c_{[j+1,k],0}}(B)A^t(Ae^\delta_j - \xi)\|^2 = \Big\|V\sum_{i=1}^{n}\Pi_{J^c_{[j+1,k],0}}(n^{-1}\Sigma^t\Sigma)\Sigma^t\big((a_i, e^\delta_j) - \xi_i\big)b_i\Big\|^2$$

$$= \sum_{i=1}^{n}\|V\Pi_{J^c_{[j+1,k],0}}(n^{-1}\Sigma^t\Sigma)\Sigma^t\big((a_i, e^\delta_j) - \xi_i\big)b_i\|^2 = \sum_{i=1}^{n}\|\Pi_{J^c_{[j+1,k],0}}(B)A^t\big((a_i, e^\delta_j) - \xi_i\big)b_i\|^2.$$

Thus, by the bias-variance decomposition and the definitions of the notation $\bar{A}$ and $\bar{\xi}$, etc., we have, for any $j = 1,\dots,k$,

$$\mathbb{E}[\|\Pi_{J^c_{[j+1,k],0}}(B)A^tN^\delta_j\|^2|\mathcal{F}_j] = \mathbb{E}[\|\Pi_{J^c_{[j+1,k],0}}(B)A^t\tilde{N}^\delta_j\|^2|\mathcal{F}_j] - \|\Pi_{J^c_{[j+1,k],0}}(B)A^t\mathbb{E}[\tilde{N}^\delta_j|\mathcal{F}_j]\|^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}\|\Pi_{J^c_{[j+1,k],0}}(B)A^t\big((a_i, e^\delta_j) - \xi_i\big)b_i\|^2 - \|\Pi_{J^c_{[j+1,k],0}}(B)\frac{A^t}{n}(Ae^\delta_j - \xi)\|^2$$

$$= (n-1)\|\Pi_{J^c_{[j+1,k],0}}(B)(Be^\delta_j - \bar{A}^t\bar{\xi})\|^2.$$

By the Cauchy–Schwarz inequality, the identity $\|\Pi_{J^c_{[j+1,k],0}}(B)\bar{A}^t\|^2 = \|\Pi_{J^c_{[j+1,k],0}}(B)B^{\frac{1}{2}}\|^2$, and the triangle inequality, we deduce from (2.7) that

$$\begin{aligned}
\mathbb{E}[\|x^\delta_{k+1} - \mathbb{E}[x^\delta_{k+1}]\|^2] &= \sum_{j=1}^k \eta_j^2 \mathbb{E}\big[\mathbb{E}[\|\Pi_{J^c_{[j+1,k],0}}(B)A^t N^\delta_j\|^2|\mathcal{F}_j]\big] \\
=&(n-1)\sum_{j=1}^k \eta_j^2 \mathbb{E}[\|\Pi_{J^c_{[j+1,k],0}}(B)(Be^\delta_j - \bar{A}^t\bar{\xi})\|^2] \\
\le&2(n-1)\sum_{j=1}^k \eta_j^2 \Big(\mathbb{E}[\|\Pi_{J^c_{[j+1,k],0}}(B)Be^\delta_j\|^2] + \|\Pi_{J^c_{[j+1,k],0}}(B)\bar{A}^t\bar{\xi}\|^2\Big) \\
\le&2(n-1)\sum_{j=1}^k \eta_j^2 \mathbb{E}[\|\Pi_{J^c_{[j+1,k],0}}(B)Be^\delta_j\|^2] + 2(n-1)\bar{\delta}^2\sum_{j=1}^k \eta_j^2 \|\Pi_{J^c_{[j+1,k],0}}(B)B^{\frac{1}{2}}\|^2.
\end{aligned}$$

By the definitions of $\mathrm{I}^\delta_{0,1}$, $\mathrm{I}^\delta_{0,2}$, and $(\mathrm{I}^\delta_0)^c$, we have

$$\mathbb{E}[\|e^\delta_{k+1}\|^2] \le \mathrm{I}^\delta_{0,1} + \mathrm{I}^\delta_{0,2} + (\mathrm{I}^\delta_0)^c.$$

Next further expanding $\{e^\delta_j\}_{j=2}^k$ in the expression of $(\mathrm{I}^\delta_0)^c$ using (2.4) gives

$$\begin{aligned}
(\mathrm{I}^\delta_0)^c =& 2(n-1)\eta_1^2\|\Pi_{J^c_{[2,k],0}}(B)Be^\delta_1\|^2 \\
&+ 2(n-1)\sum_{j_1=2}^k \eta_{j_1}^2 \mathbb{E}\Big[\Big\|\Pi_{J^c_{[j_1+1,k],0}}(B)B\Big(\Pi_{J^c_{[1,j_1-1],0}}(B)e^\delta_1 + \sum_{j_2=1}^{j_1-1}\big(\Pi_{J^c_{[j_2+1,j_1-1],0}}(B)\eta_{j_2}H_{j_2}\big)\Big)\Big\|^2\Big].
\end{aligned}$$

Now using the definition of $H_j$ in (2.5), we obtain

$$\begin{aligned}
\sum_{j_2=1}^{j_1-1}\big(\Pi_{J^c_{[j_2+1,j_1-1],0}}(B)\eta_{j_2}H_{j_2}\big) =& \sum_{j_2=1}^{j_1-1}\big(\eta_{j_2}\Pi_{J^c_{[j_2+1,j_1-1],0}}(B)(B - a_{i_{j_2}}a^t_{i_{j_2}})e^\delta_{j_2}\big) \\
&+ \sum_{j_2=1}^{j_1-1}\big(\eta_{j_2}\Pi_{J^c_{[j_2+1,j_1-1],0}}(B)\xi_{i_{j_2}}a_{i_{j_2}}\big).
\end{aligned}$$

Thus, we can further bound $(\mathrm{I}^\delta_0)^c$ by

$$\begin{aligned}
(\mathrm{I}^\delta_0)^c \le&2(n-1)\eta_1^2\|\Pi_{J^c_{[2,k],0}}(B)Be^\delta_1\|^2 \\
&+ 4(n-1)\sum_{j_1=2}^k \eta_{j_1}^2 \mathbb{E}\Big[\Big\|\Pi_{J^c_{[j_1+1,k],0}}(B)B\sum_{j_2=1}^{j_1-1}\big(\eta_{j_2}\Pi_{J^c_{[j_2+1,j_1-1],0}}(B)\xi_{i_{j_2}}a_{i_{j_2}}\big)\Big\|^2\Big] \\
&+ 4(n-1)\sum_{j_1=2}^k \eta_{j_1}^2 \mathbb{E}\Big[\Big\|\Pi_{J^c_{[j_1+1,k],0}}(B)B\Big(\Pi_{J^c_{[1,j_1-1],0}}(B)e^\delta_1
\end{aligned}$$

$$+\sum_{j_2=1}^{j_1-1}\big(\eta_{j_2}\Pi_{J^c_{[j_2+1,j_1-1],0}}(B)(B-a_{i_{j_2}}a_{i_{j_2}}^t)e^\delta_{j_2}\big)\Big\|^2\Big]$$

$$\le 2(n-1)\eta_1^2\|\Pi_{J^c_{[2,k],0}}(B)Be_1^\delta\|^2+4(n-1)\sum_{j_1=2}^{k}\eta_{j_1}^2\underbrace{\mathbb{E}\Big[\Big\|\sum_{j_2=1}^{j_1-1}\big(\eta_{j_2}\Pi_{J^c_{[j_2+1,k],1}}(B)B\xi_{i_{j_2}}a_{i_{j_2}}\big)\Big\|^2\Big]}_{\mathrm{II}_1}$$

$$+4(n-1)\sum_{j_1=2}^{k}\eta_{j_1}^2\underbrace{\mathbb{E}\Big[\Big\|\Pi_{J^c_{[1,k],1}}(B)Be_1^\delta+\sum_{j_2=1}^{j_1-1}\big(\eta_{j_2}\Pi_{J^c_{[j_2+1,k],1}}(B)B(B-a_{i_{j_2}}a_{i_{j_2}}^t)e^\delta_{j_2}\big)\Big\|^2\Big]}_{\mathrm{II}_2}.$$

Next we simplify the two terms $\mathrm{II}_1$ and $\mathrm{II}_2$. Under Assumption 1.1(iii), direct computation gives, for any $j_1=2,\dots,k$ and $j,j'=1,\dots,j_1-1$,

$$\begin{aligned}&\mathbb{E}[\langle\Pi_{J^c_{[j'+1,k],1}}(B)B\xi_{i_{j'}}a_{i_{j'}},\Pi_{J^c_{[j+1,k],1}}(B)B\xi_{i_j}a_{i_j}\rangle]\\&=\begin{cases}n\langle\Pi_{J^c_{[j'+1,k],1}}(B)B\bar{A}^t\bar{\xi},\Pi_{J^c_{[j+1,k],1}}(B)B\bar{A}^t\bar{\xi}\rangle, & j'=j,\\ \langle\Pi_{J^c_{[j'+1,k],1}}(B)B\bar{A}^t\bar{\xi},\Pi_{J^c_{[j+1,k],1}}(B)B\bar{A}^t\bar{\xi}\rangle, & j'\neq j.\end{cases}\end{aligned}\tag{2.9}$$

Indeed, the case $j'\neq j$ follows directly. Meanwhile, under Assumption 1.1(iii), we have

$$\Pi_{J^c_{[j+1,k],1}}(B)BA^t\xi=V\Pi_{J^c_{[j+1,k],1}}(n^{-1}\Sigma^t\Sigma)(n^{-1}\Sigma^t\Sigma)\Sigma^t\sum_{i=1}^{n}\xi_ib_i,$$

and $V\Sigma^t\xi_ib_i=A^t\xi_ib_i=a_i\xi_i$. This and the column orthonormality of the matrix $V$ imply

$$\begin{aligned}&n\|\Pi_{J^c_{[j+1,k],1}}(B)B\bar{A}^t\bar{\xi}\|^2=n^{-1}\Big\|V\Pi_{J^c_{[j+1,k],1}}(n^{-1}\Sigma^t\Sigma)(n^{-1}\Sigma^t\Sigma)\Sigma^t\sum_{i=1}^{n}\xi_ib_i\Big\|^2\\&=n^{-1}\sum_{i=1}^{n}\|V\Pi_{J^c_{[j+1,k],1}}(n^{-1}\Sigma^t\Sigma)(n^{-1}\Sigma^t\Sigma)\Sigma^t\xi_ib_i\|^2=n^{-1}\sum_{i=1}^{n}\|\Pi_{J^c_{[j+1,k],1}}(B)Ba_i\xi_i\|^2\\&=\mathbb{E}[\|\Pi_{J^c_{[j+1,k],1}}(B)B\xi_{i_j}a_{i_j}\|^2].\end{aligned}$$

This and the bias-variance decomposition imply that the term $\mathrm{II}_1$ can be simplified to

$$\begin{aligned}\mathrm{II}_1&=\Big\|\sum_{j_2=1}^{j_1-1}\eta_{j_2}\Pi_{J^c_{[j_2+1,k],1}}(B)B\bar{A}^t\bar{\xi}\Big\|^2+(n-1)\sum_{j_2=1}^{j_1-1}\eta_{j_2}^2\|\Pi_{J^c_{[j_2+1,k],1}}(B)B\bar{A}^t\bar{\xi}\|^2\\&\le\bar{\delta}^2\Big\|\sum_{j_2=1}^{j_1-1}\eta_{j_2}\Pi_{J^c_{[j_2+1,k],1}}(B)B^{\frac{3}{2}}\Big\|^2+(n-1)\bar{\delta}^2\sum_{j_2=1}^{j_1-1}\eta_{j_2}^2\|\Pi_{J^c_{[j_2+1,k],1}}(B)B^{\frac{3}{2}}\|^2.\end{aligned}$$

Further, by the measurability of $e_j^\delta$ with respect to $\mathcal{F}_j$, we have

$$\mathbb{E}[\langle e_1^\delta,(B-a_{i_j}a_{i_j}^t)e_j^\delta\rangle]=\langle e_1^\delta,\mathbb{E}[\mathbb{E}[(B-a_{i_j}a_{i_j}^t)e_j^\delta|\mathcal{F}_j]]\rangle=0\quad\forall j,\tag{2.10}$$

since $e_1^\delta$ is deterministic, and similarly,

$$\begin{aligned}&\mathbb{E}[\langle (B-a_{i_{j'}}a_{i_{j'}}^t)e_{j'}^\delta, (B-a_{i_j}a_{i_j}^t)e_j^\delta\rangle]\\ =&\mathbb{E}[\langle (B-a_{i_{j'}}a_{i_{j'}}^t)e_{j'}^\delta, \mathbb{E}[(B-a_{i_j}a_{i_j}^t)e_j^\delta|\mathcal{F}_j]\rangle] = 0 \quad \forall j' < j.\end{aligned} \tag{2.11}$$

Consequently, there holds

$$\mathrm{II}_2 = \|\Pi_{J^c_{[1,k],1}}(B)Be_1^\delta\|^2 + \sum_{j_2=1}^{j_1-1}\eta_{j_2}^2\mathbb{E}[\|\Pi_{J^c_{[j_2+1,k],1}}(B)B(B-a_{i_{j_2}}a_{i_{j_2}}^t)e_{j_2}^\delta\|^2].$$

Combining these estimates with the definitions of the quantities $\mathrm{I}_{1,1}^\delta$, $\mathrm{I}_{1,2}^\delta$, and $(\mathrm{I}_1^\delta)^c$ gives

$$\begin{aligned}(\mathrm{I}_0^\delta)^c \le& 2(n-1)\eta_1^2\|\Pi_{J^c_{[2,k],0}}(B)Be_1^\delta\|^2\\ &+4(n-1)\bar\delta^2\sum_{j_1=2}^k\eta_{j_1}^2\Big(\Big\|\sum_{j_2=1}^{j_1-1}\eta_{j_2}\Pi_{J^c_{[j_2+1,k],1}}(B)B^{\frac32}\Big\|^2+(n-1)\sum_{j_2=1}^{j_1-1}\eta_{j_2}^2\|\Pi_{J^c_{[j_2+1,k],1}}(B)B^{\frac32}\|^2\Big)\\ &+4(n-1)\sum_{j_1=2}^k\eta_{j_1}^2\Big(\|\Pi_{J^c_{[1,k],1}}(B)Be_1^\delta\|^2+\sum_{j_2=1}^{j_1-1}\eta_{j_2}^2\mathbb{E}[\|\Pi_{J^c_{[j_2+1,k],1}}(B)B(B-a_{i_{j_2}}a_{i_{j_2}}^t)e_{j_2}^\delta\|^2]\Big)\\ \le&4(n-1)\sum_{j_1=1}^k\eta_{j_1}^2\|\Pi_{J^c_{[1,k],1}}(B)Be_1^\delta\|^2\\ &+4(n-1)\bar\delta^2\sum_{j_1=2}^k\eta_{j_1}^2\Big(\Big\|\sum_{j_2=1}^{j_1-1}\eta_{j_2}\Pi_{J^c_{[j_2+1,k],1}}(B)B^{\frac32}\Big\|^2+(n-1)\sum_{j_2=1}^{j_1-1}\eta_{j_2}^2\|\Pi_{J^c_{[j_2+1,k],1}}(B)B^{\frac32}\|^2\Big)\\ &+4(n-1)\sum_{j_1=2}^k\eta_{j_1}^2\Big(\sum_{j_2=1}^{j_1-1}\eta_{j_2}^2\mathbb{E}[\|\Pi_{J^c_{[j_2+1,k],1}}(B)B(B-a_{i_{j_2}}a_{i_{j_2}}^t)e_{j_2}^\delta\|^2]\Big)\\ =&\mathrm{I}_{1,1}^\delta+\mathrm{I}_{1,2}^\delta+(\mathrm{I}_1^\delta)^c.\end{aligned}$$

Similar to the analysis of $(\mathrm{I}_0^\delta)^c$, by repeating the argument, we obtain

$$(\mathrm{I}_1^\delta)^c = 4(n-1)^2\sum_{j_1=2}^k\sum_{j_2=1}^{j_1-1}\eta_{j_1}^2\eta_{j_2}^2\mathbb{E}[\|\Pi_{J^c_{[j_2+1,k],1}}(B)B^2e_{j_2}^\delta\|^2].$$

In general, we can derive

$$(\mathrm{I}_\ell^\delta)^c = 2^{\ell+1}(n-1)^\ell\sum_{J_{\ell+1}\in\mathcal{J}_{[1,k],\ell+1}}\prod_{t=1}^{\ell+1}\eta_{j_t}^2\mathbb{E}[\|\Pi_{J^c_{[j_{\ell+1}+1,k],\ell}}(B)B^\ell(B-a_{i_{j_{\ell+1}}}a_{i_{j_{\ell+1}}}^t)e_{j_{\ell+1}}^\delta\|^2].$$

Then repeating the preceding argument and noting the relation $e_1^\delta = e_1$ complete the proof. ■

*Remark* 2.6. In Theorem 2.5, Assumption 1.1(iii) plays a central role in the refined error decomposition at two places, i.e., (2.8) and (2.9). Intuitively, the condition essentially assumes low correlation between the rows of the matrix $A$, in analogy to the mutual coherence condition in compressed sensing [6]. The numerical experiments in section 4 indicate that SGD performs comparably with or without this assumption.

