ON THE CONVERGENCE OF STOCHASTIC GRADIENT DESCENT FOR NONLINEAR ILL-POSED PROBLEMS

BANGTI JIN*, ZEHUI ZHOU†, AND JUN ZOU†

Abstract. In this work, we analyze the regularizing property of stochastic gradient descent for the numerical solution of a class of nonlinear ill-posed inverse problems in Hilbert spaces. At each step of the iteration, the method randomly chooses one equation from the nonlinear system to obtain an unbiased stochastic estimate of the gradient, and then performs a "descent" step with the estimated gradient. It is a randomized version of the classical Landweber method for nonlinear inverse problems, and it is highly scalable to the problem size and holds significant potentials for solving large-scale inverse problems. Under the canonical tangential cone condition, we prove the regularizing property for *a priori* stopping rules, and further, establish the convergence rates under suitable sourcewise condition and range invariance condition.

Key words. stochastic gradient descent, regularizing property, nonlinear inverse problems, convergence rates

AMS subject classifications. 65J20, 65J22, 47J06

1. Introduction.

This work is concerned with the numerical solution of the system of nonlinear ill-posed operator equations

$$F_i(x) = y_i^\dagger, \quad i = 1, \ldots, n, \tag{1.1}$$

where each $F_i : \mathcal{D}(F_i) \to Y$ is a nonlinear mapping with its domain $\mathcal{D}(F_i) \subset X$, and X and Y are Hilbert spaces with inner products $\langle \cdot, \cdot \rangle$ and norms $\|\cdot\|$, respectively. The number n of nonlinear equations in (1.1) can potentially be large. The notation $y_i^\dagger \in Y$ denotes the exact data (corresponding to the reference solution $x^\dagger \in X$ to be defined below). Equivalently, (1.1) can be rewritten as

$$F(x) = y^\dagger, \tag{1.2}$$

with $F : X \to Y^n$ (Y^n denotes the product space $Y \times \cdots \times Y$) and $y^\dagger \in Y^n$ defined by

$$F(x) = \frac{1}{\sqrt{n}} \begin{pmatrix} F_1(x) \\ \cdots \\ F_n(x) \end{pmatrix} \quad \text{and} \quad y^\dagger = \frac{1}{\sqrt{n}} \begin{pmatrix} y_1^\dagger \\ \cdots \\ y_n^\dagger \end{pmatrix},$$

respectively. The scaling $n^{-\frac{1}{2}}$ is introduced for the convenience of later discussions. In practice, we have access only to the noisy data y^δ of a noise level $\delta \geq 0$, i.e.,

$$\|y^\delta - y^\dagger\| = \delta.$$

Nonlinear inverse problems of the form (1.1) arise naturally in many real-world applications, especially parameter identifications for partial differential equations, e.g., electrical impedance tomography and diffuse optical spectroscopy. Due to the *ill-posed* nature of problem (1.1), i.e., a solution may not exist and even if it does exist, the solution may

*Department of Computer Science, University College London, Gower Street, London WC1E 6BT, UK (b.jin@ucl.ac.uk,bangti.jin@gmail.com). The work of BJ is supported by UK EPSRC grant EP/T000864/1.

†Department of Mathematics, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong. (zhzhou@math.cuhk.edu.hk, zou@math.cuhk.edu.hk). The work of JZ was substantially supported by Hong Kong RGC General Research Fund (Project 14304517) and National Natural Science Foundation of China/Hong Kong Research Grants Council Joint Research Scheme 2016/17 (project N CUHK437/16).

be nonunique and highly unstable with respect to the perturbation in the noisy data y^δ, regularization is often needed for their stable and accurate numerical solutions, and many effective techniques have been proposed over the past few decades (see, e.g., [5, 15, 23, 12, 24]). Among existing techniques, iterative regularization represents a very powerful class of solvers for problem (1.1), including Landweber method, (regularized) Gauss-Newton method, conjugate gradient methods, and Leverberg-Marquardt method etc; see the monographs [15] and [24] for overviews on iterative regularization methods in Hilbert spaces and Banach spaces, respectively. In this work, we are interested in the convergence analysis of stochastic gradient descent (SGD) for problem (1.1) with noisy data y^δ. The basic version of SGD reads: given the initial guess $x_1^\delta = x_1$, uppdate the iterate x_k^δ by

$$x_{k+1}^\delta = x_k^\delta - \eta_k F_{i_k}'(x_k^\delta)^*(F_{i_k}(x_k^\delta) - y_{i_k}^\delta); \quad k = 1, 2, \ldots, \tag{1.3}$$

where the index i_k is drawn uniformly from the index set $\{1, \ldots, n\}$, and $\eta_k > 0$ is the corresponding step size. SGD was pioneered by Robbins and Monro in statistical inference [22] (see the monograph [17] for asymptotic convergence results). It has demonstrated encouraging numerical results on diffuse optical tomography [2]. Further, a variant of SGD, i.e., randomized Kaczmarz method (RKM), has been successful in the computed tomography community [9, 10] with revived interest in linear regression and phase retrieval [25, 27]. Algorithmically, SGD is a randomized version of the classical Landweber method [18]

$$x_{k+1}^\delta = x_k^\delta - \eta_k F'(x_k^\delta)^*(F(x_k^\delta) - y^\delta), \tag{1.4}$$

which may be obtained from gradient descent applied to the functional

$$J(x) = \frac{1}{2}\|F(x) - y^\delta\|^2 = \frac{1}{n}\sum_{i=1}^n \frac{1}{2}\|F_i(x) - y_i^\delta\|^2. \tag{1.5}$$

Compared with the Landweber method, SGD requires only evaluating one randomly selected (nonlinear) equation at each iteration, instead of the whole nonlinear system, which substantially reduces the computational cost per iteration and enables excellent scalability to truly massive data sets (i.e., large n), which are increasingly common in practice due to advances in data acquisition technologies. This highly desirable property has attracted much recent interest in machine learning, where currently SGD and its variants are the workhorse for many challenging training tasks involving deep neural networks [32, 26, 16, 1].

Note that due to the ill-posed nature of problem (1.1) (in the sense that the minimizer depends sensitively on the data perturbation), the minimization problem (1.5) is also *ill-posed*, and due to the inevitable presence of noise in the observational data y^δ, the global minimizer (if it exists at all!) often represents a poor approximation to the exact solution $x^\dagger$ and thus is not of interest. The goal of iterative regularization is to iteratively construct an approximate minimizer that converges to the exact solution $x^\dagger$ as the noise level $\delta \to 0^+$, and further, to derive convergence rates in terms of δ. This is achieved by equipping an iterative algorithm, e.g., Landweber method or SGD, with an early stopping strategy. Early stopping allows properly balancing the deleterious effect of the perturbation δ and the approximation error of the iterates for the perturbed data y^δ, which respectively grows and decreases as the iteration proceeds. Thus the setting differs greatly from *well-posed* optimization problems that are extensively studied in the optimization and machine learning literature.

For a class of nonlinear inverse problems, the Landweber method is relatively well understood in terms of the regularizing property, since the influential work [8] (see also [20, 30] for linear inverse problems), and the results were refined and extended in different aspects [15]. In contrast, the stochastic counterparts, e.g., SGD, remains largely under-explored for inverse problems, despite their computational appeals. The theoretical analysis of stochastic iterative methods for inverse problems has just started, and some first theoretical results were obtained in [13, 14] for linear inverse problems. The regularizing property of SGD for linear inverse problems was proved in [14], by drawing on relevant developments in statistical learning theory [31, 4, 19], whereas in [13], the preasymptotic convergence behavior of RKM was analyzed. In this work, we study in depth the regularizing property and convergence rates of SGD for a class of nonlinear inverse problems, under an *a priori* choice of the stopping index and standard assumptions on the nonlinear operator F; see section 2 for further details and discussions. The analysis borrows techniques from the works [14, 8], i.e., handling iteration noise [14] and coping with the nonlinearity of forward map [8]. To the best of our knowledge, this work gives a first thorough analysis of SGD for nonlinear ill-posed inverse problems in the lens of iterative regularization.

There is a vast literature on the convergence of SGD and its variants in optimization and machine learning; see [1, Section 4] for a comprehensive overview; see also [7] and references therein for recent results and [6] for recent results in a Hilbert space setting. For general nonconvex optimization problems, most of the results are concerned with the convergence in terms of either expected optimality gap or expected norm of its gradient, with respect to the iteration index k. However, these works focus on *well-posed* optimization problems, and the ultimate goal is to find a global minimizer. This differs substantially from the setting of *ill-posed* problems, e.g., (1.5). In particular, the existing convergence results of SGD cannot be applied directly to deduce convergence (and rate) for problem (1.5), due to its least-squares structure and different assumptions (on the forward map, instead of the objective functional J; see Remark 2.1 below for further discussions. More closely related to this work are the works [31, 28, 4, 19] on generalization error in statistical learning. Ying and Pontil [31] studied an online least-squares gradient descent algorithm in a reproducing kernel Hilbert space (RKHS), and derived bounds on the generalization error. Lin and Rosasco [19] analyzed the influence of batch size on the convergence of mini-batch SGD. See also the recent work [4] on averaged SGD for nonparametric regression in RKHS. There are also major differences between these interesting works and this study. First, in these prior works, the noise arises mainly due to finite sampling, whereas for inverse problems, it arises from imperfect data acquisition process and enters into the data y^δ directly. Second, the main focus of these works is to bound the generalization error, instead of error estimates on the iterate. Third, these prior works analyzed only linear problems (similar to [14]), instead of nonlinear problems of this work. Nonetheless, our proof strategy of decomposing the mean squared error into the bias and variance components shares similarity with these works.

Throughout, we denote the iterate for the exact data $y^\dagger$ by x_k. The notation $\mathcal{F}_k$ denotes the filtration generated by the random indices $\{i_1, \ldots, i_{k-1}\}$ up to the $(k-1)$th iteration. 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 noise level δ and the iteration number k. We shall abuse $\|\cdot\|$ for the operator norm on Y^n and from X to Y (or Y^n). The rest of the paper is organized as follows. In section 2, we state the main results and provide relevant discussions. Then in section 3 and section 4, we give the proofs on the regularizing property and convergence rate, respectively. The paper concludes with further discussions in section 5. In the appendix, we collect some useful inequalities.

2. Main results and discussions.

To analyze SGD for nonlinear inverse problems, suitable conditions are needed. For example, for Tikhonov regularization, both nonlinearity and source conditions are often employed to derive convergence rates [5, 11, 24, 12]. Below we make a number of assumptions on the nonlinear operators F_i and the reference solution $x^\dagger$. Since the solution to problem (1.1) may be nonunique, the reference solution $x^\dagger$ is taken to be the minimum norm solution (with respect to the initial guess x_1), which is known to be unique under Assumption 2.1(ii) below [8].

ASSUMPTION 2.1. *The following conditions hold:*

(i) *The operator $F : X \to Y^n$ is continuous, with a continuous and uniformly bounded Frechét derivative on X.*

(ii) *There exists an $\eta \in (0, \frac{1}{2})$ such that for any $x, \tilde{x} \in X$,*

$$\|F(x) - F(\tilde{x}) - F'(\tilde{x})(x - \tilde{x})\| \leq \eta \|F(x) - F(\tilde{x})\|. \tag{2.1}$$

(iii) *There are a family of uniformly bounded operators R_x^i such that for any $x \in X$, $F_i'(x) = R_x^i F_i'(x^\dagger)$ and $R_x = \mathrm{diag}(R_x^i) : Y^n \to Y^n$, with*

$$\|R_x - I\| \leq c_R \|x - x^\dagger\|.$$

(iv) *The source condition holds: there exist some $\nu \in (0, \frac{1}{2})$ and $w \in X$ such that*

$$x^\dagger - x_1 = (F'(x^\dagger)^* F'(x^\dagger))^\nu w.$$

The conditions in Assumption 2.1 are standard for analyzing iterative regularization methods for nonlinear inverse problems [8, 15]. (i) is smilar to the λ-smoothness commonly used in optimization. (ii)–(iii) have been verified for a class of nonlinear inverse problems [8], e.g., parameter identification for PDEs and nonlinear integral equations. The inequality (2.1) is often known as tangential cone condition, and it controls the degree of nonlinearity of the operator F. Roughly speaking, it requires the map F be not far from a linear map; see Lemma 3.1 for the consequences. The fractional power $(F'(x^\dagger)^* F'(x^\dagger))^\nu$ in (iv) is defined by spectral decomposition (e.g., via Dunford-Taylor integral). Customarily, it represents a certain smoothness condition on the exact solution $x^\dagger$ (relative to the initial guess x_1). The restriction $\nu < \frac{1}{2}$ is due to technical reasons. It is worth noting that most results require only (i)–(ii), especially the convergence of SGD, whereas (iii)–(iv) are only needed for proving the convergence rate of SGD.

REMARK 2.1. *It is instructive to compare Assumption 2.1 with the canoical conditions for the usual finite-sum optimization:*

$$\mathcal{F}(x) = n^{-1} \sum_{i=1}^{n} f_i(x). \tag{2.2}$$

Clearly problem (1.5) is a special case of (2.2), with the choice $f_i(x) = \frac{1}{2}\|F_i(x) - y_i^\delta\|^2$. In the literature on SGD for problem (2.2), the following two conditions are often adopted

- *L-smoothness: $\|\mathcal{F}'(x) - \mathcal{F}'(\tilde{x})\| \leq L\|x - \tilde{x}\|$*
- *λ-convexity: $\mathcal{F}(x) \geq \mathcal{F}(\tilde{x}) + (\mathcal{F}'(\tilde{x}), x - \tilde{x}) + \frac{\lambda}{2}\|x - \tilde{x}\|^2$.*

Under these conditions, various convergence results have been established; see [1, Section 4].

Assumption 2.1(i) imposes boundness and continuity on the derivative $F'(u)$, which does not imply directly the L-smoothness condition. Nonetheless, the Lipschitz continuity of $F'(u)$ can be verified for a number of inverse problems, which then implies the L-smoothness condition. Assumption 2.1(ii) requires the forward map being not too far from a linear map, and thus one might expect a link with the λ-convexity, which, however, seems not evident. Straightforward computation gives $\nabla^2 J(x) = F'(x)^* F'(x) + \nabla^2 F(x)^*(F(x) - y^\delta)$. First, the map F is not assumed a priori twice differentiable so that $J(x)$ admits a Hessian $\nabla^2 J(x)$. Second, if the Hessian $\nabla^2 F$ does exist, then Taylor expansion gives

$$\|F(x) - F(\tilde{x}) - F'(\tilde{x})(x - \tilde{x})\| = \|\tfrac{1}{2}\nabla^2 F(\tilde{x})(x - \tilde{x})^2 + \mathcal{O}(|x - \tilde{x}|^3)\| \leq \eta \|F(x) - F(\tilde{x})\|.$$

Unfortunately it does not imply directly that $\nabla^2 F$ is small. Further, $F'(x)^*F'(x)$ is usually only positive semidefinitive, since the linearized operator $F'(x)$ is degenerate (e.g. compact) for most ill-posed inverse problems, so even if $\nabla^2 F(\tilde{x})$ is small, generally one cannot ensure $\nabla^2 J(x) \geq 0$, i.e., the convexity. In sum, (2.1) does not imply the λ-convexity condition. Thus Assumption 2.1 is not directly comparable with standard assumptions for SGD, and the convergence results in [1] cannot be applied directly.

We also need suitable assumptions on the step size schedule $\{\eta_k\}_{k=1}^\infty$. The choice is viable since $\max_i \sup_{x\in X} \|F_i'(x)\| < \infty$, by Assumption 2.1(i). The choice in Assumption 2.2(i) is more general than (ii). The latter choice is often known as a polynomially decaying step size schedule in the literature.

ASSUMPTION 2.2. *The step sizes $\{\eta_k\}_{k\geq 1}$ satisfy one of the following conditions.*
(i) $\eta_k \max_i \sup_{x\in X} \|F_i'(x)\|^2 < 1$ *and* $\sum_{k=1}^\infty \eta_k = \infty$.
(ii) $\eta_k = \eta_0 k^{-\alpha}$, *with* $\alpha \in (0,1)$ *and* $\eta_0 \leq (\max_i \sup_{x\in X} \|F_i'(x)\|^2)^{-1}$.

Due to the random choice of the index i_k, the SGD iterate x_k^δ is random. There are several different ways to measure the convergence. We shall employ the mean squared norm defined by $\mathbb{E}[\|\cdot\|^2]$, where the expectation $\mathbb{E}[\cdot]$ is with respect to the filtration $\mathcal{F}_k$. Clearly, the iterate x_k^δ is measurable with respect to $\mathcal{F}_k$. The first result gives the regularizing property of SGD for problem (1.1) under *a priori* parameter choice. The notation $\mathcal{N}(\cdot)$ denotes the kernel of a linear operator.

THEOREM 2.1 (convergence for noisy data). *Let Assumption 2.1(i)-(ii) and Assumption 2.2(i) be fulfilled. If the stopping index $k(\delta) \in \mathbb{N}$ satisfies $\lim_{\delta\to 0^+} k(\delta) = \infty$ and $\lim_{\delta\to 0^+} \delta^2 \sum_{i=1}^{k(\delta)} \eta_i = 0$, then there exists a solution $x^* \in X$ to problem (1.1) such that*

$$\lim_{\delta\to 0^+} \mathbb{E}[\|x_{k(\delta)}^\delta - x^*\|^2] = 0.$$

Further, if $\mathcal{N}(F'(x^\dagger)) \subset \mathcal{N}(F'(x))$, then

$$\lim_{\delta\to 0^+} \mathbb{E}[\|x_{k(\delta)}^\delta - x^\dagger\|^2] = 0.$$

REMARK 2.2. *The conditions on $k(\delta)$ in Theorem 2.1 are identical with that for the Landweber method [8, Theorem 2.4]. Note that consistency does not require a monotonically decreasing step size schedule, and holds for a constant step size.*

Next we make an assumption on the nonlinearity of the operator F in a stochastic sense.

ASSUMPTION 2.3. *There exist some $\theta \in (0,1]$ and $c_R > 0$ such that for any function $G: X \to Y^n$ and $z_t = tx_k^\delta + (1-t)x^\dagger$, $t \in [0,1]$, there hold*

$$\mathbb{E}[\|(I - R_{z_t})G(x_k^\delta)\|^2]^{\frac{1}{2}} \leq c_R \mathbb{E}[\|x_k^\delta - x^\dagger\|^2]^{\frac{\theta}{2}} \mathbb{E}[\|G(x_k^\delta)\|^2]^{\frac{1}{2}},$$
$$\mathbb{E}[\|(I - R_{z_t}^*)G(x_k^\delta)\|^2]^{\frac{1}{2}} \leq c_R \mathbb{E}[\|x_k^\delta - x^\dagger\|^2]^{\frac{\theta}{2}} \mathbb{E}[\|G(x_k^\delta)\|^2]^{\frac{1}{2}}.$$

Assumption 2.3 is a stochastic version of Assumption 2.1(iii), and strengthens the corresponding estimate in the sense of expectation. The case $\theta = 0$ follows trivially from Assumption 2.1(iii), by the boundedness of the operator R_x, whereas with $\theta = 1$, it recovers the latter when specialized to a Dirac measure. It will play a role in the convergence rate analysis, by taking $G(x) = F(x) - y^\delta$ and $G(x) = F'(x^\dagger)(x - x^\dagger)$ (see the proofs in Lemma 4.1 and Lemma 4.6), and it enables bounding the terms involving conditional dependence.

The next result gives a convergence rate under *a priori* parameter choice, i.e., bound on the error $e_k^\delta := x_k^\delta - x^\dagger$, in terms of δ and k etc. The notation $[\cdot]$ denotes taking the integral part of a real number, provided that $\|F'(x^\dagger)^*F'(x^\dagger)\| \leq 1$ and $\eta_0 \leq 1$. The assumptions in Theorem 2.2 are identical with that for the Landweber method [8], except Assumption 2.3. The strategy of the error analysis is to split the mean squared error $\mathbb{E}[\|e_k^\delta\|^2]$ using bias-variance decomposition: with bias $\|\mathbb{E}[e_k^\delta]\|^2$ and variance $\mathbb{E}[\|e_k^\delta - \mathbb{E}[e_k^\delta]\|^2]$,

$$\mathbb{E}[\|e_k^\delta\|^2] = \|\mathbb{E}[e_k^\delta]\|^2 + \mathbb{E}[\|e_k^\delta - \mathbb{E}[e_k^\delta]\|^2]. \tag{2.3}$$

The former contains the approximation error and data error, whereas the latter arises from the random choice of the index i_k. Due to the nonlinearity of the operator F, the two terms interact with each other (and also $\mathbb{E}[\|F'(x^\dagger)e_k^\delta\|^2]$); see Theorem 4.4 and Theorem 4.7. This leads to a coupled system of recursive inequalities for $\mathbb{E}[\|e_k^\delta\|^2]$ and $\mathbb{E}[\|F'(x^\dagger)e_k^\delta\|^2]$, and thus the analysis differs substantially from that for linear inverse problems in [14] and the Landweber method for nonlinear inverse problems [8].

Theorem 2.2. *Let Assumption 2.1, Assumption 2.2(ii) and Assumption 2.3 be fulfilled with $\|w\|$ and η_0 being sufficiently small, and x_k^δ be the SGD iterate defined in (1.3). Then for all $k \leq k^* = [(\frac{\delta}{\|w\|})^{-\frac{2}{(2\nu+1)(1-\alpha)}}]$ and small $\epsilon \in (0, \frac{\alpha}{2})$, there hold*

$$\mathbb{E}[\|e_k^\delta\|^2] \leq c^* k^{-\min(2\nu(1-\alpha),\alpha-\epsilon)}\|w\|^2 \quad \text{and} \quad \mathbb{E}[\|F'(x^\dagger)e_k^\delta\|^2] \leq c^* k^{-\min((1+2\nu)(1-\alpha),1-\epsilon)}\|w\|^2,$$

where the constant c^* depends on ν, α, η_0, n and θ, but is independent of k and δ.

Remark 2.3. *When $\alpha \in (0,1)$ is close to 1, setting $k = k^*$ gives*

$$\mathbb{E}[\|e_{k^*}^\delta\|^2] \leq c^*\|w\|^{\frac{2}{2\nu+1}}\delta^{\frac{4\nu}{2\nu+1}} \quad \text{and} \quad \mathbb{E}[\|F'(x^\dagger)e_{k^*}^\delta\|^2] \leq c^*\|w\|^{\frac{4\nu}{2\nu+1}}\delta^{\frac{2}{2\nu+1}}.$$

These rates are comparable with that for the Landweber method for nonlinear inverse problems [8, Theorem 3.2] and SGD for linear inverse problems [14, Theorem 2.2]. The restriction $O(k^{-(\alpha-\epsilon)})$ is due to the computational variance arising from the random index i_k, and for small α, the convergence rate may suffer from a loss. It is noteworthy that for $\nu > 1/2$, the convergence rate is suboptimal, just as the classical Landweber method, and thus SGD may suffer from a saturation phenomenon. It is an interesting open question to remove the saturation phenomenon.

Remark 2.4. *In practice, the domain $\mathcal{D}(F) \subset X$ is often not the whole space X, especially for parameter identifications for PDEs, where box constraints arise naturally due to physical constraints. When the domain $\mathcal{D}(F) \subset X$ is a closed convex set, e.g., box constraints, it can be incorporated into the algorithm by a projection operator P [29], i.e.,*

$$x_{k+1}^\delta = P(x_k^\delta - \eta_k F_{i_k}'(x_k^\delta)^*(F_{i_k}(x_k^\delta) - y_{i_k}^\delta)).$$

However, the presence of the projection P significantly complicates the analysis. The extension to the constrained case is an interesting open question.

3. Convergence of SGD.

Now we analyze the convergence of SGD, and give the proof of Theorem 2.1. We first recall a useful characterization of an exact solution x^* [8, Proposition 2.1].

Lemma 3.1. *The following statements hold under Assumption 2.1(i)–(ii).*

(i) *The following upper and lower bounds hold:*

$$\tfrac{1}{1+\eta}\|F'(x)(x-\tilde{x})\| \leq \|F(x) - F(\tilde{x})\| \leq \tfrac{1}{1-\eta}\|F'(x)(x-\tilde{x})\|.$$

(ii) *If x^* is a solution of problem (1.1), then any other solution $\tilde{x}^*$ satisfies $x^* - \tilde{x}^* \in \mathcal{N}(F'(x^*))$, and vice versa.*

The next result gives a crucial monotonicity result of the mean squared error.

PROPOSITION 3.1. *Under Assumption 2.1(i)-(ii) and Assumption 2.2(i), for any solution x^* to problem (1.1), there holds*

$$\begin{aligned}\mathbb{E}[\|x^* - x_{k+1}^\delta\|^2] - \mathbb{E}[\|x^* - x_k^\delta\|^2] \leq & -(1-2\eta)\eta_k \mathbb{E}[\|F(x_k^\delta) - y^\delta\|^2] \\ & + 2\eta_k(1+\eta)\delta \mathbb{E}[\|F(x_k^\delta) - y^\delta\|^2]^{\frac{1}{2}}.\end{aligned}$$

Proof. Completing the square using the definition of the iterate x_k^δ in (1.3) gives

$$\begin{aligned}&\|x^* - x_{k+1}^\delta\|^2 - \|x^* - x_k^\delta\|^2 \\ =& -2\eta_k\langle F'_{i_k}(x_k^\delta)(x_k^\delta - x^*), F_{i_k}(x_k^\delta) - y_{i_k}^\delta\rangle + \eta_k^2\|F'_{i_k}(x_k^\delta)^*(F_{i_k}(x_k^\delta) - y_{i_k}^\delta)\|^2.\end{aligned}$$

Using the splitting $F'_{i_k}(x_k^\delta)(x_k^\delta - x^*) = (F_{i_k}(x_k^\delta) - y_{i_k}^\delta) + (y_{i_k}^\delta - y_{i_k}^\dagger) + (y_{i_k}^\dagger - F_{i_k}(x_k^\delta) - F'_{i_k}(x_k^\delta)(x^* - x_k^\delta))$, by the condition $\eta_k\|F'_{i_k}(x)\|^2 < 1$ in Assumption 2.2(i), we obtain

$$\begin{aligned}&\|x^* - x_{k+1}^\delta\|^2 - \|x^* - x_k^\delta\|^2 \\ =& -2\eta_k\langle F_{i_k}(x_k^\delta) - y_{i_k}^\delta, F_{i_k}(x_k^\delta) - y_{i_k}^\delta\rangle + \eta_k^2\|F'_{i_k}(x_k^\delta)^*(F_{i_k}(x_k^\delta) - y_{i_k}^\delta)\|^2 \\ & -2\eta_k\langle y_{i_k}^\delta - y_{i_k}^\dagger, F_{i_k}(x_k^\delta) - y_{i_k}^\delta\rangle \\ & -2\eta_k\langle y_{i_k}^\dagger - F_{i_k}(x_k^\delta) - F'_{i_k}(x_k^\delta)(x^* - x_k^\delta), F_{i_k}(x_k^\delta) - y_{i_k}^\delta\rangle \\ \leq& -\eta_k\langle F_{i_k}(x_k^\delta) - y_{i_k}^\delta, F_{i_k}(x_k^\delta) - y_{i_k}^\delta\rangle - 2\eta_k\langle y_{i_k}^\delta - y_{i_k}^\dagger, F_{i_k}(x_k^\delta) - y_{i_k}^\delta\rangle \\ & -2\eta_k\langle y_{i_k}^\dagger - F_{i_k}(x_k^\delta) - F'_{i_k}(x_k^\delta)(x^* - x_k^\delta), F_{i_k}(x_k^\delta) - y_{i_k}^\delta\rangle.\end{aligned}$$

Next, by the measurability of x_k with respect to $\mathcal{F}_k$, Cauchy-Schwarz inequality and Assumption 2.1(i), we have

$$\begin{aligned}&\mathbb{E}[\|x^* - x_{k+1}^\delta\|^2 - \|x^* - x_k^\delta\|^2|\mathcal{F}_k] \\ \leq& -\eta_k\|F(x_k^\delta) - y^\delta\|^2 - 2\eta_k\langle y^\delta - y^\dagger, F(x_k^\delta) - y^\delta\rangle \\ & -2\eta_k\langle y^\dagger - F(x_k^\delta) - F'(x_k^\delta)(x^* - x_k^\delta), F(x_k^\delta) - y^\delta\rangle \\ \leq& -\eta_k\|F(x_k^\delta) - y^\delta\|^2 + 2\eta_k\delta\|F(x_k^\delta) - y^\delta\| + 2\eta_k\eta\|F(x_k^\delta) - y^\dagger\|\|F(x_k^\delta) - y^\delta\| \\ \leq& \eta_k\|F(x_k^\delta) - y^\delta\|\big((2\eta - 1)\|F(x_k^\delta) - y^\delta\| + 2(1+\eta)\delta\big).\end{aligned}$$

Last, taking full conditional yields the desired assertion ∎

Below we analyze the convergence of SGD for exact and noisy data separately.

3.1. Convergence for exact data. The next result is direct from Proposition 3.1.

COROLLARY 3.2. *Let Assumption 2.1(i)-(ii) and Assumption 2.2(i) be fulfilled. Then for the exact data $y^\dagger$, any solution x^* to problem (1.1) satisfies*

$$\mathbb{E}[\|x^* - x_{k+1}\|^2] - \mathbb{E}[\|x^* - x_k\|^2] \leq -(1-2\eta)\eta_k\mathbb{E}[\|F(x_k) - y^\dagger\|^2],$$

$$\sum_{k=1}^{\infty}\eta_k\mathbb{E}[\|F(x_k) - y^\dagger\|^2] \leq \tfrac{1}{1-2\eta}\|x^* - x_1\|^2.$$

REMARK 3.1. *Corollary 3.2 shows that the mean squared error $\mathbb{E}[\|x_k - x^*\|^2]$ is monotonically decreasing, but the mean squared residual $\mathbb{E}[\|F(x_k) - y^\dagger\|^2]$ is not necessarily so. The latter reflects the fact that the estimated gradient is not guaranteed to be descent.*

The next result shows that the sequence $\{x_k\}_{k\geq 1}$ is a Cauchy sequence.