*Remark* 2.7. It is instructive to see the obstruction in extending the argument of Theorem 2.5 to a general matrix $A$ with exact data (i.e., $\xi = 0$) in the absence of Assumption 1.1(iii). Let the singular value decomposition of $A$ be $A = U\Sigma V^t$, with $\Sigma \in \mathbb{R}^{n\times m}$ being diagonal with positive diagonal entries $\{\sigma_i\}_{i=1}^r$ (with $r \le \min(m,n)$ being the rank, ordered nonincreasingly) and $U = [u_1, \ldots, u_n] \in \mathbb{R}^{n\times n}$ and $V = [v_1, \ldots, v_m] \in \mathbb{R}^{m\times m}$ being column orthonormal. Now consider the right-hand side and left-hand side, denoted by RHS and LHS, respectively, of the crucial identity (2.8) with a random index set $J$ and a random vector $e \in \mathbb{R}^m$ (by suppressing the subscripts). In view of the identity $a_i^t = b_i^t A$, we have

$$\text{LHS} = \|V\Pi_J(n^{-1}\Sigma^t\Sigma)\Sigma^t U^t Ae\|^2 = \|DU^t Ae\|^2 = \sum_{j=1}^n (d_j u_j^t(Ae))^2 = \sum_{j=1}^r d_j^2 (u_j^t(Ae))^2,$$

$$\text{RHS} = \sum_{i=1}^n \|V\Pi_J(n^{-1}\Sigma^t\Sigma)\Sigma^t U^t b_i b_i^t Ae\|^2 = \sum_{i=1}^n \|DU^t b_i (Ae)_i\|^2 = \sum_{j=1}^r d_j^2 \sum_{i=1}^n (u_{ji}(Ae)_i)^2,$$

with the diagonal matrix $D$ given by $D = \Pi_J(n^{-1}\Sigma^t\Sigma)\Sigma^t := \operatorname{diag}(d_1, \ldots, d_n)$, with the first $r$ entries being strictly positive. Since the index set $J$ is arbitrary, the existence of a constant $c$ (independent of $J$) such that RHS $\le c$LHS essentially requires

$$\sum_{i=1}^n (u_{ji}(Ae)_i)^2 \le c(u_j^t(Ae))^2, \quad j = 1, \ldots, r.$$

Since $Ae = \sum_{\ell=1}^r \sigma_\ell u_\ell v_\ell^t e$, the above inequality is equivalent to

$$\sum_{i=1}^n (u_{ji}(Ae)_i)^2 \le c(\sigma_j v_j^t e)^2. \tag{2.12}$$

When Assumption 1.1(iii) does not hold, there exist some $j \le r$ and two nonzero elements $u_{ji_1}, u_{ji_2}$. Now we take any $e \in \mathbb{R}^m$ such that $v_j^t e = 0$ and $(Ae)_{i_1} \ne 0$ or $(Ae)_{i_2} \ne 0$. Then the left-hand side of (2.12) is strictly positive, and the right-hand side vanishes. Thus, there is no constant $c$ such that this inequality holds. This shows the delicacy of the analysis for a general matrix $A$. Nonetheless, the numerical experiments in section 4 indicate that the saturation phenomenon actually also does not occur for a general matrix as long as the stepsize $c_0$ is sufficiently small. Thus, we believe that the restriction is due to the limitation of the proof technique. Note that the convergence analysis in section 3 remains valid provided that relaxed versions of the identities (2.8) and (2.9) hold but with different constants in the final estimate.

The proof of Theorem 2.5 also gives the following error decomposition for exact data $y^\dagger$.

Corollary 2.8. *For any $0 \le \ell < k$, the following error decomposition holds:*

$$\mathbb{E}[\|e_{k+1}\|^2] \le \sum_{i=0}^{\ell} \mathrm{I}_i + (\mathrm{I}_\ell)^c, \tag{2.13}$$

*where the terms $\mathrm{I}_i$, $i = 0, 1, \ldots, \ell$, are defined by*

$$\mathrm{I}_0 = \|\Pi_{J^c_{[1,k],0}}(B)e_1\|^2,$$

$$\mathrm{I}_i = (n-1)^i \sum_{J_i \in \mathcal{J}_{[1,k],i}} \prod_{t=1}^{i} \eta_{j_t}^2 \|\Pi_{J^c_{[1,k],i}}(B)B^i e_1\|^2 \quad \forall 1 \le i \le \ell,$$

$$(\mathrm{I}_\ell)^c = (n-1)^{\ell+1} \sum_{J_{\ell+1} \in \mathcal{J}_{[1,k],\ell+1}} \prod_{t=1}^{\ell+1} \eta_{j_t}^2 \mathbb{E}[\|\Pi_{J^c_{[j_{\ell+1}+1,k],\ell}}(B)B^{\ell+1}e_{j_{\ell+1}}\|^2].$$

In view of Theorem 2.5, the error $\mathbb{E}[\|e_{k+1}^\delta\|^2]$ can be decomposed into three components: approximation error $\sum_{i=0}^{\ell} \mathrm{I}_{i,1}^\delta$, propagation error $\sum_{i=0}^{\ell} \mathrm{I}_{i,2}^\delta$, and stochastic error $(\mathrm{I}_\ell^\delta)^c$. Here we have slightly abused the terminology for approximation and propagation errors, since the approximation error only depends on the regularity of the exact solution $x^\dagger$ (indicated by the source condition (1.5) in Assumption 1.1(ii)), whereas the propagation error is determined by the noise level. With the choice $\ell = 0$, the decomposition recovers that in [18, 19]. When compared with the classical error decomposition for the Landweber method, the summands for $\ell \ge 1$ arise from the stochasticity of the iterates (due to the random row index at each iteration), and so does the stochastic error $(\mathrm{I}_\ell^\delta)^c$. This refined decomposition is crucial in analyzing the saturation phenomenon (under suitable conditions on the initial stepsize). Below we first derive bounds on the first two terms in Propositions 3.1 and 3.5, and then we prove optimal convergence rates of SGD by mathematical induction in section 3.3.

## 3. Convergence rate analysis.

In this section, we present the convergence rate analysis and establish Theorem 1.2. The proof proceeds by first analyzing the approximation error and propagation error in sections 3.1 and 3.2, respectively, and then bounding the mean squared error $\mathbb{E}[\|e_k^\delta\|^2]$ via mathematical induction. We also give an alternative (simplified) convergence analysis for the case $\alpha = 0$ in section 3.4.

### 3.1. Bound on the approximation error.

We begin with bounding the approximation error $\sum_{i=0}^{\ell} \mathrm{I}_{i,1}^\delta$ for any fixed $\ell \ge \nu$. The summand $\mathrm{I}_{0,1}^\delta$ is the usual approximation error (for the Landweber method), and the remaining terms arise from the random row index. Thus, the approximation error decays at the optimal rate.

Proposition 3.1. *Let Assumption 1.1 be fulfilled, and*

$$h_0(k) = 2(\nu+\ell)^2 n\phi(2\alpha)k^{-2(1-\alpha)+\max(1-2\alpha,0)}.$$

*Then for any integer $\ell \ge \nu$, $\alpha \in [0, 1)$, and $k \ge 2\ell$, there holds*

$$\sum_{i=0}^{\ell} \mathrm{I}_{i,1}^\delta \le c_{\nu,\ell,\alpha,n} c_0^{-2\nu} k^{-2\nu(1-\alpha)} \|w\|^2,$$

*with the constant*

$$c_{\nu,\ell,\alpha,n} = \begin{cases} 4(\nu+\ell)^{2\nu} & \text{if } h_0(k) \le \frac{1}{2}, \\ 2(\nu+\ell)^{2\nu} \sum_{i=0}^{\ell} (h_0(2\ell))^i & \text{otherwise.} \end{cases}$$

*Proof.* In view of the source condition (1.5) and Lemma 2.3, we have

$$\mathrm{I}_{0,1}^{\delta} = 2\|\Pi_{J^c_{[1,k],0}}(B)e_1\|^2 \le 2\|\Pi_{J^c_{[1,k],0}}(B)B^{\nu}\|^2\|w\|^2 \le 2\nu^{2\nu}(ec_0)^{-2\nu}k^{-2\nu(1-\alpha)}\|w\|^2.$$

Similarly, for any $1 \le i \le \ell$,

$$\begin{aligned}\prod_{t=1}^{i}\eta_{j_t}^2\|\Pi_{J^c_{[1,k],i}}(B)B^i e_1\|^2 &\le \prod_{t=1}^{i}\eta_{j_t}^2\|\Pi_{J^c_{[1,k],i}}(B)B^{\nu+i}\|^2\|w\|^2 \\ &\le (\tfrac{\nu+i}{e})^{2(\nu+i)}c_0^{-2\nu}k^{2(\nu+i)\alpha}\|w\|^2\prod_{t=1}^{i}j_t^{-2\alpha}(k-i)^{-2(\nu+i)}.\end{aligned}$$

By the definition of $\mathrm{I}_{i,1}^{\delta}$, since $k \ge 2\ell$, $k - i \ge \frac{k}{2}$ for $i = 1, \dots, \ell$, by Lemma 2.4(i),

$$\begin{aligned}\mathrm{I}_{i,1}^{\delta} &\le 2^{i+1}n^i(\tfrac{\nu+i}{e})^{2(\nu+i)}c_0^{-2\nu}k^{2(\nu+i)\alpha}\|w\|^2\sum_{J_i\in\mathcal{J}_{[1,k],i}}\prod_{t=1}^{i}j_t^{-2\alpha}(k-i)^{-2(\nu+i)} \\ &\le 2(2e^{-1})^{2\nu}(\nu+i)^{2\nu}c_0^{-2\nu}k^{-2\nu(1-\alpha)}\|w\|^2\Big[8e^{-2}(\nu+i)^2n\phi(2\alpha)k^{-2(1-\alpha)+\max(1-2\alpha,0)}\Big]^i.\end{aligned}$$

Clearly, the quantity in the square brackets is bounded by $h_0(k)$. Next we treat the two cases $h_0(k) \le \frac{1}{2}$ and $h_0(k) > \frac{1}{2}$ separately. If $h_0(k) \le \frac{1}{2}$, we deduce

$$\sum_{i=0}^{\ell}\mathrm{I}_{i,1}^{\delta} \le 2(\nu+\ell)^{2\nu}c_0^{-2\nu}k^{-2\nu(1-\alpha)}\|w\|^2\sum_{i=0}^{\ell}h_0(k)^i \le c_{\nu,\ell,\alpha,n}c_0^{-2\nu}k^{-2\nu(1-\alpha)}\|w\|^2.$$

Further, when $h_0(k) > \frac{1}{2}$, since $k \ge 2\ell$, we have $h_0(k) \le h_0(2\ell)$, and thus we obtain

$$\sum_{i=0}^{\ell}\mathrm{I}_{i,1}^{\delta} \le 2(\nu+\ell)^{2\nu}c_0^{-2\nu}k^{-2\nu(1-\alpha)}\|w\|^2\sum_{i=0}^{\ell}h_0(2\ell)^i \le c_{\nu,\ell,\alpha,n}c_0^{-2\nu}k^{-2\nu(1-\alpha)}\|w\|^2.$$

Finally, combining the last two estimates completes the proof. ∎

*Remark* 3.2. For any $k$ satisfying $h_0(k) \le \frac{1}{2}$, the constant $c_{\nu,\ell,\alpha,n}$ is actually independent of $\alpha$ and $n$. Further, if $k < 2\ell$, then by setting $\ell$ to 0, we obtain

$$\mathrm{I}_{0,1}^{\delta} \le 2^{1-2\nu}\nu^{2\nu}c_0^{-2\nu}k^{-2\nu(1-\alpha)}\|w\|^2.$$

**3.2. Bound on the propagation error.** Now we bound the propagation error $\sum_{i=0}^{\ell} \mathrm{I}_{i,2}^{\delta}$, which arises from the presence of the data noise $\xi$. The summands for $\ell \geq 1$ arise from the stochasticity of the SGD iterates $x_k^{\delta}$. We bound each summand $\mathrm{I}_{i,2}^{\delta}$, $i = 0, \ldots, \ell$, separately, or equivalently the following two quantities for $k \geq 4i$:

$$\mathrm{I}(i,k) := \sum_{J_i \in \mathcal{J}_{[2,k],i}} \prod_{t=1}^{i} \eta_{j_t}^2 \Big\| \sum_{j_{i+1}=1}^{j_i-1} \eta_{j_{i+1}} \Pi_{J^c_{[j_{i+1}+1,k],i}}(B) B^{i+\frac{1}{2}} \Big\|^2, \tag{3.1}$$

$$\mathrm{II}(i,k) := \sum_{J_i \in \mathcal{J}_{[2,k],i}} \prod_{t=1}^{i} \eta_{j_t}^2 \sum_{j_{i+1}=1}^{j_i-1} \eta_{j_{i+1}}^2 \|\Pi_{J^c_{[j_{i+1}+1,k],i}}(B) B^{i+\frac{1}{2}}\|^2, \tag{3.2}$$

with the convention $\sum_{J_0 \in \mathcal{J}_{[2,k],0}} \prod_{t=1}^{0} \eta_{j_t}^2 = 1$ and $j_0 = k+1$. The condition $k \geq 4i$ implies the following two elementary estimates:

$$k - j_{i+1} - i \geq \tfrac{k}{4}, \quad j_{i+1} = 1, 2, \ldots, [\tfrac{k}{2}], \tag{3.3}$$

$$k - j_{i+1} \leq (i+1)(k - j_{i+1} - i), \quad j_{i+1} = [\tfrac{k}{2}] + 1, \ldots, k - i - 1. \tag{3.4}$$

First we bound $\mathrm{I}(i,k)$. The notation $[\cdot]$ denotes taking the integral part of a real number.