Lemma 3.3. *Under Assumption 2.1(i)-(ii) and Assumption 2.2(i), for the exact data $y^\dagger$, the sequence $\{x_k\}_{k\geq 1}$ generated by SGD (1.3) is a Cauchy sequence.*

Proof. The argument below follows closely [8, Theorem 2.3], which can be traced back to [21]. Let x^* be any solution to problem (1.1), and let $e_k := x_k - x^*$. By Corollary 3.2, $\mathbb{E}[\|e_k\|^2]$ is monotonically decreasing to some $\epsilon \geq 0$. Next we show that the sequence $\{x_k\}_{k\geq 1}$ is actually a Cauchy sequence. First we note that $\mathbb{E}[\langle\cdot,\cdot\rangle]$ defines an inner product. For any $j \geq k$, choose an index ℓ with $j \geq \ell \geq k$ such that

$$\mathbb{E}[\|y^\dagger - F(x_\ell)\|^2] \leq \mathbb{E}[\|y^\dagger - F(x_i)\|^2], \quad \forall k \leq i \leq j. \tag{3.1}$$

By the inequality $\mathbb{E}[\|e_j - e_k\|^2]^{\frac{1}{2}} \leq \mathbb{E}[\|e_j - e_\ell\|^2]^{\frac{1}{2}} + \mathbb{E}[\|e_\ell - e_k\|^2]^{\frac{1}{2}}$ and the identities

$$\begin{aligned}\mathbb{E}[\|e_j - e_\ell\|^2] &= 2\mathbb{E}[\langle e_\ell - e_j, e_\ell\rangle] + \mathbb{E}[\|e_j\|^2] - \mathbb{E}[\|e_\ell\|^2],\\ \mathbb{E}[\|e_\ell - e_k\|^2] &= 2\mathbb{E}[\langle e_\ell - e_k, e_\ell\rangle] + \mathbb{E}[\|e_k\|^2] - \mathbb{E}[\|e_\ell\|^2],\end{aligned} \tag{3.2}$$

it suffices to prove that both $\mathbb{E}[\|e_j - e_\ell\|^2]$ and $\mathbb{E}[\|e_\ell - e_k\|^2]$ tend to zero as $k \to \infty$. For $k \to \infty$, the last two terms on each of the right-hand side of (3.2) tend to $\epsilon - \epsilon = 0$, by the monotone convergence of $\mathbb{E}[\|e_k\|^2]$ to ϵ, cf. Corollary 3.2. Next we show that the term $\mathbb{E}[\langle e_\ell - e_k, e_\ell\rangle]$ also tends to zero as $k \to \infty$. Actually, by the definition of x_k, we have

$$e_\ell - e_k = \sum_{i=k}^{\ell-1}(e_{i+1} - e_i) = \sum_{i=k}^{\ell-1}\eta_i F'_{i_i}(x_i)^*(y^\dagger_{i_i} - F_{i_i}(x_i)).$$

By triangle inequality and Cauchy-Schwarz inequality, we have

$$\begin{aligned}|\mathbb{E}[\langle e_\ell - e_k, e_\ell\rangle]| &\leq \sum_{i=k}^{\ell-1}\eta_i|\mathbb{E}[\langle F'_{i_i}(x_i)^*(y^\dagger_{i_i} - F_{i_i}(x_i)), e_\ell\rangle]|\\ &= \sum_{i=k}^{\ell-1}\eta_i|\mathbb{E}[\langle y^\dagger_{i_i} - F_{i_i}(x_i), F'_{i_i}(x_i)(x^* - x_i + x_i - x_\ell)\rangle]|\\ &= \sum_{i=k}^{\ell-1}\eta_i|\mathbb{E}[\langle y^\dagger - F(x_i), F'(x_i)(x^* - x_i + x_i - x_\ell)\rangle]|\\ &\leq \sum_{i=k}^{\ell-1}\eta_i\mathbb{E}[\|y^\dagger - F(x_i)\|^2]^{\frac{1}{2}}\mathbb{E}[\|F'(x_i)(x^* - x_i)\|^2]^{\frac{1}{2}}\\ &\quad + \sum_{i=k}^{\ell-1}\eta_i\mathbb{E}[\|y^\dagger - F(x_i)\|^2]^{\frac{1}{2}}\mathbb{E}[\|F'(x_i)(x_i - x_\ell)\|^2]^{\frac{1}{2}} := \mathrm{I} + \mathrm{II}.\end{aligned}$$

By Assumption 2.1(ii) and Lemma 3.1(i), we bound the first term I by

$$\begin{aligned}\mathrm{I} &\leq (1+\eta)\sum_{i=k}^{\ell-1}\eta_i\mathbb{E}[\|y^\dagger - F(x_i)\|^2]^{\frac{1}{2}}\mathbb{E}[\|F(x^*) - F(x_i)\|^2]^{\frac{1}{2}}\\ &= (1+\eta)\sum_{i=k}^{\ell-1}\eta_i\mathbb{E}[\|y^\dagger - F(x_i)\|^2].\end{aligned}$$

Likewise, we bound the term II by triangle inequality and the choice of ℓ in (3.1) as:

$$\mathrm{II} \leq (1+\eta)\sum_{i=k}^{\ell-1}\eta_i\mathbb{E}[\|y^\dagger - F(x_i)\|^2]^{\frac{1}{2}}\mathbb{E}[\|(F(x_\ell)-y^\dagger)+(y^\dagger - F(x_i))\|^2]^{\frac{1}{2}}$$
$$\leq 2(1+\eta)\sum_{i=k}^{\ell-1}\eta_i\mathbb{E}[\|y^\dagger - F(x_i)\|^2].$$

The last two estimates together imply $|\mathbb{E}[\langle e_\ell - e_k, e_\ell\rangle]| \leq 3(1+\eta)\sum_{i=k}^{\ell-1}\eta_i\mathbb{E}[\|y^\dagger - F(x_i)\|^2]$. Similarly, one can deduce $\mathbb{E}[\langle e_j - e_\ell, e_\ell\rangle]| \leq 3(1+\eta)\sum_{i=\ell}^{j-1}\eta_i\mathbb{E}[\|y^\dagger - F(x_i)\|^2]$. These two estimates and Corollary 3.2 imply that the right-hand sides of (3.2) tend to zero as $k\to\infty$. Hence both $\{e_k\}_{k\geq 1}$ and $\{x_k\}_{k\geq 1}$ are Cauchy sequences. ∎

LEMMA 3.4. *Under Assumption 2.1(i)-(ii) and Assumption 2.2(i), there holds*

$$\lim_{k\to\infty}\mathbb{E}[\|F(x_k)-y^\dagger\|^2] = 0.$$

Proof. Lemma 3.3 implies that $\{x_k\}_{k\geq 1}$ is a Cauchy sequence. By Assumption 2.2(i), $\sup_{x\in X}\|F'(x)\| \leq c_F$ for some $c_F > 0$. Further, for any $x, \tilde{x} \in X$, there holds

$$\|F(x) - F(\tilde{x})\| \leq (1-\eta)^{-1}\|F'(x)(x-\tilde{x})\| \leq c_F(1-\eta)^{-1}\|x-\tilde{x}\|.$$

Thus, $\{F(x_k)-y^\dagger\}_{k\geq 1}$ is a Cauchy sequence, and $\mathbb{E}[\|F(x_k)-y^\dagger\|^2]$ converges. Now we proceed by contradiction, and assume that $\lim_{k\to\infty}\mathbb{E}[\|F(x_k)-y^\dagger\|^2] > 0$. Then there exist some $\epsilon > 0$ and $k^* \in \mathbb{N}$, such that $\mathbb{E}[\|F(x_k)-y^\dagger\|^2] \geq \epsilon$ for all $k \geq k^*$. Hence, by Assumption 2.2(i),

$$\sum_{k=1}^{\infty}\eta_k\mathbb{E}[\|F(x_k)-y^\dagger\|^2] \geq \sum_{k=k^*}^{\infty}\eta_k\mathbb{E}[\|F(x_k)-y^\dagger\|^2] \geq \epsilon\sum_{k=k^*}^{\infty}\eta_k = \infty,$$

which contradicts the inequality $\sum_{k=1}^{\infty}\eta_k\mathbb{E}[\|F(x_k)-y^\dagger\|^2] < \infty$ from Corollary 3.2. ∎

Now we can state the convergence of SGD for the exact data $y^\dagger$. Below $x^\dagger$ denotes the unique solution to problem (1.1) of minimal distance to x_1.

THEOREM 3.5 (Convergence for exact data). *Let Assumption 2.1(i)-(ii) and Assumption 2.2(i) be fulfilled. Then for the exact data $y^\dagger$, the sequence $\{x_k\}_{k\geq 1}$ generated by SGD converges to a solution x^* of problem (1.1):*

$$\lim_{k\to\infty}\mathbb{E}[\|x_k - x^*\|^2] = 0.$$

Further, if $\mathcal{N}(F'(x^\dagger)) \subset \mathcal{N}(F'(x))$, then

$$\lim_{k\to\infty}\mathbb{E}[\|x_k - x^\dagger\|^2] = 0.$$

Proof. Since $\{x_k\}_{k\geq 1}$ is a Cauchy sequence, it has a limit, denoted by x^*. Further, x^* is a solution, since by Lemma 3.4, the mean squared residual $\mathbb{E}[\|y^\dagger - F(x_k)\|^2]$ converges to zero as $k\to\infty$. Note that problem (1.1) has a unique solution of minimal distance to the initial guess x_1 that satisfies $x^\dagger - x_1 \in \mathcal{N}(F'(x^\dagger))^\perp$; see Lemma 3.1. If $\mathcal{N}(F'(x^\dagger)) \subset \mathcal{N}(F'(x_k))$ for all $k = 1, 2, \ldots$, then clearly, $x_k - x_1 \in \mathcal{N}(F'(x^\dagger))^\perp$, $k = 1, 2, \ldots$. Hence, $x^\dagger - x^* = x^\dagger - x_1 + x_1 - x^* \in \mathcal{N}(F'(x^\dagger))^\perp$. This and Lemma 3.1 imply $x^* = x^\dagger$. ∎

REMARK 3.2. *Theorem 3.5 does not impose any constraint on the step size schedule $\{\eta_k\}_{k=1}^\infty$ directly, apart from the fact that it should not decay too fast to zero. In particular, it can be taken to be a constant step size. This result slightly improves that in [14, Theorem 2.1], where a decreasing step size is required (for linear inverse problems). The improvement is achieved by exploiting the quadratic structure of the functional $J(x)$ in (1.5) (and the tangential cone condition in Assumption 2.1(i)), whereas in [14] the consistency is derived by means of bias-variance decomposition.*

3.2. Convergence for noisy data. The next result gives the stability of the SGD iterate x_k^δ with respect to the noise level δ (at $\delta = 0$).

LEMMA 3.6. *Let Assumption 2.1(i) be fulfilled. For any fixed $k \in \mathbb{N}$ and any path $(i_1, \ldots, i_{k-1}) \in \mathcal{F}_k$, let x_k and x_k^δ be the SGD iterates along the path for exact data $y^\dagger$ and noisy data y^δ, respectively. Then*

$$\lim_{\delta\to 0^+} \mathbb{E}[\|x_k^\delta - x_k\|^2] = 0.$$

Proof. We prove the assertion by mathematical induction. It holds trivially for $k = 1$. Now suppose that it holds for all indices up to k and any path in $\mathcal{F}_k$. By the definition, for any fixed path $(i_1, \ldots, i_k)$, we have

$$\begin{aligned} x_{k+1}^\delta - x_{k+1} = (x_k^\delta - x_k) - \eta_k\big((F'_{i_k}(x_k^\delta)^* - F'_{i_k}(x_k)^*)(F_{i_k}(x_k^\delta) - y_{i_k}^\delta) \\ + F'_{i_k}(x_k)^*((F_{i_k}(x_k^\delta) - y_{i_k}^\delta) - (F_{i_k}(x_k) - y_{i_k}^\dagger))\big). \end{aligned}$$

Thus, by triangle inequality,

$$\begin{aligned} \|x_{k+1}^\delta - x_{k+1}\| \le \|x_k^\delta - x_k\| + \eta_k\|F'_{i_k}(x_k^\delta)^* - F'_{i_k}(x_k)^*\|\|F_{i_k}(x_k^\delta) - y_{i_k}^\delta\| \\ + \eta_k\|F'_{i_k}(x_k)^*\|\|(F_{i_k}(x_k^\delta) - y_{i_k}^\delta) - (F_{i_k}(x_k) - y_{i_k}^\dagger)\|. \end{aligned} \tag{3.3}$$

Next we show that for any fixed k, $\sup_{(i_1,\ldots,i_{k-1})\in\mathcal{F}_k} \|x_k\|$ is bounded. Indeed, by Assumption 2.1(i), $\max_i \sup_{x\in X} \|F'_i(x)\| \le c_F$ for some $c_F > 0$. Then, by Lemma 3.1(i)

$$\|x_{k+1} - x^*\| \le \|x_k - x^*\| + \eta_k\|F'_{i_k}(x_k)^*\|\|F_{i_k}(x_k) - y_{i_k}^\dagger\| \le (1 + \eta_k \tfrac{c_F^2}{1-\eta})\|x_k - x^*\|.$$

This and an induction argument show that the claim. Similarly,

$$\begin{aligned} \|F_{i_k}(x_k^\delta) - y_{i_k}^\delta\| &\le \|F_{i_k}(x_k^\delta) - F_{i_k}(x_k)\| + \|F_{i_k}(x_k) - y_{i_k}^\dagger\| + \|y_{i_k}^\dagger - y_{i_k}^\delta\| \\ &\le \tfrac{c_F}{1-\eta}\big(\|x_k^\delta - x_k\| + \|x_k - x^*\|\big) + \delta, \end{aligned}$$

and consequently,

$$\begin{aligned} \|x_{k+1}^\delta - x_{k+1}\| &\le \|x_k^\delta - x_k\| + \eta_k(\tfrac{c_F}{1-\eta}(\|x_k^\delta - x_k\| + \|x_k - x^*\|) + \delta)\|F'_{i_k}(x_k^\delta)^* - F'_{i_k}(x_k)^*\| \\ &\quad + c_F\|((F_{i_k}(x_k^\delta) - y_{i_k}^\delta) - (F_{i_k}(x_k) - y_{i_k}^\dagger))\| \\ &\le \|x_k^\delta - x_k\| + 2\eta_k c_F(\tfrac{c_F}{1-\eta}(\|x_k^\delta - x_k\| + \|x_k - x^*\|) + \delta) \\ &\quad + c_F\|((F_{i_k}(x_k^\delta) - y_{i_k}^\delta) - (F_{i_k}(x_k) - y_{i_k}^\dagger))\|, \end{aligned}$$

This and mathematical induction shows that for any fixed k, $\sup_{(i_1,\ldots,i_{k-1})\in\mathcal{F}_k} \|x_k^\delta - x_k\|$ is uniformly bounded. Let $c = \frac{c_F}{1-\eta}\sup_{(i_1,\ldots,i_{k-1})\in\mathcal{F}_k}(\|x_k^\delta - x_k\| + \|x_k - x^*\|) + \delta$. Then it follows from (3.3) that

$$\lim_{\delta\to 0^+} \|x_{k+1}^\delta - x_{k+1}\| \le \lim_{\delta\to 0^+} \|x_k^\delta - x_k\|^2 + c\eta_k \lim_{\delta\to 0^+} \|F'_{i_k}(x_k^\delta)^* - F'_{i_k}(x_k)^*\|$$

$$+ c_F \lim_{\delta\to 0^+} \|(F_{i_k}(x_k^\delta) - y_{i_k}^\delta) - (F_{i_k}(x_k) - y_{i_k}^\dagger)\|.$$

Then the desired assertion follows from the continuity of the operators F_i and F_i' in Assumption 2.1(i), the induction hypothesis, and taking full expectation. ∎

Now we can prove Theorem 2.1 on the regularizing property of SGD.

Proof of Theorem 2.1. Let $\{\delta_n\}_{n\geq 1} \subset \mathbb{R}$ be a sequence converging to zero, and $y_n := y^{\delta_n}$ a corresponding sequence of noisy data. For each pair (δ_n, y_n), we denote by $k_n = k(\delta_n)$ the stopping index. Further, we may assume that k_n increases strictly monotonically with n. By Proposition 3.1 and Young's inequality $2ab \leq \epsilon a^2 + \epsilon^{-1} b^2$, with the choice $a = \mathbb{E}[\|F(x_k^\delta) - y^\delta\|^2]^{\frac{1}{2}}$, $b = (1+\eta)\delta$ and $\epsilon = 1 - 2\eta > 0$:

$$\begin{aligned}\mathbb{E}[\|x^* - x_{k+1}^\delta\|^2] - \mathbb{E}[\|x^* - x_k^\delta\|^2] \leq & -(1-2\eta)\eta_k \mathbb{E}[\|F(x_k^\delta) - y^\delta\|^2] \\ & + 2\eta_k(1+\eta)\delta \mathbb{E}[\|F(x_k^\delta) - y^\delta\|^2]^{\frac{1}{2}} \leq \frac{(1+\eta)^2}{1-2\eta}\eta_k \delta^2.\end{aligned}$$

Then for any $m < n$, summing the inequality with $\delta = \delta_n$ from k_m to $k_n - 1$ and applying triangle inequality lead to

$$\begin{aligned}\mathbb{E}[\|x_{k_n}^{\delta_n} - x^*\|^2] &\leq \mathbb{E}[\|x_{k_m}^{\delta_n} - x^*\|^2] + \frac{(1+\eta)^2}{1-2\eta}\delta_n^2 \sum_{j=k_m}^{k_n-1} \eta_j \\ &\leq 2\mathbb{E}[\|x_{k_m}^{\delta_n} - x_{k_m}\|^2] + 2\mathbb{E}[\|x_{k_m} - x^*\|^2] + \frac{(1+\eta)^2}{1-2\eta}\delta_n^2 \sum_{j=1}^{k_n-1} \eta_j.\end{aligned}$$

By Theorem 3.5, we can fix a large m so that the term $\mathbb{E}[\|x_{k_m} - x^*\|^2]$ is sufficiently small. Since the index k_m is fixed, we may apply Lemma 3.6 to conclude that the term $\mathbb{E}[\|x_{k_m}^{\delta_n} - x_{k_m}\|^2]$ tends to zero as $n \to \infty$. The last term also tends to zero under the condition $\lim_{n\to\infty} \delta_n^2 \sum_{i=1}^{k_n} \eta_i = 0$. This completes the proof of the first assertion. The case $\mathcal{N}(F'(x^\dagger)) \subset \mathcal{N}(F'(x))$ follows similarly as Theorem 3.5. ∎

4. Convergence rates. Now we prove convergence rates for SGD under Assumption 2.1, Assumption 2.2(ii) and Assumption 2.3; see Theorem 4.8 and Theorem 2.2 for the results for exact and noisy data, respectively. We employ some shorthand notation. Let

$$K_i = F_i'(x^\dagger), \quad K = \frac{1}{\sqrt{n}} \begin{pmatrix} K_1 \\ \vdots \\ K_n \end{pmatrix} \quad \text{and} \quad B = K^*K = \frac{1}{n}\sum_{i=1}^n K_i^* K_i.$$

Further, we frequently adopt the shorthand notation

$$\Pi_j^k(B) = \prod_{i=j}^k (I - \eta_i B), \tag{4.1}$$

with the convention $\Pi_j^k(B) = I$ for $j > k$, and for $s \geq 0$ and $j \in \mathbb{N}$, we define,

$$\tilde{s} = s + \tfrac{1}{2} \quad \text{and} \quad \phi_j^s = \|B^s \Pi_{j+1}^k(B)\|.$$

The rest of this section is organized as follows. By bias variance decomposition, we first derive two important recursions for the mean $\|B^s \mathbb{E}[e_k^\delta]\|$ and variance $\mathbb{E}[\|B^s(e_k^\delta - \mathbb{E}[e_k^\delta])\|^2]$, for any $s \geq 0$, in subsection 4.1 and subsection 4.2, respectively, and then use the recursions to derive convergence rates under *a priori* parameter choice in subsection 4.3.

4.1. Recursion on the bias. First, we derive a recursion on the bias of the SGD iterate x_k^δ. The following bound on the linearization error is useful.

LEMMA 4.1. *Under Assumption 2.1(iii), there holds*

$$\|F(x)-F(x^\dagger)-K(x-x^\dagger)\| \le \tfrac{c_R}{2}\|K(x-x^\dagger)\|\|x-x^\dagger\|.$$

Further, under Assumption 2.3, there holds

$$\mathbb{E}[\|F(x_k^\delta)-F(x^\dagger)-K(x_k^\delta-x^\dagger)\|^2]^{\frac12} \le \tfrac{c_R}{1+\theta}\mathbb{E}[\|K(x_k^\delta-x^\dagger)\|^2]^{\frac12}\mathbb{E}[\|x_k^\delta-x^\dagger\|^2]^{\frac{\theta}{2}}.$$

Proof. Let $z_t = tx+(1-t)x^\dagger$. By the mean value theorem and Assumption 2.1(iii),

$$\begin{aligned}\|F(x)-F(x^\dagger)-K(x-x^\dagger)\| &\le \|\int_0^1 (F'(z_t)-K)(x-x^\dagger)\mathrm{d}t\|\\ \le \int_0^1 \|(R_{z_t}-I)K(x-x^\dagger)\|\mathrm{d}t &\le \frac{c_R}{2}\|K(x-x^\dagger)\|\|x-x^\dagger\|.\end{aligned}$$

This shows the first estimate. Similarly, using Assumption 2.1(iii) and Assumption 2.3 with the choice $G(x)=K(x-x^\dagger)$, we obtain

$$\begin{aligned}\mathbb{E}[\|F(x_k^\delta)-F(x^\dagger)-K(x_k^\delta-x^\dagger)\|^2]^{\frac12} &\le \int_0^1 \mathbb{E}[\|(R_{z_t}-I)K(x_k^\delta-x^\dagger)\|^2]^{\frac12}\mathrm{d}t\\ \le c_R\mathbb{E}[\|K(x_k^\delta-x^\dagger)\|^2]^{\frac12}\int_0^1 \mathbb{E}[\|z_t-x^\dagger\|^2]^{\frac{\theta}{2}}\mathrm{d}t &\le \frac{c_R}{1+\theta}\mathbb{E}[\|K(x_k^\delta-x^\dagger)\|^2]^{\frac12}\mathbb{E}[\|x_k^\delta-x^\dagger\|^2]^{\frac{\theta}{2}}.\end{aligned}$$

This completes the proof of the lemma. ∎

The next result gives a useful representation of the mean $\mathbb{E}[e_k^\delta]$ of the error $e_k^\delta \equiv x_k^\delta - x^\dagger$.

LEMMA 4.2. *Under Assumption 2.1(iii), the error e_k^δ satisfies*

$$\mathbb{E}[e_{k+1}^\delta] = \Pi_1^k(B)e_1 + \sum_{j=1}^k \eta_j \Pi_{j+1}^k(B)K^*(-(y^\dagger-y^\delta)+\mathbb{E}[v_j]),$$

with the vector $v_k\in Y^n$ given by

$$v_k = -(F(x_k^\delta)-F(x^\dagger)-K(x_k^\delta-x^\dagger)) + (I-R_{x_k^\delta}^*)(F(x_k^\delta)-y^\delta). \tag{4.2}$$

Proof. The definition of the SGD iterate x_k^δ in (1.3) and the relation $F_{i_k}'(x_k^\delta)^* = (R_{x_k^\delta}^{i_k}F_{i_k}'(x^\dagger))^* = K_{i_k}^*R_{x_k^\delta}^{i_k*}$ from Assumption 2.1(iii) directly imply

$$e_{k+1}^\delta = e_k^\delta - \eta_k K_{i_k}^*K_{i_k}(x_k^\delta-x^\dagger) - \eta_k K_{i_k}^*(y_{i_k}^\dagger - y_{i_k}^\delta) + \eta_k K_{i_k}^* v_{k,i_k},$$

with the random variable $v_{k,i}$ defined by

$$v_{k,i} = -(F_i(x_k^\delta)-F_i(x^\dagger)-K_i(x_k^\delta-x^\dagger)) + (I-R_{x_k^\delta}^{i*})(F_i(x_k^\delta)-y_i^\delta). \tag{4.3}$$

Thus, by the measurability of x_k^δ (and thus e_k^δ) with respect to $\mathcal{F}_k$, $\mathbb{E}[e_{k+1}^\delta|\mathcal{F}_k]$ is given by

$$\mathbb{E}[e_{k+1}^\delta|\mathcal{F}_k] = (I-\eta_k B)e_k^\delta - \eta_k K^*(y^\dagger-y^\delta) + \eta_k K^* v_k.$$

Then taking full conditional and applying the recursion repeatedly complete the proof. ∎

Remark 4.1. *The term v_k in (4.2) includes both the linearization error $(F(x_k^\delta)-F(x^\dagger)-K(x_k^\delta-x^\dagger))$ of the nonlinear operator F and the range invariance of the derivative $F'(x)$ in Assumption 2.1(ii)–(iii).*

The next result gives a useful bound on $\mathbb{E}[v_j]$.

Lemma 4.3. *Under Assumption 2.1(i)–(iii), for v_j defined in (4.2), there holds*

$$\|\mathbb{E}[v_j]\| \le \tfrac{(3-\eta)c_R}{2(1-\eta)}\mathbb{E}[\|e_j^\delta\|^2]^{\frac12}\mathbb{E}[\|B^{\frac12}e_j^\delta\|^2]^{\frac12} + c_R\mathbb{E}[\|e_j^\delta\|^2]^{\frac12}\delta.$$

Proof. By the triangle inequality, there holds

$$\|\mathbb{E}[v_j]\| \le \|\mathbb{E}[F(x_j^\delta)-F(x^\dagger)-K(x_j^\delta-x^\dagger)]\| + \|\mathbb{E}[(I-R_{x_j^\delta}^*)(F(x_j^\delta)-y^\delta)]\| := \mathrm{I}+\mathrm{II}.$$

The bound on I follows from Lemma 4.1 and Cauchy-Schwarz inequality as

$$\mathrm{I} \le \tfrac{c_R}{2}\mathbb{E}[\|e_j^\delta\|\|Ke_j^\delta\|] \le \tfrac{c_R}{2}\mathbb{E}[\|e_j^\delta\|^2]^{\frac12}\mathbb{E}[\|Ke_j^\delta\|^2]^{\frac12}.$$

For the term II, by triangle inequality, Cauchy-Schwarz inequality and Lemma 3.1,

$$\begin{aligned}\mathrm{II} &:= \|\mathbb{E}[(I-R_{x_j^\delta})(y^\delta-F(x_j^\delta))]\| \le \mathbb{E}[\|(I-R_{x_j^\delta})(y^\delta-F(x_j^\delta))\|]\\ &\le \tfrac{c_R}{1-\eta}\mathbb{E}[\|e_j^\delta\|\|Ke_j^\delta\|] + c_R\mathbb{E}[\|e_j^\delta\|]\delta \le \mathbb{E}[\|e_j^\delta\|^2]^{\frac12}(\tfrac{c_R}{1-\eta}\mathbb{E}[\|Ke_j^\delta\|^2]^{\frac12}+c_R\delta).\end{aligned}$$

Combining these estimates with the identity $\|Ke_j^\delta\| = \|B^{\frac12}e_j^\delta\|$ gives the assertion. □

Last, we bound the error $\mathbb{E}[e_k^\delta]$ in a weighted norm. The cases $s=0$ and $s=\frac12$ will be employed in the convergence analysis.

Theorem 4.4. *Under Assumption 2.1, for any $s\ge 0$, there holds*

$$\|B^s\mathbb{E}[e_{k+1}^\delta]\| \le \phi_0^{s+\nu}\|w\| + \sum_{j=1}^k \eta_j\phi_j^{\tilde s}\Big(\tfrac{(3-\eta)c_R}{2(1-\eta)}\mathbb{E}[\|e_j^\delta\|^2]^{\frac12}\mathbb{E}[\|B^{\frac12}e_j^\delta\|^2]^{\frac12} + c_R\mathbb{E}[\|e_j^\delta\|^2]^{\frac12}\delta+\delta\Big).$$

Proof. By Lemma 4.2 and triangle inequality,

$$\|B^s\mathbb{E}[e_{k+1}^\delta]\| \le \mathrm{I} + \sum_{j=1}^k \eta_j\mathrm{II}_j.$$

with $\mathrm{I} = \|B^s\Pi_1^k(B)(x_1-x^\dagger)\|$ and $\mathrm{II}_j = \|B^s\Pi_{j+1}^k(B)K^*(\mathbb{E}[v_j]-(y^\dagger-y^\delta))\|$. It suffices to bound the terms I and II_j. By Assumption 2.1(iv),

$$\mathrm{I} = \|B^s\Pi_1^k(B)B^\nu w\| \le \|\Pi_1^k(B)B^{s+\nu}\|\|w\|.$$

To bound the terms II_j, we have

$$\mathrm{II}_j \le \|B^s\Pi_{j+1}^k(B)K^*(\mathbb{E}[v_j]-(y^\dagger-y^\delta))\| \le \|B^{s+\frac12}\Pi_{j+1}^k(B)\|(\|\mathbb{E}[v_j]\|+\delta).$$

This, Lemma 4.3 and the notation ϕ_j^s complete the proof. □

Remark 4.2. *The bound on $\mathbb{E}[e_k^\delta]$ depends on the variance of the iterate x_k^δ (via the terms like $\mathbb{E}[\|e_k^\delta\|^2]$ etc.), which differs from the linear case [14]. This is one of the complications for nonlinear inverse problems. The weighted norm $\|B^s\mathbb{E}[e_k^\delta]\|$ is useful since the upper bound in Theorem 4.4 involves $\mathbb{E}[\|B^{\frac12}e_k^\delta\|^2]$, i.e., $s=\frac12$. For linear inverse problems, $R_x = I$ and $c_R=0$, and the recursion simplifies to $\|B^s\mathbb{E}[e_{k+1}^\delta]\| \le \phi_0^{s+\nu}\|w\| + \sum_{j=1}^k\eta_j\phi_j^{\tilde s}\delta$, i.e., the approximation error and data error, respectively.*

4.2. Recursion on variance. Now we turn to the computational variance $\mathbb{E}[\|B^s(x_k^\delta - \mathbb{E}[x_k^\delta])\|^2]$, which arises from the random index i_k. First, we bound on the variance in terms of iteration noises $N_{j,1}$ and $N_{j,2}$ (defined in (4.4) below).

LEMMA 4.5. *Under Assumption 2.1(iii), for the SGD iterate x_k^δ, there holds*

$$\begin{aligned}\mathbb{E}[\|B^s(x_{k+1}^\delta - \mathbb{E}[x_{k+1}^\delta])\|^2] \leq \sum_{j=1}^k \eta_j^2(\phi_j^{\tilde{s}})^2\mathbb{E}[\|N_{j,1}\|^2] + 2\sum_{i=1}^k\sum_{j=i}^k \eta_i\eta_j\phi_i^{\tilde{s}}\phi_j^{\tilde{s}}\mathbb{E}[\|N_{i,1}\|\|N_{j,2}\|]\\ + \sum_{i=1}^k\sum_{j=1}^k \eta_i\eta_j\phi_i^{\tilde{s}}\phi_j^{\tilde{s}}\mathbb{E}[\|N_{i,2}\|\|N_{j,2}\|],\end{aligned}$$

with the random variables $N_{j,1}$ and $N_{j,2}$ respectively given by

$$\begin{aligned}N_{j,1} &= (K(x_j^\delta - x^\dagger) - K_{i_j}(x_j^\delta - x^\dagger)\varphi_{i_j}) + ((y^\dagger - y^\delta) - (y_i^\dagger - y_i^\delta)\varphi_{i_j}),\\ N_{j,2} &= -\mathbb{E}[v_j] + v_{j,i_j}\varphi_{i_j},\end{aligned} \tag{4.4}$$

where v_k and $v_{k,i}$ are given in (4.2) and (4.3), and $\varphi_i = (0,\ldots,0,n^{\frac{1}{2}},0,\ldots,0)$ denotes the canonical ith Cartesian basis vector in $\mathbb{R}^n$ scaled by $n^{\frac{1}{2}}$.