Lemma 3.3. *Let $\mathrm{I}(i,k)$ be defined as in (3.1), and let Assumption 1.1 be fulfilled. Then for any fixed $i \in \mathbb{N}$ and $k \geq 4i$, the following estimate holds:*

$$\begin{aligned}\mathrm{I}(i,k) \leq & 2\Big(2^{2\alpha-1} e^{-1} (2i+1) c_0 \big((2e^{-1})^2 (2i+1)^2 \phi(2\alpha) k^{-2(1-\alpha)+\max(1-2\alpha,0)}\big)^i \\ & + 25 c_0^{\max(i,1)} \big(2^{2\alpha-1} e^{-1} (i+2)^2 k^{-\alpha}\big)^i\Big) \phi(\alpha)^2 k^{1-\alpha}.\end{aligned}$$

*Proof.* We abbreviate $\mathrm{I}(i,k)$ as I. By the triangle inequality and Lemma 2.3, for any $s \in (0, i+\frac{1}{2}]$ and any $j_{i+1} \neq k-i$ (when $j_{i+1} = k-i$, $j_i = k-i+1, \ldots, j_1 = k$), we have

$$\begin{aligned}\Big\| \sum_{j_{i+1}=1}^{j_i-1} \eta_{j_{i+1}} \Pi_{J^c_{[j_{i+1}+1,k],i}}(B) B^s \Big\| \leq & \sum_{j_{i+1}=1}^{j_i-1} \eta_{j_{i+1}} \|\Pi_{J^c_{[j_{i+1}+1,k],i}}(B) B^s\| \\ \leq & s^s (ec_0)^{-s} k^{\alpha s} \sum_{j_{i+1}=1}^{j_i-1} \eta_{j_{i+1}} (k - j_{i+1} - i)^{-s}.\end{aligned}$$

By the identity (2.1), and since the quantity $\big(\sum_{j_{i+1}=1}^{k-i-1} \eta_{j_{i+1}} s^s (ec_0)^{-s} (k - j_{i+1} - i)^{-s} k^{\alpha s}\big)^2$ is independent of the indices $\{j_1, \ldots, j_i\}$, there holds

$$\begin{aligned}\mathrm{I}^{\frac{1}{2}} \leq & \Big( \sum_{J_{i+1} \in \mathcal{J}_{[k-i,k],i+1}} \prod_{t=1}^{i+1} \eta_{j_t}^2 \|B^s\|^2 \Big)^{\frac{1}{2}} \\ & + s^s (ec_0)^{-s} k^{\alpha s} \sum_{j_{i+1}=1}^{k-i-1} \eta_{j_{i+1}} (k - j_{i+1} - i)^{-s} \Big( \sum_{J_i \in \mathcal{J}_{[j_{i+1}+1,k],i}} \prod_{t=1}^{i} \eta_{j_t}^2 \Big)^{\frac{1}{2}} \\ = & c_0^{i+1} \prod_{j=k-i}^{k} j^{-\alpha} \|B^s\| + s^s (ec_0)^{-s} k^{\alpha s} \sum_{j_{i+1}=1}^{k-i-1} \eta_{j_{i+1}} (k - j_{i+1} - i)^{-s} \Big( \sum_{J_i \in \mathcal{J}_{[j_{i+1}+1,k],i}} \prod_{t=1}^{i} \eta_{j_t}^2 \Big)^{\frac{1}{2}}.\end{aligned}$$

The two terms on the right are denoted by $\mathrm{I}_0'$ and $\mathrm{I}'$. For $i \geq 1$, setting $s = \frac{i}{2} + 1 \leq i + \frac{1}{2}$ in the first term, the inequalities $k - i \geq \frac{3}{4}k$ and $c_0\|B\| \leq (2e)^{-1}$ imply that

$$\mathrm{I}_0' \leq c_0^{\frac{i}{2}} (2e)^{-\frac{i}{2}-1} (\tfrac{4}{3})^{(i+1)\alpha} k^{-(i+1)\alpha} \leq e^{-1} \big(2^{2\alpha-1} e^{-1} c_0 k^{-\alpha}\big)^{\frac{i}{2}}.$$

Likewise, for $i = 0$, setting $s = i + \frac{1}{2}$ gives $\mathrm{I}_0' \leq (2e)^{-\frac{1}{2}} c_0^{\frac{1}{2}} k^{-\alpha}$. Next we split $\mathrm{I}'$ into two summations $\mathrm{I}_1'$ and $\mathrm{I}_2'$ over the index $j_{i+1}$, one from 1 to $[\frac{k}{2}]$ and the other from $[\frac{k}{2}] + 1$ to $k - i - 1$, respectively. It suffices to bound $\mathrm{I}_1'$ and $\mathrm{I}_2'$. First, setting $s$ to $i + \frac{1}{2}$ in $\mathrm{I}_1'$ and then applying the inequality

$$\sum_{J_i \in \mathcal{J}_{[j_{i+1}+1,k],i}} \prod_{t=1}^{i} j_t^{-2\alpha} \leq \sum_{J_i \in \mathcal{J}_{[1,k],i}} \prod_{t=1}^{i} j_t^{-2\alpha}$$

and the estimate (3.3) lead to

$$\mathrm{I}_1' \leq (\tfrac{2i+1}{2e})^{i+\frac{1}{2}} c_0^{\frac{1}{2}} k^{(i+\frac{1}{2})\alpha} (\tfrac{k}{4})^{-(i+\frac{1}{2})} \Big( \sum_{j_{i+1}=1}^{[\frac{k}{2}]} j_{i+1}^{-\alpha} \Big) \Big( \sum_{J_i \in \mathcal{J}_{[1,k],i}} \prod_{t=1}^{i} j_t^{-2\alpha} \Big)^{\frac{1}{2}}.$$

Then by Lemma 2.4(i) and the estimate (2.2), we obtain

$$\mathrm{I}_1' \leq 2^{\alpha-1} (2e^{-1})^{\frac{1}{2}} (2i+1)^{\frac{1}{2}} c_0^{\frac{1}{2}} \phi(\alpha) \big( (2e^{-1})^2 (2i+1)^2 \phi(2\alpha) k^{-2(1-\alpha)+\max(1-2\alpha,0)} \big)^{\frac{i}{2}} k^{\frac{1-\alpha}{2}}.$$

For the term $\mathrm{I}_2'$, we analyze the cases $i = 0$ and $i \geq 1$ separately. Since $c_0\|B\| \leq (2e)^{-1}$ (cf. Assumption 1.1), if $i = 0$, then Lemma 2.3 with $s = \frac{1}{2}$ gives

$$\mathrm{I}_2' \leq (\tfrac{1}{2e})^{\frac{1}{2}} c_0^{\frac{1}{2}} k^{\frac{\alpha}{2}} \sum_{j=[\frac{k}{2}]+1}^{k-1} j^{-\alpha} (k-j)^{-\frac{1}{2}}.$$

Now the estimate (2.2) implies

$$\sum_{j=[\frac{k}{2}]+1}^{k-1} j^{-\alpha} (k-j)^{-\frac{1}{2}} \leq (\tfrac{k}{2})^{-\alpha} 2 (\tfrac{k}{2})^{\frac{1}{2}} \leq 2 (\tfrac{k}{2})^{\frac{1}{2}-\alpha}.$$

Consequently, when $i = 0$, we have

$$\mathrm{I}_2' \leq 2^{\alpha} e^{-\frac{1}{2}} c_0^{\frac{1}{2}} k^{\frac{1-\alpha}{2}} \quad \text{and} \quad \mathrm{I}_2' + \mathrm{I}_0' \leq 2^{\alpha+1} e^{-\frac{1}{2}} c_0^{\frac{1}{2}} k^{\frac{1-\alpha}{2}}.$$

Meanwhile, when $i \geq 1$, setting $s = \frac{i}{2} + 1 \leq i + \frac{1}{2}$ in Lemma 2.3 gives

$$\mathrm{I}_2' \leq (\tfrac{\frac{i}{2}+1}{e})^{\frac{i}{2}+1} c_0^{\frac{i}{2}} k^{\alpha(\frac{i}{2}+1)} (\tfrac{k}{2})^{-(i+1)\alpha} \mathrm{I}_2'',$$

with

$$\mathrm{I}_2'' := \sum_{j_{i+1}=[\frac{k}{2}]+1}^{k-i-1} (k-j_{i+1}-i)^{-(\frac{i}{2}+1)} \Big( \sum_{J_i \in \mathcal{J}_{[j_{i+1}+1,k],i}} \prod_{t=1}^{i} 1 \Big)^{\frac{1}{2}}.$$

Now Lemma 2.4(ii) and the estimates (3.4) and (2.2) yield

$$\begin{aligned} \mathrm{I}_2'' &\le \frac{1}{i!^{\frac{1}{2}}} \sum_{j_{i+1}=[\frac{k}{2}]+1}^{k-i-1} (k-j_{i+1}-i)^{-(\frac{i}{2}+1)} (k-j_{i+1})^{\frac{i}{2}} \\ &\le \frac{(i+1)^{\frac{i}{2}}}{i!^{\frac{1}{2}}} \sum_{j_{i+1}=[\frac{k}{2}]+1}^{k-i-1} (k-j_{i+1}-i)^{-1} \le \frac{2(i+1)^{\frac{i}{2}}}{i!^{\frac{1}{2}}} \max(\ln k, 1). \end{aligned}$$

Combining the last two identities gives

$$\mathrm{I}_2' \le 2^{\alpha}(i+2) i!^{-\frac{1}{2}} e^{-1} \max(\ln k, 1) \big(2^{2\alpha-1} e^{-1} c_0 (i+2)^2 k^{-\alpha}\big)^{\frac{i}{2}}, \quad i \ge 1.$$

Now by the estimate $\sup_{i\in\mathbb{N}} \frac{i+2}{i!^{\frac{1}{2}}} \le 3$ and the elementary inequality (for $s \in (0,1]$)

$$k^{-s} \max(\ln k, 1) \le s^{-1}, \tag{3.5}$$

with $s = \frac{1-\alpha}{2}$, we obtain

$$\mathrm{I}_2' \le 12 e^{-1} \phi(\alpha) \big(2^{2\alpha-1} e^{-1} c_0 (i+2)^2 k^{-\alpha}\big)^{\frac{i}{2}} k^{\frac{1-\alpha}{2}}, \quad i \ge 1,$$

and thus for $i \ge 1$, there holds

$$\mathrm{I}_2' + \mathrm{I}_0' \le 13 e^{-1} \phi(\alpha) \big(2^{2\alpha-1} e^{-1} c_0 (i+2)^2 k^{-\alpha}\big)^{\frac{i}{2}} k^{\frac{1-\alpha}{2}} \le 5\phi(\alpha) \big(2^{2\alpha-1} e^{-1} c_0 (i+2)^2 k^{-\alpha}\big)^{\frac{i}{2}} k^{\frac{1-\alpha}{2}}.$$

The bounds on $\mathrm{I}_1'$ and $\mathrm{I}_2' + \mathrm{I}_0'$ and the triangle inequality complete the proof. ∎

The next result bounds the quantity $\mathrm{II}(i,k)$.

Lemma 3.4. *Let $\mathrm{II}(i,k)$ be defined as in (3.2), and let Assumption 1.1 hold. Then for any fixed $i \in \mathbb{N}$, and for $k \ge 4i$, the following estimate holds:*

$$\begin{aligned} \mathrm{II}(i,k) \le \Big( &\tfrac{ec_0}{2(2i+1)} (4e^{-2}(2i+1)^2 \phi(2\alpha) k^{-2(1-\alpha)+\max(1-2\alpha,0)})^{i+1} \\ &+ 3\phi(\alpha) \big(2^{2\alpha-1} e^{-1} c_0 (i+1)^2 k^{-\alpha}\big)^{i+1} \Big) k^{1-\alpha}. \end{aligned}$$

*Proof.* As before, we abbreviate $\mathrm{II}(i,k)$ to II. By (2.1), II can be rewritten as

$$\mathrm{II} = \sum_{j_{i+1}=1}^{k-i} \sum_{J_i \in \mathcal{J}_{[j_{i+1}+1,k],i}} \prod_{t=1}^{i} \eta_{j_t}^2 \eta_{j_{i+1}}^2 \|\Pi_{J_{[j_{i+1}+1,k],i}^c}(B) B^{i+\frac{1}{2}}\|^2.$$

Now we split the summation into three terms, i.e., $j_{i+1} = k-i$, one from $j_{i+1} = 1$ to $[\frac{k}{2}]$ and one from $j_{i+1} = [\frac{k}{2}]+1$ to $k-i-1$, denoted by $\mathrm{II}_0$, $\mathrm{II}_1$, and $\mathrm{II}_2$, respectively. Since $k-i \geq \frac{3}{4}k$, $\|B\| \leq 1$, and $c_0\|B\| \leq (2e)^{-1}$ (cf. Assumption 1.1(i)), we obtain that, for any $i \geq 0$,

$$\mathrm{II}_0 = c_0^{2i+2} \prod_{j=k-i}^{k} j^{-2\alpha} \|B^{i+\frac{1}{2}}\|^2 \leq c_0^{i+1}(c_0\|B\|)^{i+1}(k-i)^{-2(i+1)\alpha} \leq \left(2^{2\alpha-1}e^{-1}c_0k^{-2\alpha}\right)^{i+1}.$$

By Lemma 2.3 with $s = i+\frac{1}{2}$ and (3.3),

$$\begin{aligned}\mathrm{II}_1 \leq & (\tfrac{2i+1}{2e})^{2i+1} c_0 k^{(2i+1)\alpha} \sum_{j_{i+1}=1}^{[\frac{k}{2}]} (k-j_{i+1}-i)^{-(2i+1)} \sum_{J_i \in \mathcal{J}_{[j_{i+1}+1,k],i}} \prod_{t=1}^{i+1} j_t^{-2\alpha} \\ \leq & (\tfrac{2i+1}{2e})^{2i+1} c_0 k^{(2i+1)\alpha} (\tfrac{k}{4})^{-(2i+1)} \sum_{j_{i+1}=1}^{[\frac{k}{2}]} \sum_{J_i \in \mathcal{J}_{[j_{i+1}+1,k],i}} \prod_{t=1}^{i+1} j_t^{-2\alpha}.\end{aligned}$$

Meanwhile, Lemma 2.4 and the estimate (2.2) imply

$$\sum_{j_{i+1}=1}^{[\frac{k}{2}]} \sum_{J_i \in \mathcal{J}_{[j_{i+1}+1,k],i}} \prod_{t=1}^{i+1} j_t^{-2\alpha} \leq (\phi(2\alpha)k^{\max(1-2\alpha,0)})^{i+1}.$$

The last two estimates together imply

$$\mathrm{II}_1 \leq \tfrac{ec_0}{2(2i+1)} (4e^{-2}(2i+1)^2\phi(2\alpha)k^{-2(1-\alpha)+\max(1-2\alpha,0)})^{i+1} k^{1-\alpha}.$$

Now we bound the term $\mathrm{II}_2$. In this case, we analyze the cases $i = 0$ and $i \geq 1$ separately. When $i = 0$, by Lemma 2.3,

$$\mathrm{II}_2 = \sum_{j=[\frac{k}{2}]+1}^{k-1} \eta_j^2 \|\Pi_{J^c_{[j+1,k],0}}(B)B^{\frac{1}{2}}\|^2 \leq \frac{c_0k^\alpha}{2e} \sum_{j=[\frac{k}{2}]+1}^{k-1} j^{-2\alpha}(k-j)^{-1}.$$

The estimates (2.2) and (3.5) with $s = 1-\alpha$ imply

$$\sum_{j=[\frac{k}{2}]+1}^{k-1} j^{-2\alpha}(k-j)^{-1} \leq 2(\tfrac{k}{2})^{-2\alpha} \max(\ln k, 1) \leq 2^{2\alpha+1}\phi(\alpha)k^{1-3\alpha}.$$

The last two estimates together show that for $i = 0$, there holds

$$\mathrm{II}_2 \leq 2^{2\alpha}e^{-1}c_0\phi(\alpha)k^{1-2\alpha}.$$

Next, when $i \geq 1$, by Lemma 2.3 with $s = \frac{i+1}{2}$,

$$\begin{aligned}\mathrm{II}_2 \leq & \sum_{j_{i+1}=[\frac{k}{2}]+1}^{k-i-1} \sum_{J_i \in \mathcal{J}_{[j_{i+1}+1,k],i}} \prod_{t=1}^{i} \eta_{j_t}^2 \eta_{j_{i+1}}^2 \|\Pi_{J^c_{[j_{i+1}+1,k],i}}(B)B^{\frac{i+1}{2}}\|^2 \\ \leq & (\tfrac{i+1}{2e})^{i+1} k^{(i+1)\alpha} c_0^{i+1} (\tfrac{k}{2})^{-2(i+1)\alpha} \sum_{j_{i+1}=[\frac{k}{2}]+1}^{k-i-1} \sum_{J_i \in \mathcal{J}_{[j_{i+1}+1,k],i}} (k-j_{i+1}-i)^{-(i+1)}.\end{aligned}$$

Now Lemma 2.4(ii) and (2.2) imply

$$\sum_{j_{i+1}=[\frac{k}{2}]+1}^{k-i-1}\sum_{J_i\in\mathcal{J}_{[j_{i+1}+1,k],i}}(k-j_{i+1}-i)^{-(i+1)}\le\sum_{j_{i+1}=[\frac{k}{2}]+1}^{k-i-1}(k-j_{i+1}-i)^{-(i+1)}\frac{(k-j_{i+1})^i}{i!}$$
$$\le\frac{(i+1)^i}{i!}\sum_{j_{i+1}=[\frac{k}{2}]+1}^{k-i-1}(k-j_{i+1}-i)^{-1}\le\frac{2(i+1)^i}{i!}\max(\ln k,1).$$

Combining the last two bounds with (3.5) with $s=1-\alpha$ leads to

$$\mathrm{II}_2\le\frac{2}{i!}\phi(\alpha)\big(2^{2\alpha-1}e^{-1}c_0(i+1)^2k^{-\alpha}\big)^{i+1}k^{1-\alpha},\quad i\ge1.$$

Clearly, the preceding discussion shows that the last inequality holds actually also for $i=0$. Therefore, the bounds on $\mathrm{II}_0$, $\mathrm{II}_1$, and $\mathrm{II}_2$ complete the proof of the lemma. ∎

Now we can bound the propagation error $\sum_{i=0}^{\ell}\mathrm{I}_{2,i}^{\delta}$. The bound is largely comparable with that for the Landweber method [18, Theorem 3.2].

Proposition 3.5. *Let Assumption 1.1 be fulfilled, and let*

$$h_1(k)=2(2\ell+1)^2n\phi(2\alpha)k^{-2(1-\alpha)+\max(1-2\alpha,0)}\quad\text{and}\quad h_2(k)=2^{2\alpha-1}(\ell+2)^2nc_0k^{-\alpha}.$$

*Then for any fixed $\ell\in\mathbb{N}$, and $k\ge4\ell$, there holds*

$$\sum_{i=0}^{\ell}\mathrm{I}_{i,2}^{\delta}\le c_{\ell,\alpha,n,c_0}\bar{\delta}^2k^{1-\alpha},$$

*with the constant $c_{\ell,\alpha,n,c_0}$ given by*

$$c_{\ell,\alpha,n,c_0}=\begin{cases}(2^4(\ell+1)c_0+203)\phi(\alpha)^2 & \text{if } h_1(k),h_2(k)\le\frac12,\\ \Big(8(\ell+1)c_0\sum_{i=0}^{\ell+1}h_1(4\ell)^i+103\sum_{i=0}^{\ell+1}h_2(4\ell)^i\Big)\phi(\alpha)^2 & \text{otherwise.}\end{cases}$$

*Proof.* For $i=0,1,\dots,\ell$, we bound the summands $\mathrm{I}_{i,2}^{\delta}$ by

$$\mathrm{I}_{i,2}^{\delta}\le2^{i+1}n^i\bar{\delta}^2\sum_{J_i\in\mathcal{J}_{[2,k],i}}\prod_{t=1}^{i}\eta_{j_t}^2\|\sum_{j_{i+1}=1}^{j_i-1}\eta_{j_{i+1}}\Pi_{J^c_{[j_{i+1}+1,k],i}}(B)B^{i+\frac12}\|^2$$
$$+2^{i+1}n^{i+1}\bar{\delta}^2\sum_{J_i\in\mathcal{J}_{[2,k],i}}\prod_{t=1}^{i}\eta_{j_t}^2\sum_{j_{i+1}=1}^{j_i-1}\eta_{j_{i+1}}^2\|\Pi_{J^c_{[j_{i+1}+1,k],i}}(B)B^{i+\frac12}\|^2.$$

The two terms on the right-hand side, denoted by $\mathrm{I}_{i,2,1}^{\delta}$ and $\mathrm{I}_{i,2,2}^{\delta}$, can be bounded using Lemmas 3.3 and 3.4, respectively. Indeed, for $\mathrm{I}_{i,2,1}^{\delta}$, Lemma 3.3 yields that for any $k\ge4\ell\ge4i$,

$$\mathrm{I}_{i,2,1}^{\delta}\le4\Big(2^{2\alpha-1}e^{-1}(2i+1)c_0\big(2^3e^{-2}(2i+1)^2n\phi(2\alpha)k^{-2(1-\alpha)+\max(1-2\alpha,0)}\big)^i$$
$$+25\big(2^{2\alpha}e^{-1}(i+2)^2nc_0k^{-\alpha}\big)^i\Big)\phi(\alpha)^2\bar{\delta}^2k^{1-\alpha}.$$

Thus, by the definitions of $h_1(k), h_2(k)$, for any $k \geq 4\ell$, if $h_1(k) \leq \frac{1}{2}$ and $h_2(k) \leq \frac{1}{2}$, then the condition $i \leq \ell$ implies

$$\begin{aligned}\sum_{i=0}^{\ell} \mathrm{I}_{i,2,1}^{\delta} &\leq 4\Big(2^{2\alpha-1}e^{-1}(2\ell+1)c_0 \sum_{i=0}^{\ell} h_1(k)^i + 25\sum_{i=0}^{\ell} h_2(k)^i\Big)\phi(\alpha)^2\bar{\delta}^2 k^{1-\alpha} \\ &\leq 4\big(2^{2\alpha}e^{-1}(2\ell+1)c_0 + 50\big)\phi(\alpha)^2\bar{\delta}^2 k^{1-\alpha}.\end{aligned}$$

Meanwhile, if $k$ does not satisfy the condition, by the monotonicity of $h_1(k)$ and $h_2(k)$ in $k$, we have $h_1(k) \leq h_1(4\ell)$ and $h_2(k) \leq h_2(4\ell)$, and consequently,

$$\sum_{i=0}^{\ell} \mathrm{I}_{i,2,1}^{\delta} \leq 4\Big(2^{2\alpha-1}e^{-1}(2\ell+1)c_0 \sum_{i=0}^{\ell} h_1(4\ell)^i + 25\sum_{i=0}^{\ell} h_2(4\ell)^i\Big)\phi(\alpha)^2\bar{\delta}^2 k^{1-\alpha}.$$

Next we bound the term $\mathrm{I}_{i,2,2}^{\delta}$. Actually, by Lemma 3.4, for any $k \geq 4\ell \geq 4i$, there holds

$$\begin{aligned}\mathrm{I}_{i,2,2}^{\delta} \leq \Big(&\tfrac{ec_0}{2(2i+1)}\big(2^3e^{-2}(2i+1)^2 n\phi(2\alpha)k^{-2(1-\alpha)+\max(1-2\alpha,0)}\big)^{i+1} \\ &+ 3\phi(\alpha)\big(2^{2\alpha}e^{-1}(i+1)^2 nc_0 k^{-\alpha}\big)^{i+1}\Big)\bar{\delta}^2 k^{1-\alpha}.\end{aligned}$$

Then repeating the preceding arguments yields

$$\sum_{i=0}^{\ell} \mathrm{I}_{i,2,2}^{\delta} \leq \begin{cases} \big(2c_0 + 3\phi(\alpha)\big)\bar{\delta}^2 k^{1-\alpha} & \text{if } h_1(k), h_2(k) \leq \frac{1}{2}, \\ \Big(2c_0 \displaystyle\sum_{i=0}^{\ell} h_1(4\ell)^{i+1} + 3\phi(\alpha)\sum_{i=0}^{\ell}\big(h_2(4\ell)\big)^{i+1}\Big)\bar{\delta}^2 k^{1-\alpha} & \text{otherwise.}\end{cases}$$

Now combining the bounds on $\sum_{i=0}^{\ell} \mathrm{I}_{i,2,1}^{\delta}$ and $\sum_{i=0}^{\ell} \mathrm{I}_{i,2,2}^{\delta}$ yields the desired assertion. ∎

*Remark* 3.6. If $k < 4\ell$, we can replace $\ell$ by 0. By Assumption 1.1(i), $c_0 < 1$, repeating the argument of the proposition and Lemmas 3.3 and 3.4 yields

$$\mathrm{I}_{0,2}^{\delta} \leq 2c_0\big(n(\phi(2\alpha) + 3\phi(\alpha)) + 11\phi(\alpha)^2\big)\bar{\delta}^2 k^{1-\alpha}.$$

Note that in the conditions $h_0(k), h_1(k), h_2(k) \leq \frac{1}{2}$, $h_0$ and $h_1$, apart from the factor $\phi(2\alpha)$, do not depend sensitively on the exponent $\alpha$, but for $\alpha$ close to zero, $h_2(k) \leq \frac{1}{2}$ essentially requires a small $c_0 = O(n^{-1})$, and furthermore, the larger $\ell$ is, the smaller $c_0$ should be in order to fulfill the conditions. The latter condition also appears in the proof of Theorem 1.2 below.

**3.3. Bound on the error $\mathbb{E}[\|e_k^{\delta}\|^2]$.** To prove Theorem 1.2, we need a useful technical estimate, where the notation $k^{\max(0,0)}$ denotes $\ln k$. Note that the restricted range of $s$ is sufficient for the proof of Theorem 1.2.

Lemma 3.7. *Let Assumption 1.1 hold. Then for any $\epsilon, \eta \in [0,1]$, $s \in (-\infty, 0] \cup (\max(0, 1-2\alpha), +\infty)$, and $k \geq 4\ell$ with $\ell_\epsilon = \epsilon(\ell+1)$ and $\ell_\eta = \eta(\ell+1)$, the following two estimates hold:*

$$\begin{aligned}(3.6)\qquad &\sum_{J_{\ell+1}\in\mathcal{J}_{[k-\ell,k],\ell+1}} \prod_{t=1}^{\ell+1} \eta_{j_t}^2 \|B^{\ell+1}\|^2 (k-\ell)^{-s} \\ &\leq \max(2^s, 1)(2^{2\alpha-2\eta} e^{-2\eta} c_0^{2-2\eta} k^{-2\alpha})^{\ell+1} k^{-s},\end{aligned}$$

$$\begin{aligned}(3.7)\qquad &\sum_{j_{\ell+1}=1}^{[\frac{k}{2}]} \sum_{J_\ell\in\mathcal{J}_{[j_{\ell+1}+1,k],\ell}} \prod_{t=1}^{\ell+1} \eta_{j_t}^2 \|\Pi_{J^c_{[j_{\ell+1}+1,k],\ell}}(B) B^{\ell+1}\|^2 j_{\ell+1}^{-s} \\ &\leq c_s \Big( \big(4e^{-1}\ell_\epsilon\big)^{2\epsilon} \phi(2\alpha) c_0^{2-2\epsilon} k^{-s_\epsilon(s)} \Big)^{\ell+1} k^{-s},\end{aligned}$$

$$\begin{aligned}(3.8)\qquad &\sum_{j_{\ell+1}=[\frac{k}{2}]+1}^{k-\ell-1} \sum_{J_\ell\in\mathcal{J}_{[j_{\ell+1}+1,k],\ell}} \prod_{t=1}^{\ell+1} \eta_{j_t}^2 \|\Pi_{J^c_{[j_{\ell+1}+1,k],\ell}}(B) B^{\ell+1}\|^2 j_{\ell+1}^{-s} \\ &\leq \tfrac{\max(2^s,1)\phi(2\ell_\eta-\ell)}{(\ell+1)!} (2^{2\alpha} e^{-2\eta} (\ell+1) \ell_\eta^{2\eta} c_0^{2-2\eta} k^{-s_\eta})^{\ell+1} k^{-s},\end{aligned}$$

*with the constant $c_s$, and the exponents $s_\epsilon(s)$ and $s_\eta$ respectively defined by*

$$\begin{aligned} c_s &= \begin{cases} 2^s & \text{if } s \leq 0, \\ \frac{\phi(2\alpha+s)}{\phi(2\alpha)} & \text{if } s > \max(0, 1-2\alpha), \end{cases} \\ s_\epsilon(s) &= 2\epsilon(1-\alpha) - \max(1-2\alpha, 0) - (\ell+1)^{-1} \max\big(s - \max(1-2\alpha, 0), 0\big), \\ s_\eta &= (2-2\eta)\alpha - \max(1-2\eta, 0). \end{aligned}$$

*Proof.* The proof is similar to that of Lemma 3.4. We denote the three terms on the left-hand side by $\mathrm{I}_0$, $\mathrm{I}_1$, and $\mathrm{I}_2$, respectively. It is easy to check that, for any $\eta \in [0,1]$, with the inequalities $c_0\|B\| \leq (2e)^{-1}$, $\|B\| \leq 1$ (cf. Assumption 1.1(i)), and $k - \ell \geq \frac{3}{4}k$,

$$\mathrm{I}_0 = \sum_{J_{\ell+1}\in\mathcal{J}_{[k-\ell,k],\ell+1}} \prod_{t=1}^{\ell+1} \eta_{j_t}^2 \|B^{\ell+1}\|^2 (k-\ell)^{-s} \leq (2^{2\alpha-2\eta} e^{-2\eta} c_0^{2-2\eta} k^{-2\alpha})^{\ell+1} \max(2^s, 1) k^{-s}.$$

For $\mathrm{I}_1$, by the condition $\|B\| \leq 1$ in Assumption 1.1 and Lemma 2.3 with $s = \ell_\epsilon$, we have

$$\begin{aligned} \mathrm{I}_1 &\leq \sum_{j_{\ell+1}=1}^{[\frac{k}{2}]} \sum_{J_\ell\in\mathcal{J}_{[j_{\ell+1}+1,k],\ell}} \prod_{t=1}^{\ell+1} \eta_{j_t}^2 \|\Pi_{J^c_{[j_{\ell+1}+1,k],\ell}}(B) B^{\ell_\epsilon}\|^2 j_{\ell+1}^{-s} \\ &\leq (\tfrac{\ell_\epsilon}{e})^{2\ell_\epsilon} c_0^{2(1-\epsilon)(\ell+1)} k^{2\ell_\epsilon\alpha} \sum_{j_{\ell+1}=1}^{[\frac{k}{2}]} \sum_{J_\ell\in\mathcal{J}_{[j_{\ell+1}+1,k],\ell}} \prod_{t=1}^{\ell+1} j_t^{-2\alpha} (k - j_{\ell+1} - \ell)^{-2\ell_\epsilon} j_{\ell+1}^{-s}. \end{aligned}$$

Now by the estimates (3.3) and (2.2), and Lemma 2.4(i),

$$\begin{aligned}
&\sum_{j_{\ell+1}=1}^{[\frac{k}{2}]} \sum_{J_\ell \in \mathcal{J}_{[j_{\ell+1}+1,k],\ell}} \prod_{t=1}^{\ell+1} j_t^{-2\alpha}(k-j_{\ell+1}-\ell)^{-2\ell_\epsilon} j_{\ell+1}^{-s} \\
\le& (\tfrac{k}{4})^{-2\ell_\epsilon}\Big(\sum_{j_{\ell+1}=1}^{[\frac{k}{2}]} j_{\ell+1}^{-(2\alpha+s)}\Big)\Big(\sum_{J_\ell \in \mathcal{J}_{[1,k],\ell}} \prod_{t=1}^{\ell} j_t^{-2\alpha}\Big) \\
\le& (\tfrac{k}{4})^{-2\ell_\epsilon}\Big(\sum_{j_{\ell+1}=1}^{[\frac{k}{2}]} j_{\ell+1}^{-(2\alpha+s)}\Big)\big(\phi(2\alpha)k^{\max(1-2\alpha,0)}\big)^{\ell}.
\end{aligned}$$

Direct computation with (2.2) gives

$$\sum_{j_{\ell+1}=1}^{[\frac{k}{2}]} j_{\ell+1}^{-(2\alpha+s)} \le \begin{cases} 2^s\phi(2\alpha)(\frac{k}{2})^{\max(1-2\alpha,0)}k^{-s}, & s \le 0, \\ \frac{\phi(2\alpha+s)}{\phi(2\alpha)}(\phi(2\alpha)k^{\max(1-2\alpha,0)})k^{s-\max(1-2\alpha,0)}k^{-s}, & s > \max(0,1-2\alpha). \end{cases}$$

These two estimates together give (3.7). Similarly, Lemma 2.3 with $s = \ell_\eta$ yields

$$\mathrm{I}_2 \le (\tfrac{\ell_\eta}{e})^{2\ell_\eta} c_0^{2(1-\eta)(\ell+1)} k^{2\ell_\eta\alpha} \sum_{j_{\ell+1}=[\frac{k}{2}]+1}^{k-\ell-1} \sum_{J_\ell \in \mathcal{J}_{[j_{\ell+1}+1,k],\ell}} \prod_{t=1}^{\ell+1} j_t^{-2\alpha}(k-j_{\ell+1}-\ell)^{-2\ell_\eta} j_{\ell+1}^{-s}.$$

Note that for $j_{\ell+1} = [\frac{k}{2}]+1, \dots, k-\ell-1$, $j_{\ell+1}^{-s} \le \max(1,2^s)k^{-s}$, and thus Lemma 2.4(ii) and the estimates (2.2) and (3.4) give

$$\begin{aligned}
&\sum_{j_{\ell+1}=[\frac{k}{2}]+1}^{k-\ell-1} \sum_{J_\ell \in \mathcal{J}_{[j_{\ell+1}+1,k],\ell}} \prod_{t=1}^{\ell+1} j_t^{-2\alpha}(k-j_{\ell+1}-\ell)^{-2\ell_\eta} j_{\ell+1}^{-s} \\
&\le \max(1,2^s)(\tfrac{k}{2})^{-2(\ell+1)\alpha}k^{-s} \sum_{j_{\ell+1}=[\frac{k}{2}]+1}^{k-\ell-1} \Big(\sum_{J_\ell \in \mathcal{J}_{[j_{\ell+1}+1,k],\ell}} 1\Big)(k-j_{\ell+1}-\ell)^{-2\ell_\eta} \\
&\le \max(1,2^s)\tfrac{(\ell+1)^\ell}{\ell!}(\tfrac{k}{2})^{-2(\ell+1)\alpha}k^{-s} \sum_{j_{\ell+1}=[\frac{k}{2}]+1}^{k-\ell-1} (k-j_{\ell+1}-\ell)^{-2\ell_\eta+\ell} \\
&\le \max(1,2^s)\tfrac{(\ell+1)^\ell\phi(2\ell_\eta-\ell)}{\ell!}(\tfrac{k}{2})^{-2(\ell+1)\alpha}k^{-s}k^{\max((1-2\eta)(\ell+1),0)},
\end{aligned}$$

where the last line follows from (2.2) and the identity $-2\ell_\eta+\ell+1 = (1-2\eta)(\ell+1)$. Combining the preceding estimates yields the bound (3.8) and completes the proof of the lemma. ∎

Now, we can prove the order-optimal convergence rate of SGD in Theorem 1.2.