Proof. Similar to the proof of Lemma 4.2, we rewrite the SGD iteration (1.3) as

$$x_{k+1}^\delta = x_k^\delta - \eta_k K_{i_k}^* K_{i_k}(x_k^\delta - x^\dagger) - \eta_k K_{i_k}^*(y_{i_k}^\dagger - y_{i_k}^\delta) + \eta_k K_{i_k}^* v_{k,i_k}, \tag{4.5}$$

with $v_{k,i}$ defined in (4.3). By the definition of v_k in (4.2) and the measurability of x_k^δ with respect to $\mathcal{F}_k$, we obtain

$$\mathbb{E}[x_{k+1}^\delta|\mathcal{F}_k] = x_k^\delta - \eta_k B(x_k^\delta - x^\dagger) - \eta_k K^*(y^\dagger - y^\delta) + \eta_k K^* v_k.$$

Taking full conditional yields

$$\mathbb{E}[x_{k+1}^\delta] = \mathbb{E}[x_k^\delta] - \eta_k B\mathbb{E}[x_k^\delta - x^\dagger] - \eta_k K^*(y^\dagger - y^\delta) + \eta_k K^*\mathbb{E}[v_k]. \tag{4.6}$$

Thus, subtracting (4.6) from (4.5) shows that $z_k := x_k^\delta - \mathbb{E}[x_k^\delta]$ satisfies

$$z_{k+1} = (I - \eta_k B)z_k + \eta_k M_k, \tag{4.7}$$

with $z_1 = 0$ and the iteration noise M_j given by $M_j = M_{j,1} + M_{j,2}$, where

$$\begin{aligned}M_{j,1} &= (B(x_j^\delta - x^\dagger) - K_{i_j}^* K_{i_j}(x_j^\delta - x^\dagger)) + (K^*(y^\dagger - y^\delta) - K_{i_j}^*(y_{i_j}^\dagger - y_{i_j}^\delta)),\\ M_{j,2} &= -(K^*\mathbb{E}[v_j] - K_{i_j}^* v_{j,i_j}).\end{aligned}$$

Repeatedly applying the recursion (4.7) with $z_1 = 0$ leads to

$$z_{k+1} = \sum_{j=1}^k \eta_j \Pi_{j+1}^k(B) M_j.$$

With the decomposition of $M_j = M_{j,1} + M_{j,2}$, we directly obtain

$$\mathbb{E}[\|B^s z_{k+1}\|^2] = \sum_{i=1}^k\sum_{j=1}^k \eta_i\eta_j\mathbb{E}[\langle B^s\Pi_{i+1}^k(B)M_{i,1}, B^s\Pi_{j+1}^k(B)M_{j,1}\rangle]$$

$$+2\sum_{i=1}^{k}\sum_{j=1}^{k}\eta_i\eta_j\mathbb{E}[\langle B^s\Pi_{i+1}^k(B)M_{i,1},B^s\Pi_{j+1}^k(B)M_{j,2}\rangle]$$

$$+\sum_{i=1}^{k}\sum_{j=1}^{k}\eta_i\eta_j\mathbb{E}[\langle B^s\Pi_{i+1}^k(B)M_{i,2},B^s\Pi_{j+1}^k(B)M_{j,2}\rangle]:=\mathrm{I}+\mathrm{II}+\mathrm{III}.$$

Below we simplify the three terms. Since x_j^δ is measurable with respect to $\mathcal{F}_j$, we have $\mathbb{E}[M_{j,1}|\mathcal{F}_j]=0$, which directly implies the independence $\mathbb{E}[\langle B^sM_{i,1},B^sM_{j,1}\rangle]=0$, $i\neq j$. Indeed, for $i>j$, $\mathbb{E}[\langle B^sM_{i,1},B^sM_{j,1}\rangle|\mathcal{F}_i]=\langle B^s\mathbb{E}[M_{i,1}|\mathcal{F}_i],B^sM_{j,1}\rangle=0$, and taking full conditional yields the claim. Thus, the term I simplifies to

$$\mathrm{I}=\sum_{j=1}^{k}\eta_j^2\mathbb{E}[\|B^s\Pi_{j+1}^k(B)M_{j,1}\|^2].$$

Further, for $i>j$, a similar argument yields $\mathbb{E}[\langle B^sM_{i,1},B^sM_{j,2}\rangle]=0$ and thus

$$\mathrm{II}=2\sum_{i=1}^{k}\sum_{j=i}^{k}\eta_i\eta_j\mathbb{E}[\langle B^s\Pi_{i+1}^kM_{i,1},B^s\Pi_{j+1}^kM_{j,2}\rangle].$$

Now we further simplify $M_{j,1}$ and $M_{j,2}$. By the definitions of $N_{j,1}$ and $N_{j,2}$, with $(K^*)^\dagger$ being the pseudoinverse of K^*, we have $(K^*)^\dagger M_j=N_{j,1}+N_{j,2}$. Thus, by triangle inequality,

$$\mathbb{E}[\|B^sz_{k+1}\|^2]\leq\sum_{j=1}^{k}\eta_j^2\mathbb{E}[\|B^{s+\frac12}\Pi_{j+1}^k(B)\|^2\|N_{j,1}\|^2]$$

$$+2\sum_{i=1}^{k}\sum_{j=i}^{k}\eta_i\eta_j\|B^{s+\frac12}\Pi_{i+1}^k(B)\|\|B^{s+\frac12}\Pi_{j+1}^k(B)\|\mathbb{E}[\|N_{i,1}\|\|N_{j,2}\|]$$

$$+\sum_{i=1}^{k}\sum_{j=1}^{k}\eta_i\eta_j\|B^{s+\frac12}\Pi_{i+1}^k(B)\|\|B^{s+\frac12}\Pi_{j+1}^k(B)\|\mathbb{E}[\|N_{i,2}\|\|N_{j,2}\|].$$

This completes the proof of the lemma. ∎

The next result bounds the iteration noises $N_{j,1}$ and $N_{j,2}$.

Lemma 4.6. *Under Assumption 2.1(i)–(iii) and Assumption 2.3, for $N_{j,1}$ and $N_{j,2}$ defined in (4.4), there hold*

$$\mathbb{E}[\|N_{j,1}\|^2]^{\frac12}\leq n^{\frac12}(\mathbb{E}[\|B^{\frac12}e_j^\delta\|^2]^{\frac12}+\delta),\tag{4.8}$$

$$\mathbb{E}[\|N_{j,2}\|^2]^{\frac12}\leq n^{\frac12}\big(\tfrac{c_R(2+\theta-\eta)}{(1+\theta)(1-\eta)}\mathbb{E}[\|B^{\frac12}e_j^\delta\|^2]^{\frac12}+c_R\delta\big)\mathbb{E}[\|e_j^\delta\|^2]^{\frac{\theta}{2}}.\tag{4.9}$$

Proof. By the measurability of x_j^δ with respect to $\mathcal{F}_j$, we have $\mathbb{E}[K_{i_j}(x_j^\delta-x^\dagger)\varphi_{i_j}|\mathcal{F}_j]=K(x_j^\delta-x^\dagger)$. Then by bias-variance decomposition, we have

$$\mathbb{E}[\|(K(x_j^\delta-x^\dagger)-K_{i_j}(x_j^\delta-x^\dagger)\varphi_{i_j})\|^2|\mathcal{F}_j]\leq\mathbb{E}[\|K_{i_j}(x_j^\delta-x^\dagger)\varphi_{i_j}\|^2|\mathcal{F}_j]$$

$$=n^{-1}\sum_{i=1}^{n}\|K_i(x_j^\delta-x^\dagger)\|^2n=n\|K(x_j^\delta-x^\dagger)\|^2,$$

and then by taking full expectation, we obtain

$$\mathbb{E}[\|(K(x_j^\delta - x^\dagger) - K_{i_j}(x_j^\delta - x^\dagger)\varphi_{i_j})\|^2]^{\frac12} \le n^{\frac12}\mathbb{E}[\|K(x_j^\delta - x^\dagger)\|^2]^{\frac12}.$$

Similarly, $\mathbb{E}[\|(y^\dagger - y^\delta) - (y_{i_j}^\dagger - y_{i_j}^\delta)\varphi_{i_j}\|^2]^{\frac12} \le n^{\frac12}\delta$. This and triangle inequality show the estimate (4.8). Similarly, by the measurability of x_j^δ with respect to $\mathcal{F}_j$ and bias variance decomposition, we deduce (with $\mathbb{E}_{\mathcal{F}_j}$ denoting taking expectation in $\mathcal{F}_j$)

$$\mathbb{E}[\|(\mathbb{E}[v_j] - v_{j,i_j}\varphi_{i_j})\|^2] \le \mathbb{E}_{\mathcal{F}_j}[\mathbb{E}[\|v_{j,i_j}\varphi_{i_j}\|^2|\mathcal{F}_j]] = n\mathbb{E}[\|v_j\|^2],$$

i.e., $\mathbb{E}[\|(\mathbb{E}[v_j] - v_{j,i_j}\varphi_{i_j})\|^2]^{\frac12} \le n^{\frac12}\mathbb{E}[\|v_j\|^2]^{\frac12}$. Then by triangle inequality, Assumption 2.3 and Lemma 4.1,

$$\begin{aligned}\mathbb{E}[\|v_j\|^2]^{\frac12} &\le \mathbb{E}[\|(F(x_j^\delta) - F(x^\dagger) - K(x_j^\delta - x^\dagger))\|^2]^{\frac12} + \mathbb{E}[\|(I - R_{x_j^\delta}^*)(F(x_j^\delta) - y^\delta)\|^2]^{\frac12}\\
&\le \tfrac{c_R}{1+\theta}\mathbb{E}[\|Ke_j^\delta\|^2]^{\frac12}\mathbb{E}[\|e_j^\delta\|^2]^{\frac{\theta}{2}} + c_R(\tfrac{1}{1-\eta}\mathbb{E}[\|Ke_j^\delta\|^2]^{\frac12} + \delta)\mathbb{E}[\|e_j^\delta\|^2]^{\frac{\theta}{2}}\\
&= (\tfrac{(2+\theta-\eta)c_R}{(1+\theta)(1-\eta)}\mathbb{E}[\|Ke_j^\delta\|^2]^{\frac12} + c_R\delta)\mathbb{E}[\|e_j^\delta\|^2]^{\frac{\theta}{2}}.\end{aligned}$$

This completes the proof of the lemma. ∎

REMARK 4.3. *Note that the convergence analysis in [14] relies on the independence $\mathbb{E}[\langle B^s M_j, B^s M_\ell\rangle] = 0$ for $j \neq \ell$. This identity is no longer valid for nonlinear inverse problems, although it still holds for the linear part $M_{j,1}$: $\mathbb{E}[\langle B^s M_{j,1}, B^s M_{\ell,1}\rangle] = 0$ for $j \neq \ell$. The conditional dependence among the iteration noises $M_{j,2}$ poses one big challenge to the convergence analysis, and the splitting of the conditionally dependent and independent components will plays a role in the analysis below. Assumption 2.3 is to compensate the conditional dependence.*

REMARK 4.4. *The constants in Lemma 4.6 involve an unpleasant dependence on n as $n^{\frac12}$, due to the variance inflation of the estimated gradient. It can be reduced by various strategies, e.g., mini-batch or variance reduction.*

Last, we give a bound on the variance $\mathbb{E}[\|B^s(x_k^\delta - \mathbb{E}[x_k^\delta])\|^2]$. This result will play an important role in the error analysis in subsection 4.3.

THEOREM 4.7. *Let Assumption 2.1(i)–(iii) and Assumption 2.3 be fulfilled. Then for any $s \in [0, \frac12]$, there holds*

$$\begin{aligned}\mathbb{E}[\|B^s(\mathbb{E}[x_{k+1}^\delta] - x_{k+1}^\delta)\|^2] &\le n\sum_{j=1}^k \eta_j^2(\phi_j^{\tilde s})^2(\mathbb{E}[\|B^{\frac12}e_j^\delta\|^2]^{\frac12} + \delta)^2\\
&+ 2n\sum_{i=1}^k\sum_{j=i}^k \eta_i\eta_j\phi_i^{\tilde s}\phi_j^{\tilde s}(\mathbb{E}[\|B^{\frac12}e_i^\delta\|^2]^{\frac12} + \delta)(\tfrac{(2+\theta-\eta)c_R}{(1+\theta)(1-\eta)}\mathbb{E}[\|B^{\frac12}e_j^\delta\|^2]^{\frac12} + c_R\delta)\mathbb{E}[\|e_j^\delta\|^2]^{\frac{\theta}{2}}\\
&+ n\Big(\sum_{j=1}^k \eta_j\phi_j^{\tilde s}(\tfrac{(2+\theta-\eta)c_R}{(1+\theta)(1-\eta)}\mathbb{E}[\|B^{\frac12}e_j^\delta\|^2]^{\frac12} + c_R\delta)\mathbb{E}[\|e_j^\delta\|^2]^{\frac{\theta}{2}}\Big)^2.\end{aligned}$$

Proof. The assertion follows directly from Lemma 4.5 and Lemma 4.6. ∎

4.3. Convergence rates. This part is devoted to convergence rates analysis of SGD under Assumption 2.1(ii). We analyze the cases of exact and noisy data separately. For exact data, the bounds involve constants that are more transparent in terms of their dependence on various algorithmic parameters. First we analyze the case of exact data $y^\dagger$, and the bound boils down to the approximation error and computational variance. Further, we assume that $\|B\| \leq 1$ and $\eta_0 \leq 1$ below, which can be easily achieved by rescaling the operator F and the data $y^\dagger/y^\delta$. The analysis relies heavily on various technical estimates in Appendix A, especially Proposition A.1.

THEOREM 4.8. *Let Assumption 2.1, Assumption 2.2(ii) and Assumption 2.3 be fulfilled with $\|w\|$, θ and η_0 being sufficiently small. Then the error $e_k = x_k - x^\dagger$ satisfies*

$$\mathbb{E}[\|e_k\|^2] \leq c^* \|w\|^2 k^{-\min(2\nu(1-\alpha),\alpha-\epsilon)}, \quad \mathbb{E}[\|B^{\frac{1}{2}} e_k\|^2] \leq c^* \|w\|^2 k^{-\min((1+2\nu)(1-\alpha),1-\epsilon)},$$

where $\epsilon \in (0, \frac{\alpha}{2})$ is small, and c^* is independent of k, but depends on α, ν, η_0, n, and θ.

Proof. For any $s \geq 0$, Theorem 4.4 and Theorem 4.7 give (with $c_0 = \frac{(2+\theta-\eta)c_R}{(1+\theta)(1-\eta)}$)

$$\begin{aligned}
\mathbb{E}[\|B^s e_{k+1}\|^2] \leq & \Big(c_0 \sum_{j=1}^k \eta_j \phi_j^{\tilde{s}} \mathbb{E}[\|e_j\|^2]^{\frac{1}{2}} \mathbb{E}[\|B^{\frac{1}{2}} e_j\|^2]^{\frac{1}{2}} + \phi_0^{s+\nu} \|w\| \Big)^2 \\
& + 2nc_0 \Big(\sum_{i=1}^k \eta_i \phi_i^{\tilde{s}} \mathbb{E}[\|B^{\frac{1}{2}} e_i\|^2]^{\frac{1}{2}} \Big) \Big(\sum_{j=1}^k \eta_j \phi_j^{\tilde{s}} \mathbb{E}[\|B^{\frac{1}{2}} e_j\|^2]^{\frac{1}{2}} \mathbb{E}[\|e_j\|^2]^{\frac{\theta}{2}} \Big) \qquad (4.10) \\
& + nc_0^2 \Big(\sum_{j=1}^k \eta_j \phi_j^{\tilde{s}} \mathbb{E}[\|B^{\frac{1}{2}} e_j\|^2]^{\frac{1}{2}} \mathbb{E}[\|e_j\|^2]^{\frac{\theta}{2}} \Big)^2 + n \sum_{j=1}^k \eta_j^2 (\phi_j^{\tilde{s}})^2 \mathbb{E}[\|B^{\frac{1}{2}} e_j\|^2].
\end{aligned}$$

Under Assumption 2.2(ii), Lemma A.1 and Lemma A.2 directly give

$$\phi_0^{s+\nu} \leq \frac{(s+\nu)^{s+\nu}}{e^{s+\nu} (\sum_{i=1}^k \eta_i)^{s+\nu}} \leq \frac{(s+\nu)^{s+\nu} (1-\alpha)^{\nu+s}}{e^{s+\nu} \eta_0^{\nu+s} (1-2^{\alpha-1})^{\nu+s}} (k+1)^{-(1-\alpha)(\nu+s)}.$$

Note that the function $\frac{s^s}{e^s}$ is decreasing in s over the interval $[0,1]$, and the function $\frac{1-\alpha}{1-2^{\alpha-1}}$ is decreasing in α over the interval $[0,1]$ (and upper bounded by 2). Thus, for $\eta_0 \leq 1$ and any $0 \leq \nu, s \leq \frac{1}{2}$, there holds (with $c_\nu = \frac{2\nu^\nu}{\eta_0 e^\nu}$)

$$\phi_0^{s+\nu} \leq c_\nu (k+1)^{-(\nu+s)(1-\alpha)}. \qquad (4.11)$$

Let $a_j \equiv \mathbb{E}[\|e_j\|^2]$ and $b_j \equiv \mathbb{E}[\|B^{\frac{1}{2}} e_j\|^2]$. Since $\|B\| \leq 1$, we have $\phi_j^s \leq \phi_j^{\bar{s}}$ for any $0 \leq \bar{s} \leq s$. Then setting $s = 0$ and $s = 1/2$ in the recursion (4.10) and applying (4.11) lead to

$$\begin{aligned}
a_{k+1} \leq & \Big(c_0 \sum_{j=1}^k \eta_j \phi_j^{\frac{1}{2}} a_j^{\frac{1}{2}} b_j^{\frac{1}{2}} + c_\nu \|w\| (k+1)^{-\nu(1-\alpha)} \Big)^2 + n \sum_{j=1}^k \eta_j^2 (\phi_j^{\frac{1}{2}})^2 b_j \\
& + 2nc_0 \Big(\sum_{i=1}^k \eta_i \phi_i^{\frac{1}{2}} b_i^{\frac{1}{2}} \Big) \Big(\sum_{j=1}^k \eta_j \phi_j^{\frac{1}{2}} b_j^{\frac{1}{2}} a_j^{\frac{\theta}{2}} \Big) + nc_0^2 \Big(\sum_{j=1}^k \eta_j \phi_j^{\frac{1}{2}} b_j^{\frac{1}{2}} a_j^{\frac{\theta}{2}} \Big)^2, \qquad (4.12)
\end{aligned}$$

$$\begin{aligned}
b_{k+1} \leq & \Big(c_0 \sum_{j=1}^k \eta_j \phi_j^1 a_j^{\frac{1}{2}} b_j^{\frac{1}{2}} + c_\nu \|w\| (k+1)^{-(\frac{1}{2}+\nu)(1-\alpha)} \Big)^2 + n \Big(\sum_{j=1}^{[\frac{k}{2}]} \eta_j^2 (\phi_j^r)^2 b_j \\
& + \sum_{j=[\frac{k}{2}]+1}^k \eta_j^2 (\phi_j^{\frac{1}{2}})^2 b_j \Big) + 2nc_0 \Big(\sum_{i=1}^k \eta_i \phi_i^1 b_i^{\frac{1}{2}} \Big) \Big(\sum_{j=1}^k \eta_j \phi_j^1 b_j^{\frac{1}{2}} a_j^{\frac{\theta}{2}} \Big) + nc_0^2 \Big(\sum_{j=1}^k \eta_j \phi_j^1 b_j^{\frac{1}{2}} a_j^{\frac{\theta}{2}} \Big)^2, \qquad (4.13)
\end{aligned}$$

with $r = \min(\frac{1}{2}+\nu, \frac{1-\epsilon}{2(1-\alpha)}) \in (\frac{1}{2}, 1)$. The rest of the proof is to prove

$$a_k \leq c^* \|w\|^2 k^{-\beta} \quad \text{and} \quad b_k \leq c^* \|w\|^2 k^{-\gamma}. \qquad (4.14)$$

where $\beta=\min(2\nu(1-\alpha),\alpha-\epsilon)$ and $\gamma=\min((1+2\nu)(1-\alpha),1-\epsilon)$, and $c^*>0$ is to be specified. The proof is based on mathematical induction. When $k=1$, (4.14) holds trivially for any large c^*. Now we assume that (4.14) holds up to the case k, and prove it for the case $k+1$. Actually, it follows from (4.12) and the induction hypothesis that (with $\varrho=c^*\|w\|^2$)

$$\begin{aligned}a_{k+1}\leq&\Big(c_0\varrho\sum_{j=1}^k\eta_j\phi_j^{\frac12}j^{-\frac{\beta+\gamma}{2}}+c_\nu\|w\|(k+1)^{-\nu(1-\alpha)}\Big)^2+n\varrho\sum_{j=1}^k\eta_j^2(\phi_j^{\frac12})^2j^{-\gamma}\\ &+2nc_0\varrho^{1+\frac\theta2}\Big(\sum_{i=1}^k\eta_i\phi_i^{\frac12}i^{-\frac\gamma2}\Big)\Big(\sum_{j=1}^k\eta_j\phi_j^{\frac12}j^{-\frac{\gamma+\theta\beta}{2}}\Big)+nc_0^2\varrho^{1+\theta}\Big(\sum_{j=1}^k\eta_j\phi_j^{\frac12}j^{-\frac{\gamma+\beta\theta}{2}}\Big)^2.\end{aligned}\tag{4.15}$$

Next we bound the terms on the right-hand side. By Proposition A.1, we have

$$\sum_{j=1}^k\eta_j\phi_j^{\frac12}j^{-\frac\gamma2}\leq c_1(k+1)^{-\frac\beta2}\quad\text{and}\quad\sum_{j=1}^k\eta_j^2(\phi_j^{\frac12})^2j^{-\gamma}\leq c_2(k+1)^{-\beta},$$

with $c_1=2^{\frac\beta2}\eta_0^{\frac12}(2^{-1}B(\frac12,\zeta)+1)$, $\zeta=(\frac12-\nu)(1-\alpha)>0$, and $c_2=2^\beta\eta_0(\alpha^{-1}+2)$. Then we derive from (4.15) that

$$a_{k+1}\leq\big((c_0c_1\varrho+c_\nu\|w\|)^2+nc_2\varrho+2nc_0c_1^2\varrho^{1+\frac\theta2}+nc_0^2c_1^2\varrho^{1+\theta}\big)(k+1)^{-\beta}.\tag{4.16}$$

Next we bound b_k similarly. It follows from (4.13) (with $r=\min(\frac12+\nu,\frac{1-\epsilon}{2(1-\alpha)})\in(\frac12,1)$) and the induction hypothesis that

$$\begin{aligned}b_{k+1}\leq&\Big(c_0\varrho\sum_{j=1}^k\eta_j\phi_j^1j^{-\frac{\beta+\gamma}{2}}+c_\nu\|w\|(k+1)^{-(\frac12+\nu)(1-\alpha)}\Big)^2\\ &+n\varrho\Big(\sum_{j=1}^{[\frac k2]}\eta_j^2(\phi_j^r)^2j^{-\gamma}+\sum_{j=[\frac k2]+1}^k\eta_j^2(\phi_j^{\frac12})^2j^{-\gamma}\Big)\\ &+2nc_0\varrho^{1+\frac\theta2}\Big(\sum_{i=1}^k\eta_i\phi_i^1i^{-\frac\gamma2}\Big)\Big(\sum_{j=1}^k\eta_j\phi_j^1j^{-\frac{\gamma+\theta\beta}{2}}\Big)+nc_0^2\varrho^{1+\theta}\Big(\sum_{j=1}^k\eta_j\phi_j^1j^{-\frac{\gamma+\theta\beta}{2}}\Big)^2.\end{aligned}\tag{4.17}$$

By Proposition A.1, there hold

$$\begin{aligned}&\sum_{j=1}^k\eta_j\phi_j^1j^{-\frac{\beta+\gamma}{2}}\leq c_1'(k+1)^{-\frac\gamma2},\ \sum_{j=1}^{[\frac k2]}\eta_j^2(\phi_j^r)^2j^{-\gamma}+\sum_{j=[\frac k2]+1}^k\eta_j^2(\phi_j^{\frac12})^2j^{-\gamma}\leq c_2'(k+1)^{-\gamma},\\ &\Big(\sum_{i=1}^k\eta_i\phi_i^1i^{-\frac\gamma2}\Big)\Big(\sum_{j=1}^k\eta_j\phi_j^1j^{-\frac{\gamma+\theta\beta}{2}}\Big)\leq c_3'^2(k+1)^{-\gamma},\ \sum_{j=1}^k\eta_j\phi_j^1j^{-\frac{\gamma+\theta\beta}{2}}\leq c_4'(k+1)^{-\frac\gamma2},\end{aligned}$$

with $c_1'=2^{\frac\gamma2}(\zeta^{-1}+2\beta^{-1}+1)$, $c_2'=2^\gamma\eta_0^{2-2r}(3\alpha^{-1}+1)$, $c_3'=2^{\frac\gamma2}(((\frac12-\nu-\theta\nu)(1-\alpha))^{-1}+4(\theta\beta)^{-1}+1)$ and $c_4'=2^{\frac\gamma2}(\zeta^{-1}+2(\theta\beta)^{-1}+1)$. These estimates and (4.17) yield

$$b_{k+1}\leq((c_0c_1'\varrho+c_\nu\|w\|)^2+nc_2'\varrho+2nc_0c_3'^2\varrho^{1+\frac\theta2}+nc_0^2c_4'^2\varrho^{1+\theta})(k+1)^{-\gamma}.\tag{4.18}$$

In view of (4.16) and (4.18), upon dividing by ϱ, assertion (4.14) holds if we can show the existence of a $c^*>0$ such that

$$(c_0c_1\varrho^{\frac12}+c_\nu c^{*-\frac12})^2+nc_2+2nc_0c_1^2\varrho^{\frac{\theta}{2}}+nc_0^2c_1^2\varrho^\theta\leq 1,$$
$$(c_0c_1'\varrho^{\frac12}+c_\nu c^{*-\frac12})^2+nc_2'+2nc_0c_3'^2\varrho^{\frac{\theta}{2}}+nc_0^2c_4'^2\varrho^\theta\leq 1.$$

Since the constants c_2 and c_2' are proportional to η_0 and η_0^{2-2r} (with the exponent $1>2-2r>0$), respectively, for sufficiently small η_0, there holds $n\max(c_2,c_2')<1$. Now for sufficiently small $\|w\|$ and large c^* such that ρ is small, the above two inequalities hold. This completes the induction step and the proof of the theorem. ∎

REMARK 4.5. *$\mathbb{E}[\|B^{\frac12}e_k\|^2]$ decays as $\mathbb{E}[\|B^{\frac12}e_k\|^2]\leq ck^{-\min((1+2\nu)(1-\alpha),1-\epsilon)}$, which, for α close to unit, is comparable with that for the Landweber method [8]: $\|B^{\frac12}e_k\|\leq ck^{-(\nu+\frac12)(1-\alpha)}$. The factor $k^{-(1-\epsilon)}$ limits the fastest possible rate. This restriction arises from the computational variance, due to the random selection of the row index i_k, which limits the convergence rate $\mathbb{E}[\|e_k\|^2]$ to $O(k^{-\min(2\nu(1-\alpha),\alpha-\epsilon)})$. Thus for order optimality, the largest possible smoothness index is $\nu=\frac12$, beyond which SGD suffers from suboptimality, similar to the Landweber method for nonlinear inverse problems [8]. Further, it shows the impact of the exponent α: a smaller α may restrict the error $\mathbb{E}[\|e_k\|^2]$ to $O(k^{-(\alpha-\epsilon)})$.*

REMARK 4.6. *The exponent α in the step size schedule in Assumption 2.2(ii) enters into the constant c^* via the constants $c_1,\ldots,c_4'$ etc, and the constant c_0 is independent of α. The constants $c_1,\ldots,c_4'$ blow up either like $(1-\alpha)^{-1}$ as $\alpha\to1^-$, according to the well-known asymptotic behavior of the Beta function, or like α^{-1} as $\alpha\to0^+$. These dependencies partly exhibit the delicacy of choosing a proper step size schedule for SGD.*

REMARK 4.7. *We briefly comment on the "smallness" conditions on w, η_0 and θ in the analysis. The smallness assumption on w in the source condition in Assumption 2.1(iv) appears also for the classical Landweber method [8] and the standard Tikhonov regularization [5, 11], and thus it is not surprising. The smallness condition on η_0 is to control the influence of the computational variance, and in a slightly different context of statistical learning theory, similar conditions also appear in the convergence analysis of variants of SGD. The smallness condition on θ is only to facilitate the analysis, i.e., a concise form of the constant c_3', and the assumption can be removed at the expense of a less transparent (but more benign) expression for c_3'; see the proof in Proposition A.1 and also Remark A.1.*

Last, we prove the main result in this work, i.e., Theorem 2.2, which gives the convergence rate of SGD (1.3) for noisy data y^δ.