*Proof of Theorem 1.2.* Let $r_k = \mathbb{E}[\|e_k^\delta\|^2]$. We prove that for any $\epsilon \in (\frac{1}{2}, 1]$, there exist $c_*$ and $c_{**}$ such that

$$r_k \le c_* k^{-\beta} + c_{**}\bar{\delta}^2 k^{\gamma}, \tag{3.9}$$

with $\beta = \min(2\nu, 1 + (2\epsilon - 1)(\ell + 1))(1 - \alpha)$ and $\gamma = 1 - \alpha$. Then the desired assertion holds by choosing $\epsilon \in (\frac{1}{2}, 1)$ and $\ell \in \mathbb{N}$ such that $(2\epsilon - 1)(\ell + 1) \geq 2\nu - 1$. The proof proceeds by mathematical induction. We treat the cases (i) $\alpha \in [0, \frac{1}{2}) \cup (\frac{1}{2}, 1)$ and (ii) $\alpha = \frac{1}{2}$ separately. First we consider case (i). If $k \leq 4\ell$, the estimate (3.9) holds for any sufficiently large $c_*$ and $c_{**}$. Assume that it holds up to some $k \geq 4\ell$, and we prove it for $k + 1$. It follows from Theorem 2.5 and Propositions 3.1 and 3.5 that

$$\begin{aligned} r_{k+1} \leq & c_{\nu,\ell,\alpha,n} c_0^{-2\nu} k^{-2\nu(1-\alpha)} \|w\|^2 + c_{\ell,\alpha,n,c_0} \bar{\delta}^2 k^{1-\alpha} \\ & + (2n)^{\ell+1} \sum_{J_{\ell+1} \in \mathcal{J}_{[1,k],\ell+1}} \prod_{t=1}^{\ell+1} \eta_{j_t}^2 \|\Pi_{J^c_{[j_{\ell+1}+1,k],\ell}}(B) B^{\ell+1}\|^2 r_{j_{\ell+1}}. \end{aligned}$$

Applying the induction hypothesis $r_j \leq c_* j^{-\beta} + c_{**} \bar{\delta}^2 j^{\gamma}$, $j = 1, 2, \ldots, k$, to the recursion gives

$$r_{k+1} \leq c_{\nu,\ell,\alpha,n} c_0^{-2\nu} k^{-2\nu(1-\alpha)} \|w\|^2 + c_{\ell,\alpha,n,c_0} \bar{\delta}^2 k^{1-\alpha} + (2n)^{\ell+1} c_* \mathrm{I} + (2n)^{\ell+1} c_{**} \bar{\delta}^2 \mathrm{II},$$

with

$$\begin{aligned} \mathrm{I} &:= \sum_{J_{\ell+1} \in \mathcal{J}_{[1,k],\ell+1}} \prod_{t=1}^{\ell+1} \eta_{j_t}^2 \|\Pi_{J^c_{[j_{\ell+1}+1,k],\ell}}(B) B^{\ell+1}\|^2 j_{\ell+1}^{-\beta}, \\ \mathrm{II} &:= \sum_{J_{\ell+1} \in \mathcal{J}_{[1,k],\ell+1}} \prod_{t=1}^{\ell+1} \eta_{j_t}^2 \|\Pi_{J^c_{[j_{\ell+1}+1,k],\ell}}(B) B^{\ell+1}\|^2 j_{\ell+1}^{\gamma}. \end{aligned}$$

Using (2.1), we split each of I and II into three terms over the index $j_{\ell+1}$, one for $j_{\ell+1} = k - \ell$, one from 1 to $[\frac{k}{2}]$, and one from $[\frac{k}{2}] + 1$ to $k - \ell - 1$, respectively. Now by Lemma 3.7,

$$\begin{aligned} r_{k+1} \leq & c_{\nu,\ell,\alpha,n} c_0^{-2\nu} k^{-2\nu(1-\alpha)} \|w\|^2 + c_{\ell,\alpha,n,c_0} \bar{\delta}^2 k^{1-\alpha} \\ & + c_* \xi_1(k) (k+1)^{-\beta} + c_{**} \bar{\delta}^2 \xi_2(k) (k+1)^{\gamma}, \end{aligned} \tag{3.10}$$

where the functions $\xi_1$ and $\xi_2$ are given by (for any $\epsilon, \eta \in [\frac{1}{2}, 1]$, which implies $\frac{1}{2} \leq \ell_\eta$)

$$\begin{aligned} \xi_1(k) =& \tfrac{2^\beta \phi(2\alpha+\beta)}{\phi(2\alpha)} (2^{1+2\epsilon} \ell_\epsilon^{2\epsilon} \phi(2\alpha) n c_0^{2-2\epsilon} k^{-s_\epsilon(\beta)})^{\ell+1} + \tfrac{2^{2\beta} c_\eta}{(\ell+1)!} (2^{2\alpha+1} e^{-2\eta} (\ell+1) \ell_\eta^{2\eta} n c_0^{2-2\eta} k^{-s_\eta})^{\ell+1}, \\ \xi_2(k) =& (2^{1+2\epsilon} \ell_\epsilon^{2\epsilon} \phi(2\alpha) n c_0^{2-2\epsilon} k^{-s_\epsilon(-\gamma)})^{\ell+1} + \tfrac{c_\eta}{(\ell+1)!} \big(2^{2\alpha+1} e^{-2\eta} (\ell+1) \ell_\eta^{2\eta} n c_0^{2-2\eta} k^{-s_\eta}\big)^{\ell+1}, \end{aligned}$$

with the constants $c_\eta = 1 + \phi(2\ell_\eta - \ell)$, $\ell_\epsilon$, $\ell_\eta$, $s_\epsilon(\cdot)$, and $s_\eta$ defined in Lemma 3.7. By choosing $1 \geq \epsilon = \eta > \frac{1}{2}$ and $\ell$ such that $(2\epsilon - 1)(\ell + 1) \geq 2\nu - 1$, we have $s_\epsilon(\beta), s_\epsilon(-\gamma), s_\eta \geq 0$, and

$$\begin{aligned} \xi_1(k) \leq c_1 :=& \tfrac{2^\beta \phi(2\alpha+\beta)}{\phi(2\alpha)} (2^{1+2\epsilon} \ell_\epsilon^{2\epsilon} \phi(2\alpha) n c_0^{2-2\epsilon})^{\ell+1} + \tfrac{2^{2\beta} c_\eta}{(\ell+1)!} (2^{2\alpha+1} e^{-2\eta} (\ell+1) \ell_\eta^{2\eta} n c_0^{2-2\eta})^{\ell+1}, \\ \xi_2(k) \leq c_2 :=& (2^{1+2\epsilon} \ell_\epsilon^{2\epsilon} \phi(2\alpha) n c_0^{2-2\epsilon})^{\ell+1} + \tfrac{c_\eta}{(\ell+1)!} \big(2^{2\alpha+1} e^{-2\eta} (\ell+1) \ell_\eta^{2\eta} n c_0^{2-2\eta}\big)^{\ell+1}. \end{aligned}$$

For small $c_0$, $c_1, c_2 \leq \frac{1}{2}$ hold, and then (3.9) follows by setting $c_* = 2^{2\nu+1} c_0^{-2\nu} c_{\nu,\ell,\alpha,n} \|w\|^2$ and $c_{**} = 2 c_{\ell,\alpha,n,c_0}$. This proves the theorem for case (i). In case (ii), repeating the preceding

argument, by choosing $1 \geq \epsilon = \eta > \frac{1}{2}$ and $\ell \geq 1$ such that $(2\epsilon - 1)(\ell + 1) \geq 2\nu - 1$, gives

$$k^{-s_\epsilon(\beta)} = k^{-\epsilon + \max(0,0) + (\ell+1)^{-1}\max(\beta - \max(0,0),0)} \leq k^{-\epsilon + (\ell+1)^{-1}\beta} \ln k \leq k^{-\frac{1}{4}} \ln k \leq 4,$$
$$k^{-s_\epsilon(-\gamma)} = k^{-\epsilon + \max(0,0) + (\ell+1)^{-1}\max(-\gamma - \max(0,0),0)} \leq k^{-\epsilon} \ln k \leq \epsilon^{-1} \quad \text{and} \quad k^{-s_\eta} \leq 1.$$

Then repeating the preceding argument shows that the assertion holds when

$$\xi_1(k) \leq c_1 := \tfrac{2^\beta \phi(2\alpha+\beta)}{\phi(2\alpha)} (2^{3+2\epsilon} \ell_\epsilon^{2\epsilon} \phi(2\alpha) n c_0^{2-2\epsilon})^{\ell+1} + \tfrac{2^{2\beta} c_\eta}{(\ell+1)!} (2^{2\alpha+1} e^{-2\eta} (\ell+1) \ell_\eta^{2\eta} n c_0^{2-2\eta})^{\ell+1} \leq \tfrac{1}{2},$$
$$\xi_2(k) \leq c_2 := (2^{1+2\epsilon} \epsilon^{-1} \ell_\epsilon^{2\epsilon} \phi(2\alpha) n c_0^{2-2\epsilon})^{\ell+1} + \tfrac{c_\eta}{(\ell+1)!} \big(2^{2\alpha+1} e^{-2\eta} (\ell+1) \ell_\eta^{2\eta} n c_0^{2-2\eta}\big)^{\ell+1} \leq \tfrac{1}{2},$$

which can be satisfied with sufficiently small $c_0$. This completes the proof of the theorem. ∎

*Remark* 3.8. In practice, it is desirable to take small $\alpha$. With the choice $\alpha = 0$ and $\epsilon = \eta \in (\frac{1}{2}, 1)$, the proof requires that the initial stepsize $c_0$ satisfy

$$2^\beta \phi(\beta) (2^{1+2\epsilon} \ell_\epsilon^{2\epsilon} n c_0^{2-2\epsilon})^{\ell+1} + \tfrac{2^{2\beta} c_\eta}{(\ell+1)!} (2e^{-2\eta} (\ell+1) \ell_\eta^{2\eta} n c_0^{2-2\eta})^{\ell+1} \leq \tfrac{1}{2},$$
$$(2^{1+2\epsilon} \ell_\epsilon^{2\epsilon} n c_0^{2-2\epsilon})^{\ell+1} + \tfrac{c_\eta}{(\ell+1)!} \big(2e^{-2\eta} (\ell+1) \ell_\eta^{2\eta} n c_0^{2-2\eta}\big)^{\ell+1} \leq \tfrac{1}{2}.$$

These two conditions are fulfilled provided that $n c_0^{2-2\epsilon}$ and $n c_0^{2-2\eta}$ are small constants. In particular, with $\epsilon = \eta$ close to $\frac{1}{2}$, the conditions essentially amount to $c_0 = O(n^{-1})$, agreeing with the condition in Remark 3.6. Under this condition, by choosing an a priori stopping index $k_*(\delta) = O((\|w\|\delta^{-1})^{\frac{2}{1+2\nu}})$, we obtain the following bound: $\mathbb{E}[\|e^\delta_{k_*(\delta)}\|^2]^{\frac{1}{2}} \leq c\delta^{\frac{2\nu}{1+2\nu}}$. This result is essentially identical to that for the Landweber method [7, Chapter 6], and higher than the existing convergence rate $O(\delta^{\frac{\min(2\nu,1)}{\min(2\nu,1)+1}})$ [18] for SGD. In particular, it proves that SGD with small initial stepsizes is actually order-optimal.

*Remark* 3.9. One may slightly refine Theorem 1.2. Indeed, for any $\alpha \in (0,1)$, let

$$H_1(k) := 2^3 \max(\nu, \ell+1)^2 n \phi(2\alpha) k^{-(1-\alpha)+\max(1-2\alpha,0)},$$
$$H_2(k) := 2^{2\alpha-1} (\ell+2)^2 n c_0 k^{-\alpha} \ln k.$$

Note that the following inequalities hold: $h_0(k) \leq H_1(k)$, $h_1(k) \leq H_1(k)$, and $h_2(k) \leq H_2(k)$. Then there exists some $k_0$, dependent of $\alpha$, $n$, $\ell$, $\nu$, and $c_0$, such that $H_1(k), H_2(k) \leq \frac{1}{2}$ for any $k \geq k_0$. The claim (3.9) shows that the assertion holds for any $k \leq k_0$ with sufficiently large $c_*$ and $c_{**}$. Then we refine the estimate by mathematical induction. Assume that the assertion holds up to some $k \geq k_0$, and prove it for $k+1$. Since $k \geq k_0$, $h_i(k) \leq \frac{1}{2}$, $i = 1, 2, 3$, it follows from the estimate (3.10), with $\beta = \min(2\nu, 1 + (2\epsilon - 1)(\ell+1))(1-\alpha)$, $\gamma = 1 - \alpha$, $\epsilon = 1$, and $\eta = \frac{1}{2}$, and Lemma 3.7 that

$$r_{k+1} \leq c_{\nu,\ell,\alpha,n} c_0^{-2\nu} k^{-2\nu(1-\alpha)} \|w\|^2 + c_{\ell,\alpha,n,c_0} \bar{\delta}^2 k^{1-\alpha}$$
$$+ c_* \xi_1(k) (k+1)^{-\beta} + c_{**} \bar{\delta}^2 \xi_2(k) (k+1)^{1-\alpha},$$

with the functions $\xi_1(k)$ and $\xi_2(k)$ given by

$$\begin{aligned}\xi_1(k) =&\tfrac{2^\beta\phi(2\alpha+\beta)}{\phi(2\alpha)}\big(2^3(\ell+1)^2\phi(2\alpha)nk^{-2(1-\alpha)+\max(1-2\alpha,0)+(\ell+1)^{-1}(\beta-\max(1-2\alpha,0))}\big)^{\ell+1}\\ &+\tfrac{3\cdot 2^{2\beta}}{(\ell+1)!}\big(2^{2\alpha-1}(\ell+1)^2nc_0k^{-\alpha}\ln k\big)^{\ell+1},\\ \xi_2(k) =&\big(2^3(\ell+1)^2\phi(2\alpha)nk^{-2(1-\alpha)+\max(1-2\alpha,0)}\big)^{\ell+1}+\tfrac{3}{(\ell+1)!}\big(2^{2\alpha-1}(\ell+1)^2nc_0k^{-\alpha}\ln k\big)^{\ell+1}.\end{aligned}$$

Since $(\ell+1)^{-1}(\beta-\max(1-2\alpha,0))\le 1-\alpha$, the terms on the right-hand side can be bounded by either $H_1(k)$ or $H_2(k)$, and thus for $k\ge k_0$, we have $H_1(k),H_2(k)\le\frac12$, and consequently

$$\xi_1(k)\le 2^{-(\ell+1)}\big(\tfrac{2^\beta\phi(2\alpha+\beta)}{\phi(2\alpha)}+\tfrac{3\cdot 2^{2\beta}}{(\ell+1)!}\big)\quad\text{and}\quad\xi_2(k)\le 2^{-(\ell+1)}\big(1+\tfrac{3}{(\ell+1)!}\big).$$

By choosing suitable $\ell\ge 2\nu-2$ (dependent of $\nu$), we can ensure $\xi_1(k),\xi_2(k)\le\frac12$, and then taking the constants $c_*$ and $c_{**}$ as before yields the desired assertion.

**3.4. Error analysis for $\alpha=0$.** In this section, we revisit the case $\alpha=0$ separately and derive an error bound directly with more explicit constants.