Proof of Theorem 2.2. The main proof strategy is similar to that of Theorem 4.8. Let $a_j\equiv\mathbb{E}[\|e_j^\delta\|^2]$ and $b_j\equiv\mathbb{E}[\|B^{\frac12}e_j^\delta\|^2]$. Then with $c_0=\frac{(2+\theta-\eta)c_R}{(1+\theta)(1-\eta)}$, repeating the argument of Theorem 4.8 leads to

$$\begin{aligned}a_{k+1}\leq&\Big(\sum_{j=1}^k\eta_j\phi_j^{\frac12}\big(c_0a_j^{\frac12}b_j^{\frac12}+c_Ra_j^{\frac12}\delta+\delta\big)+c_\nu\|w\|(k+1)^{-\nu(1-\alpha)}\Big)^2\\
&+n\sum_{j=1}^k\eta_j^2(\phi_j^{\frac12})^2(b_j^{\frac12}+\delta)^2+n\Big(\sum_{j=1}^k\eta_j\phi_j^{\frac12}(c_0b_j^{\frac12}+c_R\delta)a_j^{\frac{\theta}{2}}\Big)^2\\
&+2n\Big(\sum_{i=1}^k\eta_i\phi_i^{\frac12}(b_i^{\frac12}+\delta)\Big)\Big(\sum_{j=1}^k\eta_j\phi_j^{\frac12}(c_0b_j^{\frac12}+c_R\delta)a_j^{\frac{\theta}{2}}\Big),\end{aligned}$$

$$b_{k+1} \leq \Big(\sum_{j=1}^k \eta_j \phi_j^1 \big(c_0 a_j^{\frac12} b_j^{\frac12} + c_R a_j^{\frac12} \delta + \delta \big) + c_\nu \|w\| (k+1)^{-(\nu+\frac12)(1-\alpha)} \Big)^2$$
$$+ n \sum_{j=1}^k \eta_j^2 (\phi_j^1)^2 (b_j^{\frac12} + \delta)^2 + n \Big(\sum_{j=1}^k \eta_j \phi_j^1 (c_0 b_j^{\frac12} + c_R \delta) a_j^{\frac{\theta}{2}} \Big)^2$$
$$+ 2n \Big(\sum_{i=1}^k \eta_i \phi_i^1 (b_i^{\frac12} + \delta) \Big) \Big(\sum_{j=1}^k \eta_j \phi_j^1 (c_0 b_j^{\frac12} + c_R \delta) a_j^{\frac{\theta}{2}} \Big).$$

Like in the proof of Theorem 4.8, the goal is to show

$$a_k \leq c^* \|w\|^2 k^{-\beta} \quad \text{and} \quad b_k \leq c^* \|w\|^2 k^{-\gamma}, \tag{4.19}$$

for all $k \leq k^* = [(\frac{\delta}{\|w\|})^{-\frac{2}{(2\nu+1)(1-\alpha)}}]$, with $\beta = \min(2\nu(1-\alpha), \alpha - \epsilon)$ and $\gamma = \min((1+2\nu)(1-\alpha), 1-\epsilon)$, and the constant $c^* > 0$ to be specified. By the choice of k^*, for any $k \leq k^*$,

$$k^{\frac{1-\alpha}{2}} \delta \leq k^{-\nu(1-\alpha)} \|w\|. \tag{4.20}$$

Now the proof proceeds by mathematical induction. When $k = 1$, (4.19) holds trivially for any sufficiently large c^*. Now we assume (4.19) holds up to some $k < k^*$, and prove it for $k + 1 \leq k^*$. Upon substituting the induction hypothesis, with $\varrho = c^* \|w\|^2$, we obtain

$$\begin{aligned} a_{k+1} \leq & \Big(\sum_{j=1}^k \eta_j \phi_j^{\frac12} \big(c_0 \varrho j^{-\frac{\beta+\gamma}{2}} + c_R \varrho^{\frac12} j^{-\frac{\beta}{2}} \delta + \delta \big) + c_\nu \|w\| (k+1)^{-\nu(1-\alpha)} \Big)^2 \\ & + n \sum_{j=1}^k \eta_j^2 (\phi_j^{\frac12})^2 (\varrho^{\frac12} j^{-\frac{\gamma}{2}} + \delta)^2 + 2n \Big(\sum_{i=1}^k \eta_i \phi_i^{\frac12} (\varrho^{\frac12} i^{-\frac{\gamma}{2}} + \delta) \Big) \\ & \times \Big(\sum_{j=1}^k \eta_j \phi_j^{\frac12} (c_0 \varrho^{\frac12} j^{-\frac{\gamma}{2}} + c_R \delta) \varrho^{\frac{\theta}{2}} j^{-\frac{\theta\beta}{2}} \Big) + n \Big(\sum_{j=1}^k \eta_j \phi_j^{\frac12} (c_0 \varrho^{\frac12} j^{-\frac{\gamma}{2}} + c_R \delta) \varrho^{\frac{\theta}{2}} j^{-\frac{\theta\beta}{2}} \Big)^2. \end{aligned} \tag{4.21}$$

Next, using Proposition A.2, we obtain

$$\begin{aligned} a_{k+1} \leq & \Big((c_1 (c_0 \varrho + (c_R \varrho^{\frac12} + 1) \|w\|) + c_\nu \|w\|)^2 + 2n (c_2 \varrho + c_3 \|w\|^2) \\ & + 2n c_1^2 (\varrho^{\frac12} + \|w\|)(c_0 \varrho^{\frac12} + c_R \|w\|) \varrho^{\frac{\theta}{2}} + n c_1^2 (c_0 \varrho^{\frac12} + c_R \|w\|)^2 \varrho^\theta \Big) (k+1)^{-\beta}, \end{aligned} \tag{4.22}$$

with the constants $c_1, \ldots, c_3$ given in Proposition A.2. Similarly, it follows from the induction hypothesis that

$$\begin{aligned} b_{k+1} \leq & \Big(\sum_{j=1}^k \eta_j \phi_j^1 \big(c_0 \varrho j^{-\frac{\beta+\gamma}{2}} + c_R \varrho^{\frac12} j^{-\frac{\beta}{2}} \delta + \delta \big) + c_\nu \|w\| (k+1)^{-(1-\alpha)(\nu+\frac12)} \Big)^2 \\ & + n \sum_{j=1}^k \eta_j^2 (\phi_j^1)^2 (\varrho^{\frac12} j^{-\frac{\gamma}{2}} + \delta)^2 + 2n \Big(\sum_{i=1}^k \eta_i \phi_i^1 (\varrho^{\frac12} i^{-\frac{\gamma}{2}} + \delta) \Big) \\ & \times \Big(\sum_{j=1}^k \eta_j \phi_j^1 (c_0 \varrho^{\frac12} j^{-\frac{\gamma}{2}} + c_R \delta) \varrho^{\frac{\theta}{2}} j^{-\frac{\theta\beta}{2}} \Big) + n \Big(\sum_{j=1}^k \eta_j \phi_j^1 (c_0 \varrho^{\frac12} j^{-\frac{\gamma}{2}} + c_R \delta) \varrho^{\frac{\theta}{2}} j^{-\frac{\theta\beta}{2}} \Big)^2, \end{aligned} \tag{4.23}$$

from which and Proposition A.2, it follows that

$$
\begin{aligned}
b_{k+1} \leq &\Big((c_0c_1'\varrho + c_5'(c_R\varrho^{\frac{1}{2}}+1)\|w\| + c_\nu\|w\|)^2 + 2n(c_2'\varrho + c_3\|w\|^2) \\
&+ 2n(c_3'\varrho^{\frac{1}{2}} + c_5'\|w\|)(c_0c_3'\varrho^{\frac{1}{2}} + c_Rc_5'\|w\|)\varrho^{\frac{\theta}{2}} + n(c_0c_4'\varrho^{\frac{1}{2}} + c_Rc_5'\|w\|)^2\varrho^\theta\Big)(k+1)^{-\gamma},
\end{aligned} \tag{4.24}
$$

with the constants $c_1', \ldots, c_5'$ given in Proposition A.2. In view of (4.22) and (4.24), for small $\|w\|$ and η_0, repeating the argument for Theorem 4.8 (and noting that c_1,c_2, c_3, c_2' tend to zero as $\eta_0 \to 0^+$) concludes the existence of a $c^* > 0$ such that (4.19) hold. This completes the induction step and the proof of Theorem 2.2. ∎

5. Concluding remarks. In this work, we have provided a convergence analysis of stochastic gradient descent for a class of nonlinear ill-posed inverse problems. The method employs an unbiased estimate of the gradient, computed from one randomly selected equation of the nonlinear system, and admits excellent scalability to the problem size. We proved that it is regularizing under the traditional tangential cone condition with *a priori* parameter choice, and also showed a convergence rate under canonical source condition and range invariance condition (and its stochastic variant), for a polynomially decaying step size schedule. The analysis combines techniques from both nonlinear regularization theory and stochastic calculus, and the results extend the existing works [8] and [14].

There are several avenues in both theoretical and practical aspects for further research. First, it is important to verify the assumptions for concrete nonlinear inverse problems, especially nonlinearity conditions in Assumption 2.1(ii)–(iii) and Assumption 2.3, for e.g., parameter identifications for PDEs, which would justify the usage of SGD. Several important inverse problems in medical imaging are of the form (1.1), e.g., electrical impedance tomography and diffuse optical spectroscopy. These applications often involve natural physical constraints, e.g., positivity, which the algorithm should be adapted to preserve. Second, the source condition employed in the work is canonical, and alternative approaches, e.g., variational inequalities and conditional stability, should also be studied for convergence rates [24], or the Frechét differentiability of the forward operator in Assumption 2.1 may be relaxed [3]. Third, the influence of various algorithmic parameters, e.g., mini-batch, random sampling, step size schedules (including adaptive rules) and *a posteriori* stopping rule, should be analyzed to provide useful practical guidelines.

Acknowledgements. The authors are grateful to the associate editor, Professor Frank E. Curtis, and two anonymous referees for helpful comments.

Appendix A. Auxiliary estimates. In this appendix, we collect several auxiliary inequalities that have been used in the convergence rates analysis. Most estimates follow from routine but rather tedious computations. We begin with a well known estimate on operator norms (see, e.g., [19] [14, Lemma A.1]).

LEMMA A.1. *For any $j < k$, and any symmetric and positive semidefinite operator S and step sizes $\eta_j \in (0, \|S\|^{-1}]$ and $p \geq 0$, there holds*

$$
\|\prod_{i=j}^{k}(I - \eta_i S)S^p\| \leq \frac{p^p}{e^p(\sum_{i=j}^k \eta_i)^p}.
$$

Below we need the Beta function $B(a,b) = \int_0^1 s^{a-1}(1-s)^{b-1}\mathrm{d}s$ for any $a, b > 0$. Note that for fixed a, the function $B(a,\cdot)$ is monotonically decreasing.

Lemma A.2. *For $\eta_j = \eta_0 j^{-\alpha}$ with $\alpha \in (0,1)$, $r \in [0,1)$, $\beta \in [0,1]$, and $\gamma = \alpha + \beta$, the following estimates hold*

$$\sum_{i=1}^{k} \eta_i \geq (1-2^{\alpha-1})(1-\alpha)^{-1}\eta_0(k+1)^{1-\alpha},$$

$$\sum_{j=1}^{k-1} \frac{\eta_j}{(\sum_{\ell=j+1}^{k}\eta_\ell)^r} j^{-\beta} \leq \eta_0^{1-r} B(1-r, 1-\gamma) k^{r\alpha+1-r-\gamma}, \quad r \in [0,1), \gamma < 1,$$

$$\sum_{j=1}^{k-1} \frac{\eta_j}{\sum_{\ell=j+1}^{k}\eta_\ell} j^{-\beta} \leq \begin{cases} 2^\gamma(1-\gamma)^{-1}k^{-\beta}, & \gamma < 1, \\ 4k^{\alpha-1}\ln k, & \gamma = 1, \\ 2\gamma(\gamma-1)^{-1}k^{\alpha-1}, & \gamma > 1, \end{cases} + 2^{1+\gamma}k^{-\beta}\ln k.$$

Proof. The first estimate follows from the fact $1-(k+1)^{\alpha-1} \geq 1 - 2^{\alpha-1}$ for $k \geq 1$ that

$$\sum_{i=1}^{k} \eta_i \geq \eta_0 \int_1^{k+1} s^{-\alpha} \mathrm{d}s = \eta_0(1-\alpha)^{-1}((k+1)^{1-\alpha}-1) \geq \eta_0(1-\alpha)^{-1}(1-2^{\alpha-1})(k+1)^{1-\alpha}.$$

To prove the second estimate, we note $\eta_i \geq \eta_0 k^{-\alpha}$ for any $i = j+1, \ldots, k$ and thus

$$\eta_0^{-1} \sum_{i=j+1}^{k} \eta_i \geq k^{-\alpha}(k-j). \tag{A.1}$$

Thus, if $\gamma = \alpha + \beta < 1$ and $r < 1$,

$$\begin{aligned} \sum_{j=1}^{k-1} \frac{\eta_j}{(\sum_{\ell=j+1}^{k}\eta_\ell)^r} j^{-\beta} \leq& \eta_0^{1-r} k^{r\alpha} \sum_{j=1}^{k-1} (k-j)^{-r} j^{-\gamma} \leq \eta_0^{1-r}k^{r\alpha} \int_0^k (k-s)^{-r}s^{-\gamma}\mathrm{d}s \\ =& \eta_0^{1-r} B(1-r, 1-\gamma) k^{r\alpha+1-r-\gamma}. \end{aligned}$$

Similarly, if $r = 1$, it follows from (A.1) that

$$\begin{aligned} \sum_{j=1}^{k-1} \frac{\eta_j}{\sum_{\ell=j+1}^{k}\eta_\ell} j^{-\beta} \leq& k^\alpha \sum_{j=1}^{k-1} (k-j)^{-1} j^{-\gamma} \\ =& k^\alpha \sum_{j=1}^{[\frac{k}{2}]} j^{-\gamma}(k-j)^{-1} + k^\alpha \sum_{j=[\frac{k}{2}]+1}^{k-1} j^{-\gamma}(k-j)^{-1} \\ \leq& 2k^{\alpha-1} \sum_{j=1}^{[\frac{k}{2}]} j^{-\gamma} + 2^\gamma k^{-\beta} \sum_{j=[\frac{k}{2}]+1}^{k-1} (k-j)^{-1}. \end{aligned}$$

Simple computation gives

$$\sum_{j=[\frac{k}{2}]+1}^{k-1} (k-j)^{-1} \leq 2\ln k \quad \text{and} \quad \sum_{j=1}^{[\frac{k}{2}]} j^{-\gamma} \leq \begin{cases} (1-\gamma)^{-1}(\frac{k}{2})^{1-\gamma}, & \gamma \in [0,1), \\ 2\ln k, & \gamma = 1, \\ \gamma(\gamma-1)^{-1}, & \gamma > 1. \end{cases} \tag{A.2}$$

Combining the last three estimates gives the assertion for the case $r = 1$.

Next we recall two useful estimates.

LEMMA A.3. *For $\eta_j = \eta_0 j^{-\alpha}$ with $\alpha \in (0,1)$, $\beta \in [0,1]$ and $r \geq 0$, there hold*

$$\sum_{j=1}^{[\frac{k}{2}]} \frac{\eta_j^2}{(\sum_{\ell=j+1}^k \eta_\ell)^r} j^{-\beta} \leq c_{\alpha,\beta,r} k^{-r(1-\alpha)+\max(0,1-2\alpha-\beta)},$$
$$\sum_{j=[\frac{k}{2}]+1}^{k-1} \frac{\eta_j^2}{(\sum_{\ell=j+1}^k \eta_\ell)^r} j^{-\beta} \leq c'_{\alpha,\beta,r} k^{-((2-r)\alpha+\beta)+\max(0,1-r)},$$

where we slightly abuse the notation $k^{-\max(0,0)}$ for $\ln k$, and $c_{\alpha,\beta,r}$ and $c'_{\alpha,\beta,r}$ are given by

$$c_{\alpha,\beta,r} = 2^r \eta_0^{2-r} \begin{cases} \frac{2\alpha+\beta}{2\alpha+\beta-1}, & 2\alpha+\beta>1, \\ 2, & 2\alpha+\beta=1, \\ \frac{2^{2\alpha+\beta-1}}{1-2\alpha-\beta}, & 2\alpha+\beta<1, \end{cases} \quad \text{and} \quad c'_{\alpha,\beta,r} = 2^{2\alpha+\beta}\eta_0^{2-r} \begin{cases} \frac{r}{r-1}, & r>1, \\ 2, & r=1, \\ \frac{2^{r-1}}{1-r}, & r<1. \end{cases}$$

Proof. The proof is based on (A.1) and (A.2) and essentially given in [14, Lemma A.3], but the constants are corrected. ∎

The next result collects some lengthy estimates needed in the proof of Theorem 4.8.