Lemma 3.10. *Let Assumption 1.1*(i) *and* (iii) *hold with $\alpha=0$. Further, suppose that the following condition holds:*

$$2(1+\phi(2\epsilon))nc_0^{2-2\epsilon}\le 1\quad\text{for some }\epsilon\in(\tfrac12,1).\tag{3.11}$$

*Then for any $s\ge 0$, there holds*

$$\mathbb{E}[\|(I-c_0B)^sB^{-\frac12}(Be_k^\delta-\bar A^t\bar\xi)\|^2]\le 2\|(I-c_0B)^{\frac{k-1}{2}+s}B^{-\frac12}(Be_1-\bar A^t\bar\xi)\|^2.$$

*Proof.* We prove the assertion by mathematical induction. When $k=1$, the inequality holds trivially true for any $s\ge 0$. Now we assume that it holds up to some $k-1\ge 1$ and prove it for $k$. With the condition $\alpha=0$, $\eta_j=c_0$ and $\Pi_{J^c_{[j,j']},0}(B)=(I-c_0B)^{j'-j+1}$ for any $j'\ge j\ge 1$. By the definitions of $H_j$ and $N_j^\delta$, we can rewrite $H_j$ as $H_j=\bar A^t\bar\xi+A^tN_j^\delta$. Consequently, we derive from (2.4) that for any $s\ge 0$

$$\begin{aligned}&(I-c_0B)^sB^{-\frac12}(Be_k^\delta-\bar A^t\bar\xi)\\ =&(I-c_0B)^sB^{-\frac12}\Big((I-c_0B)^{k-1}Be_1^\delta-\bar A^t\bar\xi+c_0\sum_{j=1}^{k-1}(I-c_0B)^{k-j-1}B\bar A^t\bar\xi\\ &\qquad+c_0\sum_{j=1}^{k-1}(I-c_0B)^{k-j-1}BA^tN_j^\delta\Big)\\ =&(I-c_0B)^{k-1+s}B^{-\frac12}(Be_1^\delta-\bar A^t\bar\xi)+c_0\sum_{j=1}^{k-1}(I-c_0B)^{k-j-1+s}B^{\frac12}A^tN_j^\delta,\end{aligned}$$

in view of the identity $c_0\sum_{j=1}^{k-1}(I-c_0B)^{k-j-1}B = I-(I-c_0B)^{k-1}$. By the recursion (2.4) and (2.10)–(2.11), we have the following bias-variance decomposition:

$$\begin{aligned}&\mathbb{E}[\|(I-c_0B)^sB^{-\frac{1}{2}}(Be_k^\delta-\bar{A}^t\bar{\xi})\|^2]\\ =&\|(I-c_0B)^{k-1+s}B^{-\frac{1}{2}}(Be_1^\delta-\bar{A}^t\bar{\xi})\|^2+c_0^2\sum_{j=1}^{k-1}\mathbb{E}[\|(I-c_0B)^{k-j-1+s}B^{\frac{1}{2}}A^tN_j^\delta\|^2].\end{aligned}$$

Next we denote the summation by $\mathrm{I}(s)$. Then the argument for $N_j^\delta$ in the proof of Theorem 2.5 and the condition $\|B\|\le 1$ imply that for any $\epsilon\in(\frac{1}{2},1)$,

$$\begin{aligned}\mathrm{I}(s)\le& nc_0^2\sum_{j=1}^{k-1}\mathbb{E}[\|(I-c_0B)^{k-j-1+s}B^{\frac{1}{2}}(Be_j^\delta-\bar{A}^t\bar{\xi})\|^2]\\ \le& nc_0^2\|(I-c_0B)^{-1}\|\sum_{j=1}^{k-1}\|(I-c_0B)^{\frac{k-j-1}{2}}B^\epsilon\|^2\mathbb{E}[\|(I-c_0B)^{\frac{k-j}{2}+s}B^{-\frac{1}{2}}(Be_j^\delta-\bar{A}^t\bar{\xi})\|^2].\end{aligned}$$

With the identity $\|(I-c_0B)^{-1}\| = (1-c_0\|B\|)^{-1}$ and the induction hypothesis

$$\begin{aligned}&\mathbb{E}[\|(I-c_0B)^{\frac{k-j}{2}+s}B^{-\frac{1}{2}}(Be_j^\delta-\bar{A}^t\bar{\xi})\|^2]\\ \le& 2\|(I-c_0B)^{\frac{k-1}{2}+s}B^{-\frac{1}{2}}(Be_1-\bar{A}^t\bar{\xi})\|^2,\quad j=1,\dots,k-1,\end{aligned}$$

we deduce

$$\mathrm{I}(s)\le\frac{2nc_0^2}{1-c_0\|B\|}\|(I-c_0B)^{\frac{k-1}{2}+s}B^{-\frac{1}{2}}(Be_1-\bar{A}^t\bar{\xi})\|^2\sum_{j=1}^{k-1}\|(I-c_0B)^{\frac{k-j-1}{2}}B^\epsilon\|^2.$$

By Lemma 2.3 and the estimate (2.2),

$$\begin{aligned}&\sum_{j=1}^{k-1}\|(I-c_0B)^{\frac{k-j-1}{2}}B^\epsilon\|^2 = \|B^\epsilon\|^2+\sum_{j=1}^{k-2}\|(I-c_0B)^{k-j-1}B^{2\epsilon}\|\\ \le&(\tfrac{2\epsilon}{ec_0})^{2\epsilon}\sum_{j=1}^{k-1}(k-j-1)^{-2\epsilon}\le(\tfrac{2\epsilon}{ec_0})^{2\epsilon}(1+\phi(2\epsilon)).\end{aligned}$$

This, the assumption $c_0\|B\|\le(2e)^{-1}$ and the condition (3.11) imply

$$\mathrm{I}(s)\le\|(I-c_0B)^{\frac{k-1}{2}+s}B^{-\frac{1}{2}}(Be_1-\bar{A}^t\bar{\xi})\|^2.$$

This completes the induction step of the proof and thus also the proof of the lemma. ∎

Finally, we can state a refined error estimate for the case $\alpha=0$.

Theorem 3.11. *Let Assumption 1.1 hold with $\alpha=0$. Under condition (3.11), there holds*

$$\mathbb{E}[\|e_k^\delta\|^2]\le c_*^0k^{-2\nu}\|w\|^2+c_{**}^0\bar{\delta}^2k,$$

*with constants $c_*^0 = 2(\frac{2\nu}{ec_0})^{2\nu}+6nc_0(\frac{2(2\nu+1)}{ec_0})^{2\nu+1}$ and $c_{**}^0 = 3+6nc_0$.*

*Proof.* Under condition (3.11), the proof of Theorem 2.5 and Lemma 3.10 imply

$$\begin{aligned}\mathbb{E}[\|e_k^\delta\|^2] \leq & 2\|(I-c_0B)^{k-1}e_1\|^2 + 2c_0^2\bar{\delta}^2\Big\|\sum_{j=1}^{k-1}(I-c_0B)^{k-j-1}B^{\frac{1}{2}}\Big\|^2 \\ & + nc_0^2\sum_{j=1}^{k-1}\mathbb{E}[\|(I-c_0B)^{k-j-1}(Be_j^\delta - \bar{A}^t\bar{\xi})\|^2].\end{aligned}$$

Below we denote the last term by II. Note that

$$\begin{aligned}&\mathbb{E}[\|(I-c_0B)^{k-j-1}(Be_j^\delta - \bar{A}^t\bar{\xi})\|^2] \\ \leq & (1-c_0\|B\|)^{-1}\|(I-c_0B)^{\frac{k-j-1}{2}}B^{\frac{1}{2}}\|^2\mathbb{E}[\|(I-c_0B)^{\frac{k-j}{2}}B^{-\frac{1}{2}}(Be_j^\delta - \bar{A}^t\bar{\xi})\|^2] \\ \leq & 2(1-c_0\|B\|)^{-1}\|(I-c_0B)^{k-j-1}B\|\|(I-c_0B)^{\frac{k-1}{2}}B^{-\frac{1}{2}}(Be_1 - \bar{A}^t\bar{\xi})\|^2.\end{aligned}$$

By Lemma 2.3, the assumption $c_0\|B\| \leq (2e)^{-1}$, and the estimates (2.2) and (3.5), we have

$$\sum_{j=1}^{k-1}\|(I-c_0B)^{k-j-1}B\| \leq (ec_0)^{-1}\sum_{j=1}^{k-1}(k-j-1)^{-1} \leq 3(ec_0)^{-1}\max(\ln k, 1) \leq 3(ec_0)^{-1}k.$$

Meanwhile, by the Cauchy–Schwarz inequality, we have

$$\|(I-c_0B)^{\frac{k-1}{2}}(B^{\frac{1}{2}}e_1 - B^{-\frac{1}{2}}\bar{A}^t\bar{\xi})\|^2 \leq 2\big(\|e_1^t(I-c_0B)^{k-1}Be_1\| + \bar{\delta}^2\|(I-c_0B)^{\frac{k-1}{2}}\|^2\big).$$

Combining the preceding estimates with Assumption 1.1(ii) and Lemma 2.3 leads to

$$\text{II} \leq 6nc_0k\big(\|(I-c_0B)^{k-1}B^{2\nu+1}\|\|w\|^2 + \bar{\delta}^2\big) \leq 6nc_0\big((\tfrac{2(2\nu+1)}{ec_0})^{2\nu+1}k^{-2\nu}\|w\|^2 + \bar{\delta}^2k\big).$$

Similarly, there hold

$$\begin{aligned}\|(I-c_0B)^{k-1}e_1\|^2 &\leq (\tfrac{2\nu}{ec_0})^{2\nu}k^{-2\nu}\|w\|^2, \\ \Big\|\sum_{j=1}^{k-1}(I-c_0B)^{k-j-1}B^{\frac{1}{2}}\Big\|^2 &\leq \Big(\sum_{j=1}^{k-1}\|(I-c_0B)^{k-j-1}B^{\frac{1}{2}}\|\Big)^2 \leq \tfrac{3}{2}c_0^{-2}k.\end{aligned}$$

Combining the preceding estimates completes the proof of the theorem. ∎

**4. Numerical experiments and discussions.** In this section, we provide numerical experiments to complement the analysis. To this end, we employ three examples, denoted by `s-phillips` (mildly ill-posed), `s-gravity` (severely ill-posed), and `s-shaw` (severely ill-posed), adapted from `phillips`, `gravity`, and `shaw` in the public MATLAB package Regutools [10] (available at http://people.compute.dtu.dk/pcha/Regutools/, last accessed on August 20, 2020). These examples are Fredholm/Volterra integral equations of the first kind, discretized by means of either Galerkin approximation with piecewise constant basis functions or quadrature rules, and all discretized into a linear system of size $n = m = 1000$. To explicitly control the regularity index $\nu$ in the source condition (1.5), we generate the true solution $x^\dagger$ by

$$x^\dagger = \frac{(A^tA)^\nu x_e}{\|(A^tA)^\nu x_e\|_{\ell^\infty}},$$

where $x_e$ is the exact solution provided by the package, and $\|\cdot\|_{\ell^\infty}$ denotes the maximum norm of vectors. In the test, the exponent $\nu$ is taken from the set $\{0, 1, 2, 4\}$. Note that the exponent $\nu$ in the source condition (1.5) is slightly larger than $\nu$ defined above, due to the inherent regularity of $x_e$ (which is less than one half for all examples). The exact data $y^\dagger$ is generated by $y^\dagger = Ax^\dagger$ and the noise data $y^\delta$ by

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

where $\xi_i$s are i.i.d. and follow the standard Gaussian distribution, and $\epsilon > 0$ represents the relative noise level (the exact noise level being $\delta = \|y^\delta - y^\dagger\|$). SGD is always initialized with $x_1 = 0$, and the maximum number of epochs is fixed at 9e5, where one epoch refers to $n$ SGD iterations. All statistical quantities presented below are computed from 100 independent runs. To verify the order optimality of SGD, we evaluate it against an order-optimal regularization method with infinite qualification, i.e., the Landweber method [7, Chapter 6], since it is the population version of SGD, converges steadily while enjoying order optimality, and thus serves as a good benchmark for performance comparison in terms of the convergence rate. (However, one may employ any other order-optimal method.) It is initialized with $x_1 = 0$, with a constant stepsize $\frac{1}{\|A\|^2}$, which can be much larger than that taken by SGD.

**Table 1**
*Comparison between SGD and LM for* `s-phillips`.

| Method | | SGD($\alpha = 0$) | | | SGD($\alpha = 0.1$) | | | LM | |
|---|---|---|---|---|---|---|---|---|---|
| $\nu$ | $\epsilon$ | $c_0$ | $e_{\mathrm{sgd}}$ | $k_{\mathrm{sgd}}$ | $c_0$ | $e_{\mathrm{sgd}}$ | $k_{\mathrm{sgd}}$ | $e_{\mathrm{lm}}$ | $k_{\mathrm{lm}}$ |
| 0 | 1e-3 | $4c/n$ | 1.66e-2 | 4691.28 | $c/30$ | 1.67e-2 | 2176.23 | 1.65e-2 | 5851 |
| | 5e-3 | $4c/n$ | 9.35e-2 | 782.10 | $c/30$ | 9.49e-2 | 336.33 | 9.28e-2 | 1036 |
| | 1e-2 | $4c/n$ | 1.29e-1 | 204.90 | $c/30$ | 1.32e-1 | 69.69 | 1.28e-1 | 249 |
| | 5e-2 | $4c/n$ | 5.42e-1 | 108.90 | $c/30$ | 5.57e-1 | 34.11 | 5.34e-1 | 136 |
| 1 | 1e-3 | $c/n$ | 3.48e-4 | 539.19 | $c/n$ | 2.88e-4 | 2089.62 | 2.28e-4 | 157 |
| | 5e-3 | $c/n$ | 3.69e-3 | 73.44 | $c/n$ | 3.32e-3 | 218.94 | 2.74e-3 | 20 |
| | 1e-2 | $c/n$ | 6.64e-3 | 57.81 | $c/n$ | 6.12e-3 | 166.47 | 5.12e-3 | 16 |
| | 5e-2 | $c/n$ | 3.52e-2 | 29.40 | $c/n$ | 3.31e-2 | 80.79 | 3.16e-2 | 8 |
| 2 | 1e-3 | $c/(30n)$ | 7.02e-5 | 2115.54 | $c/(20n)$ | 5.48e-5 | 5912.91 | 3.22e-5 | 19 |
| | 5e-3 | $c/(30n)$ | 4.47e-4 | 1197.48 | $c/(20n)$ | 4.13e-4 | 3201.63 | 3.76e-4 | 11 |
| | 1e-2 | $c/(30n)$ | 1.09e-3 | 938.70 | $c/(20n)$ | 1.04e-3 | 2441.85 | 9.82e-4 | 8 |
| | 5e-2 | $c/(30n)$ | 2.92e-2 | 636.51 | $c/(20n)$ | 2.90e-2 | 1597.56 | 1.57e-2 | 5 |
| 4 | 1e-3 | $c/(30n)$ | 9.77e-5 | 1966.38 | $c/(20n)$ | 6.91e-5 | 3291.18 | 1.30e-5 | 8 |
| | 5e-3 | $c/(30n)$ | 7.55e-4 | 879.51 | $c/(20n)$ | 6.97e-4 | 2263.89 | 3.83e-4 | 6 |
| | 1e-2 | $c/(30n)$ | 2.56e-3 | 785.94 | $c/(20n)$ | 2.50e-3 | 1996.83 | 1.42e-3 | 5 |
| | 5e-2 | $c/(30n)$ | 5.23e-2 | 596.73 | $c/(20n)$ | 5.21e-2 | 1489.29 | 2.49e-2 | 3 |

**4.1. Numerical results for general $A$.** The numerical results for the three examples are shown in Tables 1–3, where the notation $e_{\mathrm{sgd}} = \mathbb{E}[\|x^\delta_{k_{\mathrm{sgd}}} - x^\dagger\|^2]$ denotes the mean squared error achieved at the $k_{\mathrm{sgd}}$th iteration (counted in epochs) by SGD, and $e_{\mathrm{lm}} = \|x^\delta_{k_{\mathrm{lm}}} - x^\dagger\|^2$ and $k_{\mathrm{lm}}$ denote the squared $\ell^2$ error and the stopping index for the Landweber method. The stopping indices $k_{\mathrm{sgd}}$ and $k_{\mathrm{lm}}$ are taken such that the corresponding error is smallest for SGD and the Landweber method, respectively, along the iteration trajectory. The choice of the

stopping index is motivated by a lack of provably order-optimal a posteriori stopping rules for SGD. The initial stepsize $c_0$ is also indicated in the tables, in the form of a multiple of the constant $c = \frac{1}{\max_i(\|a_i\|^2)}$. In the experiments, we consider two decay rates for the stepsize schedule, i.e., $\alpha = 0$ and $\alpha = 0.1$.

**Table 2**
*Comparison between SGD and LM for* `s-gravity`.

| Method | | SGD($\alpha = 0$) | | | SGD($\alpha = 0.1$) | | | LM | |
|---|---|---|---|---|---|---|---|---|---|
| $\nu$ | $\epsilon$ | $c_0$ | $e_{\mathrm{sgd}}$ | $k_{\mathrm{sgd}}$ | $c_0$ | $e_{\mathrm{sgd}}$ | $k_{\mathrm{sgd}}$ | $e_{\mathrm{lm}}$ | $k_{\mathrm{lm}}$ |
| 0 | 1e-3 | $c/20$ | 9.37e-2 | 1000.50 | $c/10$ | 9.39e-2 | 1894.14 | 9.39e-2 | 27201 |
| | 5e-3 | $c/20$ | 3.29e-1 | 86.43 | $c/10$ | 3.29e-1 | 134.85 | 3.27e-1 | 2515 |
| | 1e-2 | $c/20$ | 5.81e-1 | 34.11 | $c/10$ | 5.80e-1 | 34.17 | 5.73e-1 | 793 |
| | 5e-2 | $c/20$ | 2.23e0 | 5.61 | $c/10$ | 2.22e0 | 6.03 | 2.07e0 | 149 |
| 1 | 1e-3 | $c/(30n)$ | 5.90e-4 | 5604.80 | $c/(10n)$ | 5.95e-4 | 8095.17 | 5.68e-4 | 99 |
| | 5e-3 | $c/(30n)$ | 5.13e-3 | 2069.25 | $c/(10n)$ | 5.14e-3 | 2707.74 | 5.02e-3 | 37 |
| | 1e-2 | $c/(30n)$ | 1.15e-2 | 1356.87 | $c/(10n)$ | 1.15e-2 | 1688.04 | 1.12e-2 | 24 |
| | 5e-2 | $c/(30n)$ | 6.48e-2 | 613.41 | $c/(10n)$ | 6.48e-2 | 709.08 | 6.19e-2 | 11 |
| 2 | 1e-3 | $c/(50n)$ | 1.32e-4 | 2441.85 | $c/(20n)$ | 1.28e-4 | 3983.34 | 6.82e-5 | 23 |
| | 5e-3 | $c/(50n)$ | 7.74e-4 | 1274.67 | $c/(20n)$ | 7.64e-4 | 1940.70 | 6.06e-4 | 12 |
| | 1e-2 | $c/(50n)$ | 1.92e-3 | 1047.03 | $c/(20n)$ | 1.91e-3 | 1580.49 | 1.47e-3 | 10 |
| | 5e-2 | $c/(50n)$ | 2.35e-2 | 708.72 | $c/(20n)$ | 2.34e-2 | 1013.31 | 1.61e-2 | 6 |
| 4 | 1e-3 | $c/(60n)$ | 1.03e-4 | 2212.26 | $c/(30n)$ | 8.38e-5 | 3982.77 | 1.30e-5 | 10 |
| | 5e-3 | $c/(60n)$ | 4.65e-4 | 1054.53 | $c/(30n)$ | 4.24e-4 | 2002.71 | 2.04e-4 | 7 |
| | 1e-2 | $c/(60n)$ | 1.29e-3 | 941.19 | $c/(30n)$ | 1.25e-3 | 1782.99 | 6.42e-4 | 6 |
| | 5e-2 | $c/(60n)$ | 2.25e-2 | 746.67 | $c/(30n)$ | 2.26e-2 | 1398.72 | 8.58e-3 | 3 |

First we comment on the SGD results. Clearly, for each fixed $\nu$, the mean squared error $e_{\mathrm{sgd}}$ (and also $e_{\mathrm{lm}}$) decreases to zero as the noise level $\epsilon$ tends to zero, but it takes more iterations to reach the optimal error, and the decay rate depends on the regularity index $\nu$ roughly as the theoretical prediction $O(\delta^{\frac{4\nu}{2\nu+1}})$. The larger the regularity index $\nu$, the faster the error decays, and the fewer iterations it needs in order to reach the optimal error. The results obtained by SGD with $\alpha = 0$ and $\alpha = 0.1$ are largely comparable with each other, but generally the former imposes a more stringent condition on the initial stepsize $c_0$ than the latter so as to achieve comparable accuracy. This is attributed to the fact that polynomially decaying stepsize schedules have a built-in variance reduction mechanism as the iteration proceeds. Nonetheless, at the low-regularity index (indicated by $\nu = 0$ in the table), the initial stepsize can be taken independent of $n$. Next we compare the results of SGD with the Landweber method. For all regularity indices, SGD, with an either constant or decaying stepsize schedule, can achieve an accuracy comparable with that of the Landweber method, provided that the initial stepsize $c_0$ for SGD is taken to be of order $O(n^{-1})$. Generally, the larger the index $\nu$ is, the smaller the value $c_0$ should be taken in the stepsize schedule in order to fully realize the benefit of smooth solutions. This observation agrees well with the observation in Remark 3.6. These observations hold for all three examples, which have different degree of ill-posedness. Thus they are fully in line with the convergence analysis in section 3.

In order to shed further light on the convergence behavior of SGD, we present numerical

**Table 3**
*Comparison between SGD and LM for* `s-shaw`.

| Method | | SGD($\alpha = 0$) | | | SGD($\alpha = 0.1$) | | | LM | |
|---|---|---|---|---|---|---|---|---|---|
| $\nu$ | $\epsilon$ | $c_0$ | $e_{\text{sgd}}$ | $k_{\text{sgd}}$ | $c_0$ | $e_{\text{sgd}}$ | $k_{\text{sgd}}$ | $e_{\text{lm}}$ | $k_{\text{lm}}$ |
| 0 | 1e-3 | $c$ | 2.81e-1 | 2704.92 | $2c$ | 2.81e-1 | 5853.54 | 2.81e-1 | 760983 |
| | 5e-3 | $c$ | 5.37e-1 | 67.14 | $2c$ | 5.33e-1 | 94.92 | 5.25e-1 | 18588 |
| | 1e-2 | $c$ | 7.08e-1 | 42.42 | $2c$ | 6.98e-1 | 60.18 | 6.67e-1 | 12385 |
| | 5e-2 | $c$ | 3.91e0 | 10.59 | $2c$ | 3.66e0 | 14.91 | 2.91e0 | 3392 |
| 1 | 1e-3 | $2c/n$ | 1.21e-4 | 275.70 | $4c/n$ | 1.26e-4 | 453.00 | 5.95e-5 | 144 |
| | 5e-3 | $2c/n$ | 1.45e-3 | 142.05 | $4c/n$ | 1.48e-3 | 202.50 | 1.26e-3 | 71 |
| | 1e-2 | $2c/n$ | 5.75e-3 | 113.01 | $4c/n$ | 5.62e-3 | 148.11 | 5.21e-3 | 54 |
| | 5e-2 | $2c/n$ | 1.51e-1 | 64.77 | $4c/n$ | 1.54e-1 | 97.02 | 1.47e-1 | 36 |
| 2 | 1e-3 | $2c/n$ | 1.53e-4 | 255.27 | $4c/n$ | 1.29e-4 | 746.46 | 6.36e-5 | 50 |
| | 5e-3 | $2c/n$ | 2.00e-3 | 84.60 | $4c/n$ | 1.73e-3 | 235.08 | 1.51e-3 | 37 |
| | 1e-2 | $2c/n$ | 6.43e-3 | 64.77 | $4c/n$ | 6.05e-3 | 172.32 | 5.71e-3 | 30 |
| | 5e-2 | $2c/n$ | 8.17e-2 | 11.88 | $4c/n$ | 8.00e-2 | 29.49 | 7.08e-2 | 5 |
| 4 | 1e-3 | $c/(30n)$ | 5.79e-5 | 1966.38 | $c/(10n)$ | 5.92e-5 | 2863.35 | 3.13e-5 | 9 |
| | 5e-3 | $c/(30n)$ | 6.00e-4 | 941.19 | $c/(10n)$ | 6.06e-4 | 1116.81 | 3.71e-4 | 5 |
| | 1e-2 | $c/(30n)$ | 1.99e-3 | 828.45 | $c/(10n)$ | 2.00e-3 | 1002.93 | 1.01e-3 | 4 |
| | 5e-2 | $c/(30n)$ | 3.61e-2 | 645.75 | $c/(10n)$ | 3.61e-2 | 746.67 | 6.45e-3 | 1 |

results with four different values of $c_0$ (i.e., $\min(c, nc^*)$, $10c^*$, $c^*$, and $\frac{c^*}{10}$, with $c^*$ from the tables) in Figures 4.1 and 4.2 for the examples with $\nu = 1$ and exact and noisy data, respectively. In the case of exact data, the mean squared error $e_{\text{sgd}}$ consists of only approximation and stochastic errors, and it decreases to zero as the iteration proceeds. With a large initial stepsize, the error $e_{\text{sgd}}$ decreases quickly during the initial iterations, but only at a slow rate $O(k^{-(1-\alpha)})$, whereas with a small $c_0$, the initial decay is much slower. The asymptotic decay rate matches the optimal decay $O(k^{-2\nu(1-\alpha)})$ only when $c_0$ decreases to $O(n^{-1})$, which otherwise exhibits only a slower decay $O(k^{-\min(2\nu,1)(1-\alpha)})$ and thus an undesirable saturation phenomenon. Note that for small $c_0$, the asymptotic decay $O(k^{-2\nu(1-\alpha)})$ kicks in only after a sufficient number of iterations, which agrees with the condition $h_0(k) \leq \frac{1}{2}$, etc., in the analysis. Further, there is an interesting transition layer for medium $c_0$ (but still of order $O(n^{-1})$), for which it first exhibits the desired asymptotic decay and then eventually shifts back to a slower decay rate. The presence of the wide transition region indicates that the optimal convergence can still be achieved for noisy data even if the employed $c_0$ is larger than the critical value suggested by the theoretical analysis in section 3. These observations hold for both constant and polynomially decaying stepsize schedules. These numerical results show that a small initial stepsize $c_0$ is necessary for overcoming the saturation phenomenon of SGD.

These empirical observations remain largely valid also for noisy data in Figure 4.2. It is observed that the asymptotic decay rate is higher for smaller initial stepsizes, but now only up to a certain iteration number, due to the presence of the propagation error, which increases monotonically as the iteration proceeds and eventually dominates the total error. This leads to the familiar semiconvergence behavior in the second and third columns of Figure 4.2. The proper balance between the decaying approximation error and the increasing propagation error determines the attainable accuracy. One clearly observes that the larger $c_0$ is, the faster the

asymptotic decay kicks in, but also the quicker the SGD iterate starts to diverge, which can greatly compromise the attainable accuracy along the trajectory, leading to the undesirable saturation phenomenon. When the initial stepsize $c_0$ becomes smaller, the attainable accuracy improves steadily. In particular, with a sufficiently small $c_0$, the attained error is optimal (but of course at the expense of a much increased computational complexity). This observation naturally leads to the important question of whether it is possible to design novel stepsize schedules (possibly not of polynomially decaying type) that enjoy both fast preasymptotic and asymptotic convergence behavior.

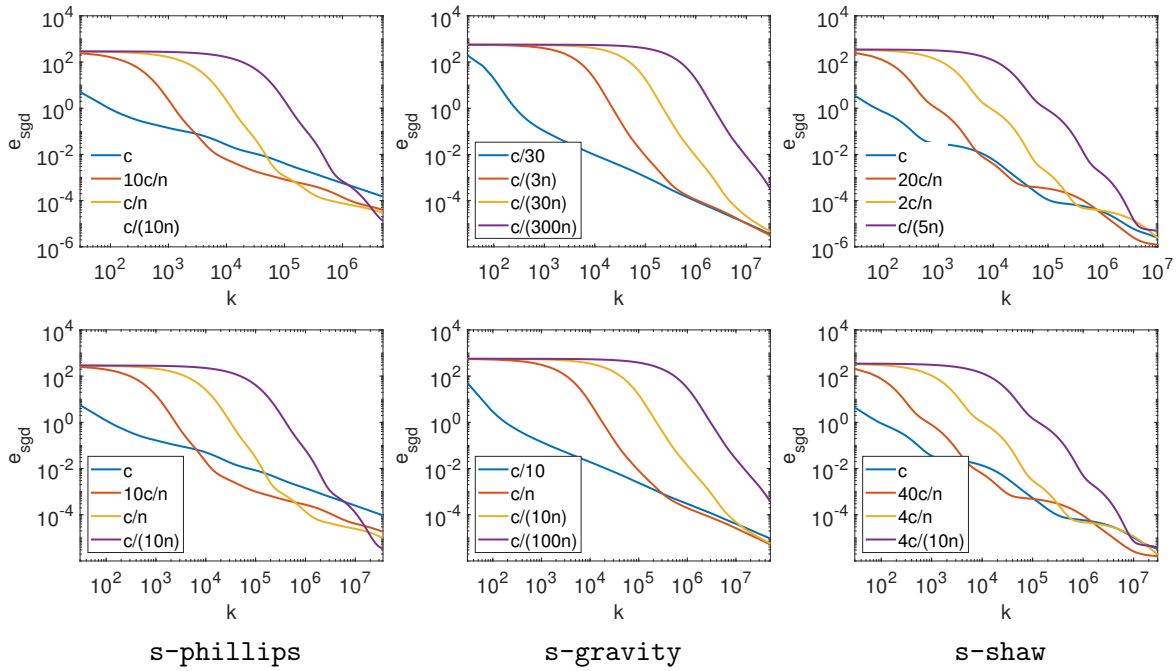


**Figure 4.1.** *The convergence trajectory of the SGD error with different initial stepsize $c_0$ for the examples with $\nu = 1$. The top and bottom rows are for $\alpha = 0$ and $\alpha = 0.1$, respectively.*

**4.2. On Assumption 1.1(iii).** The convergence analysis in section 3 requires Assumption 1.1(iii). This appears largely to be a limitation of the analysis technique. To illustrate this, we compare the results of the systems with a general matrix $A$ and of one that satisfies Assumption 1.1(iii). The latter can be constructed from the former as follows. Let $A = U\Sigma V^t$ be the singular value decomposition of $A$. Then we replace $A$ by $\tilde{A} = U^t A$ and $y^\delta$ by $\tilde{y}^\delta = U^t y^\delta$ so that $\tilde{A}$ satisfies Assumption 1.1(ii)–(iii). The numerical results for `s-phillips` are shown in Table 4. It is observed that the results obtained by SGD with $A$ and $\tilde{A}$ are largely comparable with each other for all the noise levels and smooth indices, especially when the amount of the data noise is not too small. Although not presented, the observations are identical for other examples, including multiplying the matrix $A$ by an arbitrary orthonormal matrix, as long as $c_0$ is sufficiently small. These observations are also confirmed by the corresponding trajectories: The trajectories of the mean squared error for the three examples with $\nu = 1$ for $A$ and $\tilde{A}$ nearly overlap when the data is not too small (as in most practical inverse

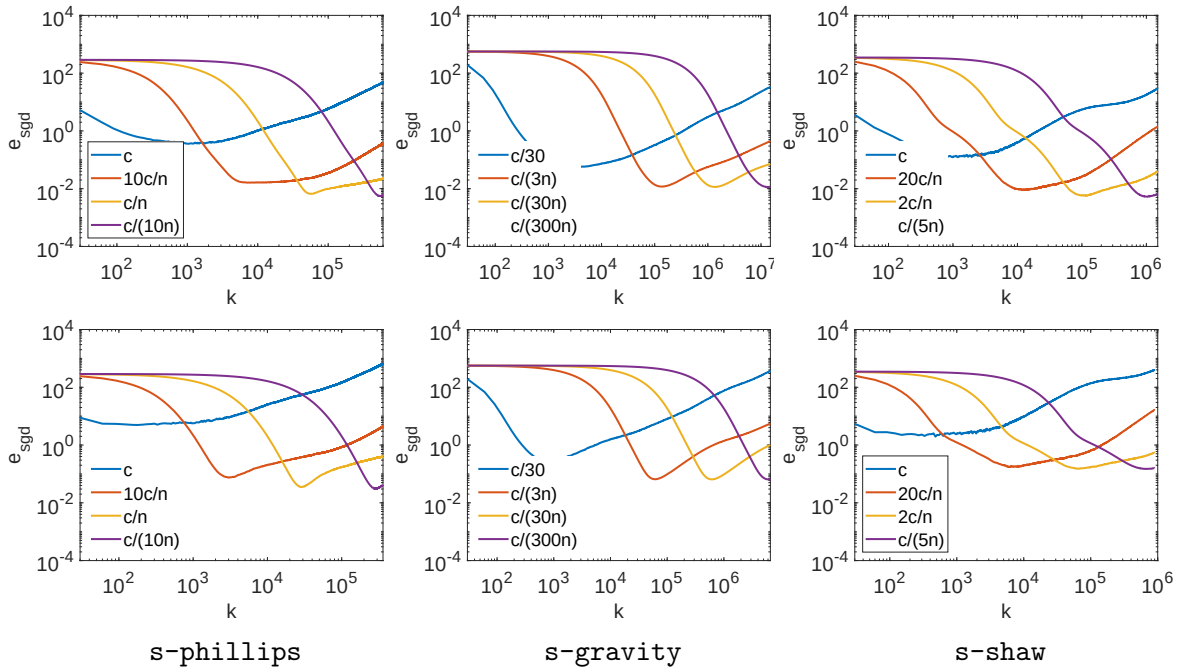


**Figure 4.2.** *The convergence trajectory of the SGD error (with $\alpha = 0$) with different initial stepsize $c_0$ for the examples with $\nu = 1$. The top and bottom rows are for $\epsilon$ =1e-2 and $\epsilon$=5e-2, respectively.*

**Table 4**

*Comparison between SGD with $\alpha = 0$ for `s-phillips` with $A$ and $\tilde{A}$.*

| Method | | | SGD with $A$ | | SGD with $\tilde{A}$ | |
|---|---|---|---|---|---|---|
| $\nu$ | $\epsilon$ | $c_0$ | $e$ | $k$ | $e$ | $k$ |
| 0 | 1e-3 | $4c/n$ | 1.66e-2 | 4691.28 | 1.65e-2 | 4738.4 |
| | 5e-3 | $4c/n$ | 9.35e-2 | 782.10 | 9.28e-2 | 835.35 |
| | 1e-2 | $4c/n$ | 1.29e-1 | 204.90 | 1.28e-1 | 198.75 |
| | 5e-2 | $4c/n$ | 5.42e-1 | 108.90 | 5.40e-1 | 111.85 |
| 1 | 1e-3 | $c/n$ | 3.48e-4 | 539.19 | 2.29e-4 | 507.55 |
| | 5e-3 | $c/n$ | 3.69e-3 | 73.44 | 2.87e-3 | 71.2 |
| | 1e-2 | $c/n$ | 6.64e-3 | 57.81 | 5.72e-3 | 57.75 |
| | 5e-2 | $c/n$ | 3.52e-2 | 29.40 | 3.84e-2 | 30.4 |
| 2 | 1e-3 | $c/(30n)$ | 7.02e-5 | 2115.54 | 3.49e-5 | 2021.6 |
| | 5e-3 | $c/(30n)$ | 4.47e-4 | 1197.48 | 3.66e-4 | 1186.10 |
| | 1e-2 | $c/(30n)$ | 1.09e-3 | 938.70 | 9.90e-4 | 934.75 |
| | 5e-2 | $c/(30n)$ | 2.92e-2 | 636.51 | 2.94e-2 | 639.60 |
| 4 | 1e-3 | $c/(30n)$ | 9.77e-5 | 1966.38 | 2.49e-5 | 1103.00 |
| | 5e-3 | $c/(30n)$ | 7.55e-4 | 879.51 | 6.34e-4 | 869.40 |
| | 1e-2 | $c/(30n)$ | 2.56e-3 | 785.94 | 2.43e-3 | 781.80 |
| | 5e-2 | $c/(30n)$ | 5.23e-2 | 596.73 | 5.24e-2 | 597.60 |

problems) (cf. Figure 4.3). For exact data, the trajectories overlap up to a certain point around 1e-4 (which depends on the value of $c_0$), and the value of leveling off is observed to

further decrease by choosing smaller $c_0$. One interesting open question is thus to establish the saturation-overcoming phenomenon without Assumption 1.1(iii), as the experimental results suggest.

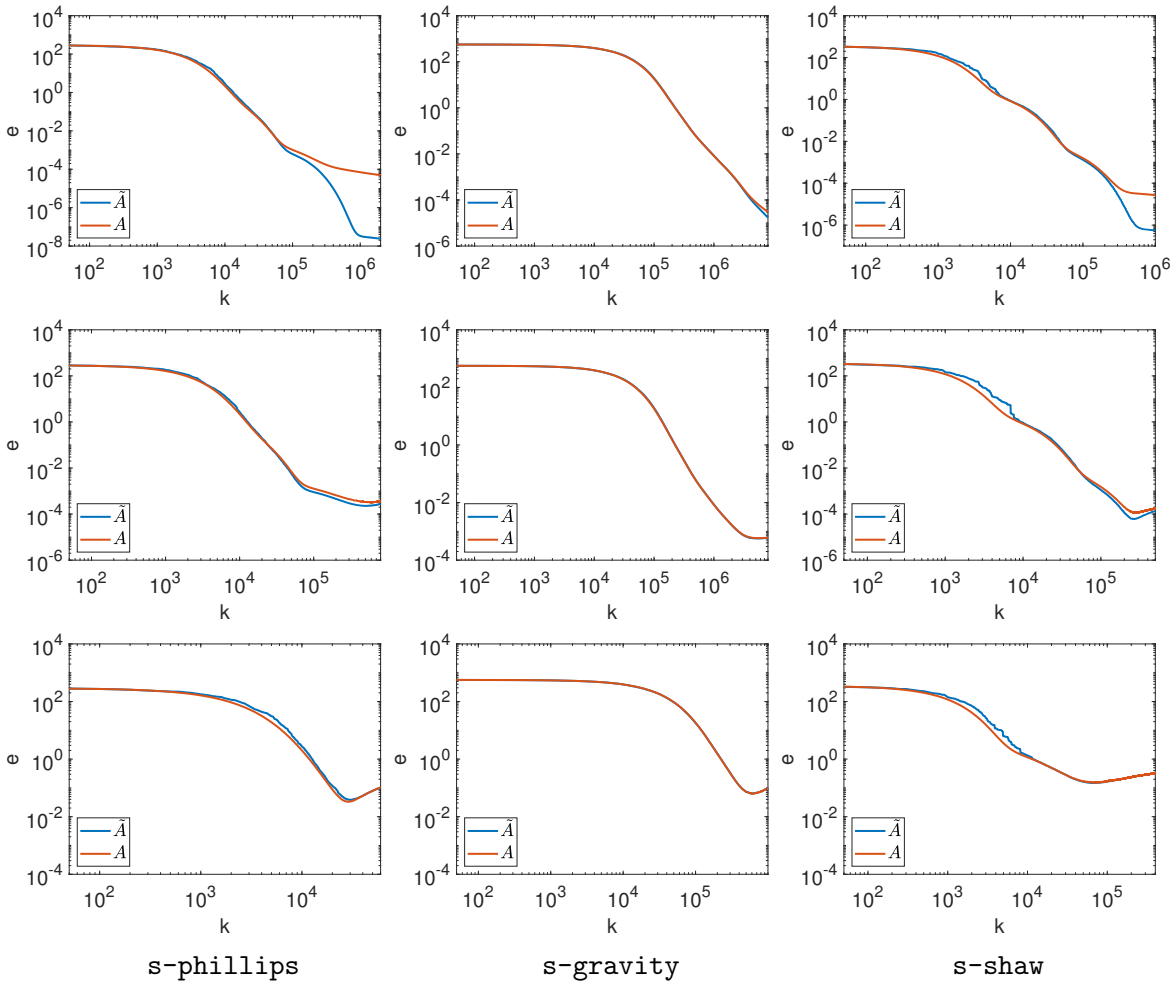


**Figure 4.3.** *The convergence of the error e versus iteration number for the examples with $\nu = 1$, computed using $A$ and $\tilde{A}$. The rows from top to bottom are for $\epsilon = 0$, $\epsilon = 1e\text{-}3$, and $\epsilon = 5e\text{-}2$, respectively.*

## 5. Concluding remarks.

In this work, we have presented a refined convergence rate analysis of stochastic gradient descent with a polynomially decaying stepsize schedule for linear inverse problems, using a finer error decomposition. The analysis indicates that the saturation phenomenon exhibited by existing analysis actually does not occur provided that the initial stepsize $c_0$ is sufficiently small. The analysis is also confirmed by several numerical experiments, which show that with a small $c_0$, the accuracy of SGD is indeed comparable to the order-optimal Landweber method.

The numerical experiments show that Assumption 1.1(iii) is actually not needed for the optimality as long as the initial stepsize $c_0$ is sufficiently small, although the analysis requires

the condition. One outstanding issue is to close the gap between the mathematical theory and practical performance. The study naturally leads to the question of whether there is a "large" stepsize schedule that can achieve optimal convergence rates. The numerical experiments indicate that within polynomially decaying stepsize schedules, a small value of $c_0$ seems to be necessary for achieving order optimality. But the analysis in this work does not cover nonpolynomial schedules, e.g., stagewise SGD [35, 9], which may potentially overcome the saturation phenomenon. Intuitively, the small initial stepsize can be viewed as a form of implicit variance reduction, and thus it is also of interest to analyze existing explicit variance reduction techniques, e.g., SVRG [20] and SAG [24]. The current work discusses only deterministic noise. Naturally it is also of interest to extend the analysis to the case of random noise; see, e.g., the work [1, 12] for relevant results for statistical inverse problems in a Hilbert space setting.

**Acknowledgment.** The authors would like to thank the two anonymous referees for their many constructive comments, which have greatly helped improve the quality of the paper.

## REFERENCES

[1] N. Bissantz, T. Hohage, A. Munk, and F. Ruymgaart, *Convergence rates of general regularization methods for statistical inverse problems and applications*, SIAM J. Numer. Anal., 45 (2007), pp. 2610–2636, https://doi.org/10.1137/060651884.

[2] L. Bottou, *Large-scale machine learning with stochastic gradient descent*, in Proceedings of COMPSTAT'2010, Y. Lechevallier and G. Saporta, eds., Springer, Heidelberg, 2010, pp. 177–186.

[3] L. Bottou, F. E. Curtis, and J. Nocedal, *Optimization methods for large-scale machine learning*, SIAM Rev., 60 (2018), pp. 223–311, https://doi.org/10.1137/16M1080173.

[4] K. Chen, Q. Li, and J.-G. Liu, *Online learning in optical tomography: A stochastic approach*, Inverse Problems, 34 (2018), 075010.

[5] A. Dieuleveut and F. Bach, *Nonparametric stochastic approximation with large step-sizes*, Ann. Statist., 44 (2016), pp. 1363–1399, https://doi.org/10.1214/15-AOS1391.

[6] D. L. Donoho and M. Elad, *Optimally sparse representation in general (nonorthogonal) dictionaries via $l^1$ minimization*, Proc. Natl. Acad. Sci. USA, 100 (2003), pp. 2197–2202, https://doi.org/10.1073/pnas.0437847100.

[7] H. W. Engl, M. Hanke, and A. Neubauer, *Regularization of Inverse Problems*, Kluwer, Dordrecht, The Netherlands, 1996, https://doi.org/10.1007/978-94-009-1740-8.

[8] B. Fehrman, B. Gess, and A. Jentzen, *Convergence rates for the stochastic gradient descent method for non-convex objective functions*, J. Mach. Learn. Res., 21 (2020), 136.

[9] R. Ge, S. M. Kakade, R. Kidambi, and P. Netrapalli, *The step decay schedule: A near optimal, geometrically decaying learning rate procedure for least squares*, in Advances in Neural Information Processing Systems 32 (NIPS 2019), Curran Associates, Red Hook, NY, 2019.

[10] P. C. Hansen, *Regularization Tools version* 4.0 *for MATLAB* 7.3, Numer. Algorithms, 46 (2007), pp. 189–194.

[11] J. HaoChen and S. Sra, *Random shuffling beats SGD after finite epochs*, in Proceedings of the 36th International Conference on Machine Learning, K. Chaudhuri and R. Salakhutdinov, eds., Proc. Mach. Learn. Res. 97, PMLR, 2019, pp. 2624–2633.

[12] B. Harrach, T. Jahn, and R. Potthast, *Beyond the Bakushinkii veto: Regularising linear inverse problems without knowing the noise distribution*, Numer. Math., 145 (2020), pp. 581–603, https://doi.org/10.1007/s00211-020-01122-2.

[13] G. T. Herman, A. Lent, and P. H. Lutz, *Relaxation method for image reconstruction*, Comm. ACM, 21 (1978), pp. 152–158, https://doi.org/10.1145/359340.359351.

[14] K. Ito and B. Jin, *Inverse Problems: Tikhonov Theory and Algorithms*, World Scientific, Hackensack,

NJ, 2015.
[15] T. Jahn and B. Jin, *On the discrepancy principle for stochastic gradient descent*, Inverse Problems, 36 (2020), 095009, https://doi.org/10.1088/1361-6420/abaa58.
[16] A. Jentzen and P. von Wurstemberger, *Lower error bounds for the stochastic gradient descent optimization algorithm: Sharp convergence rates for slowly and fast decaying learning rates*, J. Complexity, 57 (2020), 101438, https://doi.org/10.1016/j.jco.2019.101438.
[17] Y. Jiao, B. Jin, and X. Lu, *Preasymptotic convergence of randomized Kaczmarz method*, Inverse Problems, 33 (2017), 125012.
[18] B. Jin and X. Lu, *On the regularizing property of stochastic gradient descent*, Inverse Problems, 35 (2019), 015004, https://doi.org/10.1088/1361-6420/aaea2a.
[19] B. Jin, Z. Zhou, and J. Zou, *On the convergence of stochastic gradient descent for nonlinear ill-posed problems*, SIAM J. Optim., 30 (2020), pp. 1421–1450, https://doi.org/10.1137/19M1271798.
[20] R. Johnson and T. Zhang, *Accelerating stochastic gradient descent using predictive variance reduction*, in Advances in Neural Information Processing Systems 26 (NIPS 2013) (Lake Tahoe, NV), C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, eds., Curran Associates, Red Hook, NY, 2013, pp. 315–323.
[21] B. Kaltenbacher, A. Neubauer, and O. Scherzer, *Iterative Regularization Methods for Nonlinear Ill-Posed Problems*, Walter de Gruyter GmbH & Co. KG, Berlin, 2008, https://doi.org/10.1515/9783110208276.
[22] H. J. Kushner and G. G. Yin, *Stochastic Approximation and Recursive Algorithms and Applications*, 2nd ed., Springer-Verlag, New York, 2003.
[23] L. Landweber, *An iteration formula for Fredholm integral equations of the first kind*, Amer. J. Math., 73 (1951), pp. 615–624, https://doi.org/10.2307/2372313.
[24] N. Le Roux, M. Schmidt, and F. Bach, *A stochastic gradient method with an exponential convergence rate for finite training sets*, in Advances in Neural Information Processing Systems 25 (NIPS 2012), Curran Associates, Red Hook, NY, pp. 2663–2671.
[25] Y. Lei, T. Hu, and K. Tang, *Generalization performance of multi-pass stochastic gradient descent with convex loss functions*, J. Mach. Learn. Res., 22 (2021), pp. 1–41.
[26] J. Lin and L. Rosasco, *Optimal rates for multi-pass stochastic gradient methods*, J. Mach. Learn. Res., 18 (2017), 97.
[27] F. Natterer, *The Mathematics of Computerized Tomography*, Classics Appl. Math. 32, SIAM, Philadelphia, 2001, https://doi.org/10.1137/1.9780898719284.
[28] L. Pillaud-Vivien, A. Rudi, and F. Bach, *Statistical optimality of stochastic gradient descent on hard learning problems through multiple passes*, in Advances in Neural Information Processing Systems 32 (NIPS 2018), Curran Associates, Red Hook, NY, 2018.
[29] H. Robbins and S. Monro, *A stochastic approximation method*, Ann. Math. Statist., 22 (1951), pp. 400–407.
[30] I. Safran and O. Shamir, *How good is SGD with random shuffling?*, in COLT 2020, J. Abernethy and S. Agarwal, eds., Proc. Mach. Learn. Res. 125, PMLR, 2020, pp. 3250–3284.
[31] T. Strohmer and R. Vershynin, *A randomized Kaczmarz algorithm with exponential convergence*, J. Fourier Anal. Appl., 15 (2009), pp. 262–278, https://doi.org/10.1007/s00041-008-9030-4.
[32] P. Tarrès and Y. Yao, *Online learning as stochastic approximation of regularization paths: Optimality and almost-sure convergence*, IEEE Trans. Inform. Theory, 60 (2014), pp. 5716–5735, https://doi.org/10.1109/TIT.2014.2332531.
[33] B. Ying, K. Yuan, S. Vlaski, and A. H. Sayed, *Stochastic learning under random reshuffling with constant step-sizes*, IEEE Trans. Signal Process., 67 (2018), pp. 474–489.
[34] Y. Ying and M. Pontil, *Online gradient descent learning algorithms*, Found. Comput. Math., 8 (2008), pp. 561–596, https://doi.org/10.1007/s10208-006-0237-y.
[35] Z. Yuan, Y. Yan, R. Jin, and T. Yang, *Stagewise training accelerates convergence of testing error over SGD*, in Advances in Neural Information Processing Systems 32 (NIPS 2019), Curran Associates, Red Hook, NY, 2019.