PROPOSITION A.1. *Let $\beta = \min(2\nu(1-\alpha), \alpha-\epsilon)$, $\gamma = \min((1+2\nu)(1-\alpha), 1-\epsilon)$ and $r = \min(\frac{1}{2}+\nu, \frac{1-\epsilon}{2(1-\alpha)})$. Then under the conditions in Theorem 4.8, i.e., $\|B\| \leq 1$, $\eta_0 \leq 1$ and θ being sufficiently small, with $\zeta = (\frac{1}{2}-\nu)(1-\alpha)$, the following estimates hold:*

$$\text{(A.3)} \quad \sum_{j=1}^k \eta_j \phi_j^{\frac{1}{2}} j^{-\frac{\gamma}{2}} \leq c_1(k+1)^{-\frac{\beta}{2}}, \ \sum_{j=1}^k \eta_j^2 (\phi_j^{\frac{1}{2}})^2 j^{-\gamma} \leq c_2(k+1)^{-\beta},$$

$$\text{(A.4)} \quad \sum_{j=1}^{[\frac{k}{2}]} \eta_j^2 (\phi_j^r)^2 j^{-\gamma} + \sum_{j=[\frac{k}{2}]+1}^k \eta_j^2 (\phi_j^{\frac{1}{2}})^2 j^{-\gamma} \leq c_3(k+1)^{-\gamma}, \ \sum_{j=1}^k \eta_j \phi_j^1 j^{-\frac{\beta+\gamma}{2}} \leq c_4(k+1)^{-\frac{\gamma}{2}},$$

$$\text{(A.5)} \quad \Big(\sum_{i=1}^k \eta_i \phi_i^1 i^{-\frac{\gamma}{2}}\Big)\Big(\sum_{j=1}^k \eta_j \phi_j^1 j^{-\frac{\gamma+\theta\beta}{2}}\Big) \leq c_5(k+1)^{-\gamma}, \ \sum_{j=1}^k \eta_j \phi_j^1 j^{-\frac{\gamma+\theta\beta}{2}} \leq c_6(k+1)^{-\frac{\gamma}{2}}.$$

with $c_1 = 2^{\frac{\beta}{2}}\eta_0^{\frac{1}{2}}(2^{-1}B(\frac{1}{2},\zeta)+1)$, $c_2 = 2^\beta \eta_0(\alpha^{-1}+2)$, $c_3 = 2^\gamma \eta_0^{2-2r}(3\alpha^{-1}+1)$, $c_4 = 2^{\frac{\gamma}{2}}(\zeta^{-1}+2\beta^{-1}+1)$, $c_5 = 2^\gamma\big(((\frac{1}{2}-\nu-\theta\nu)(1-\alpha))^{-1}+4(\theta\beta)^{-1}+1\big)^2$ *and* $c_6 = 2^{\frac{\gamma}{2}}(\zeta^{-1}+2(\theta\beta)^{-1}+1)$.

Proof. It follows from Lemma A.1 and the condition $\|B\| \leq 1$ that

$$\sum_{j=1}^k \eta_j \phi_j^{\frac{1}{2}} j^{-\frac{\gamma}{2}} \leq (2e)^{-\frac{1}{2}} \sum_{j=1}^{k-1} \frac{\eta_j}{(\sum_{\ell=1}^k \eta_\ell)^{\frac{1}{2}}} j^{-\frac{\gamma}{2}} + \eta_0 k^{-\alpha-\frac{\gamma}{2}}$$
$$\leq (\eta_0^{\frac{1}{2}} 2^{-1} B(\tfrac{1}{2}, 1-\alpha-\tfrac{\gamma}{2}) + \eta_0) k^{\frac{1-\alpha}{2}-\frac{\gamma}{2}}.$$

By the definitions of β and γ, we have $\frac{1-\alpha}{2} - \frac{\gamma}{2} = -\frac{\beta}{2}$, and $1-\alpha-\frac{\gamma}{2} \geq (\frac{1}{2}-\nu)(1-\alpha) := \zeta$. Thus, the monotonicity of the Beta function, and the inequality $2k \geq k+1$ for $k \geq 1$ imply the first inequality of (A.3). Now by Lemma A.1 and Lemma A.3,

$$\text{(A.6)} \qquad \sum_{j=1}^k \eta_j^2 (\phi_j^{\frac{1}{2}})^2 j^{-\gamma} \leq (2e)^{-1} \sum_{j=1}^{k-1} \frac{\eta_j^2}{\sum_{\ell=j+1}^k \eta_j} j^{-\gamma} + \eta_0^2 \|B^{\frac{1}{2}}\|^2 k^{-2\alpha-\gamma}$$

$$\leq \eta_0\Big((2e)^{-1}\frac{2(2\alpha+\gamma)}{2\alpha+\gamma-1}k^{-(1-\alpha)}+(2e)^{-1}2^{1+2\alpha+\gamma}k^{-\alpha-\gamma}\ln k+\eta_0\|B^{\frac12}\|^2k^{-2\alpha-\gamma}\Big).$$

Now, for any $r>0$, there holds

$$s^{-r}\ln s\leq (er)^{-1},\quad \forall s\geq 0, \tag{A.7}$$

and thus $k^{-\alpha-\gamma}\ln k=k^{-\beta}(k^{-1}\ln k)\leq e^{-1}k^{-\beta}$. Further, by the definition of γ, $2\alpha+\gamma\leq \min(2,1+2\alpha)\leq 2$, and since $\epsilon<\frac{\alpha}{2}$, $2\alpha+\gamma-1\geq\alpha$,

$$\tfrac{2\alpha+\gamma}{2\alpha+\gamma-1}=1+\tfrac{1}{2\alpha+\gamma-1}\leq 1+\alpha^{-1}. \tag{A.8}$$

Then, the last three estimates (with $\|B\|\leq 1$) imply

$$\sum_{j=1}^{k}\eta_j^2(\phi_j^{\frac12})^2j^{-\gamma}\leq 2^{\beta}\eta_0\big(\alpha^{-1}+2\big)(k+1)^{-\beta}\,.$$

This proves the second inequality in (A.3).

Next, by letting $r=\min(\frac12+\nu,\frac{1-\epsilon}{2(1-\alpha)})\in(\frac12,1)$, and using (A.7) and (A.8), Lemma A.1 and Lemma A.3 and the monotonicity of $\frac{s^s}{e^s}$ for $s\in[0,1]$, the first part of (A.4) follows from

$$\sum_{j=1}^{[\frac{k}{2}]}\eta_j^2(\phi_j^r)^2j^{-\gamma}+\sum_{j=[\frac{k}{2}]+1}^{k}\eta_j^2(\phi_j^{\frac12})^2j^{-\gamma}$$

$$\leq (2e)^{-1}\Big(\sum_{j=1}^{[\frac{k}{2}]}\frac{\eta_j^2}{(\sum_{\ell=1}^{j}\eta_\ell)^{2r}}j^{-\gamma}+\sum_{j=[\frac{k}{2}]+1}^{k-1}\frac{\eta_j^2}{\sum_{\ell=j+1}^{k}\eta_\ell}j^{-\gamma}\Big)+\eta_0^2k^{-2\alpha-\gamma}$$

$$\leq \eta_0^{2-2r}\frac{2^{2r}(2\alpha+\gamma)}{2e(2\alpha+\gamma-1)}k^{-\gamma}+\frac{2^{1+2\alpha+\gamma}}{2e}\eta_0k^{-(\alpha+\gamma)}\ln k+\eta_0^2k^{-2\alpha-\gamma}\leq c_3(k+1)^{-\gamma}.$$

Now, we bound the sum $\sum_{j=1}^{k}\eta_j\phi_j^1j^{-\sigma}$ for any $\sigma\in[\frac{\gamma}{2},\frac{\gamma+\beta}{2}]$, and then set σ to $\frac{\gamma}{2}$, $\frac{\gamma+\theta\beta}{2}$ and $\frac{\gamma+\beta}{2}$ to complete the proof. By Lemma A.1 and Lemma A.2, there hold

$$\sum_{j=1}^{[\frac{k}{2}]}\eta_j\phi_j^1j^{-\sigma}\leq e^{-1}\begin{cases}\frac{2^{\alpha+\sigma}}{1-\alpha-\sigma}k^{-\sigma}, & \alpha+\sigma<1,\\ 4k^{\alpha-1}\ln k, & \alpha+\sigma=1,\\ \frac{2(\alpha+\sigma)}{\alpha+\sigma-1}k^{\alpha-1}, & \alpha+\sigma>1,\end{cases} \tag{A.9}$$

$$\sum_{[\frac{k}{2}]+1}^{k}\eta_j\phi_j^1j^{-\sigma}\leq e^{-1}2^{1+\alpha+\sigma}k^{-\sigma}\ln k+\eta_0k^{-\sigma}. \tag{A.10}$$

First, we choose $\sigma=\frac{\beta+\gamma}{2}$. By (A.7), since $(1-\alpha-\frac{\gamma}{2})^{-1}\leq\zeta^{-1}$, $\alpha+\frac{\gamma}{2}<1$, $\|B\|\leq 1$ and $\eta_0\leq 1$, we obtain

$$\sum_{j=1}^{k}\eta_j\phi_j^1j^{-\frac{\beta+\gamma}{2}}\leq\sum_{j=1}^{[\frac{k}{2}]}\eta_j\phi_j^1j^{-\frac{\gamma}{2}}+\sum_{j=[\frac{k}{2}]+1}^{k}\eta_j\phi_j^1j^{-\frac{\beta+\gamma}{2}}$$

$$\leq 2^{\alpha+\frac{\gamma}{2}}e^{-1}(1-\alpha-\tfrac{\gamma}{2})^{-1}k^{-\frac{\gamma}{2}}+2^{1+\alpha+\frac{\gamma+\beta}{2}}e^{-1}k^{-\frac{\gamma+\beta}{2}}\ln k+\eta_0k^{-\frac{\gamma}{2}}\leq c_4(k+1)^{-\frac{\gamma}{2}},$$

due to the inequality $2^{1+\alpha+\frac{\beta+\gamma}{2}} < e^2$, from the definitions of the exponents β and γ. This shows the second inequality of (A.4). Since θ is small, we may assume $\theta < \frac{1}{2\nu} - 1 \leq \frac{1-\alpha}{\beta} - 1$. Then by the relations $\gamma = 1-\alpha+\beta$ and $\beta \leq 2\nu(1-\alpha)$, direct computation shows $1-\alpha-\frac{\gamma+\theta\beta}{2} \geq (\frac{1}{2}-\nu-\theta\nu)(1-\alpha) > 0$. Further, since $\theta < \frac{1-\alpha}{\beta}-1$, $\min(\frac{\theta\beta}{2}, 1-\alpha-\frac{\gamma}{2}) = \frac{\theta\beta}{2}$. Hence, it follows from (A.9) and (A.10), with $\sigma = \frac{\gamma}{2}$ and $\frac{\gamma+\theta\beta}{2}$ that

$$
\begin{aligned}
\Big(\sum_{i=1}^k \eta_i \phi_i^1 i^{-\frac{\gamma}{2}}\Big)\Big(\sum_{j=1}^k \eta_j \phi_j^1 j^{-\frac{\gamma+\theta\beta}{2}}\Big) \leq{}& \Big(\frac{2^{\alpha+\frac{\gamma}{2}}}{e(1-\alpha-\frac{\gamma}{2})} + \frac{2^{1+\alpha+\frac{\gamma}{2}}}{e}\ln k + 1\Big) \\
&\times \Big(\frac{2^{\alpha+\frac{\gamma+\theta\beta}{2}}}{e(1-\alpha-\frac{\gamma+\theta\beta}{2})} k^{-\min(\frac{\theta\beta}{2}, 1-\alpha-\frac{\gamma}{2})} + \frac{2^{1+\alpha+\frac{\gamma+\theta\beta}{2}}}{e} k^{-\frac{\theta\beta}{2}}\ln k + k^{-\frac{\theta\beta}{2}}\Big) k^{-\gamma}.
\end{aligned}
$$

Then we move one factor $k^{-\frac{\theta\beta}{4}}$ from the second bracket to the first and bound by (A.7):

$$
\begin{aligned}
\Big(\sum_{i=1}^k \eta_i \phi_i^1 i^{-\frac{\gamma}{2}}\Big)\Big(\sum_{j=1}^k \eta_j \phi_j^1 j^{-\frac{\gamma+\theta\beta}{2}}\Big) \leq{}& \Big(\frac{2^{\alpha+\frac{\gamma}{2}}}{e(1-\alpha-\frac{\gamma}{2})} + \frac{2^{1+\alpha+\frac{\gamma}{2}}}{e} k^{-\frac{\theta\beta}{4}}\ln k + 1\Big) \\
&\times \Big(\frac{2^{\alpha+\frac{\gamma+\theta\beta}{2}}}{e(1-\alpha-\frac{\gamma+\theta\beta}{2})} + \frac{2^{1+\alpha+\frac{\gamma+\theta\beta}{2}}}{e} k^{-\frac{\theta\beta}{4}}\ln k + 1\Big) k^{-\gamma} \\
&\leq 2^{\gamma}\big(((\tfrac{1}{2}-\nu-\theta\nu)(1-\alpha))^{-1} + 4(\theta\beta)^{-1} + 1\big)^2 (k+1)^{-\gamma},
\end{aligned}
$$

proving the first inequality of (A.5). The other estimate in (A.5) follows similarly by choosing $\sigma = \frac{\gamma+\theta\beta}{2}$, and hence omitted. ∎

Remark A.1. *The proof of Proposition A.1 implies* $\sum_{j=1}^{k-1} \eta_j \phi_j^1 j^{-\frac{\gamma}{2}} \leq (\zeta^{-1} + 2\ln k) k^{-\frac{\gamma}{2}}$. *The log factor* $\ln k$ *seems not removable, and precludes a direct application of mathematical induction in the proof of Theorem 4.8. The extra factor* $j^{-\frac{\theta\beta}{2}}$ *due to Assumption 2.3 gracefully compensates the log factor* $\ln k$ *using (A.7). The smallness condition on* θ *can be removed at the expense of less transparent dependence. Specifically, by Lemma A.2, with* $\sigma = \alpha + \frac{\gamma+\theta\beta}{2}$*, there holds*

$$
\sum_{j=1}^k \eta_j \phi_j^1 j^{-\frac{\gamma+\theta\beta}{2}} \leq \frac{1}{ek^{\frac{\gamma}{2}}} \begin{cases} \frac{2^{\sigma}}{1-\sigma} k^{-\frac{\theta\beta}{2}}, & \sigma < 1 \\ 4k^{-(1-\alpha-\frac{\gamma}{2})}\ln k, & \sigma = 1 \\ \frac{2\sigma}{\sigma-1} k^{-(1-\alpha-\frac{\gamma}{2})}, & \sigma > 1 \end{cases} + 2^{1+\sigma} e^{-1} k^{-\frac{\gamma}{2}-\frac{\theta\beta}{2}} \ln k + k^{-(\alpha+\frac{\gamma+\theta\beta}{2})}.
$$

Instead of applying (A.7) directly, we rearrange the terms and discuss the cases $\sigma < 1$, $\sigma = 1$ *and* $\sigma > 1$ *separately with the argument in the proof of Proposition A.1 and obtain*

$$
\Big(\sum_{i=1}^k \eta_i \phi_i^1 i^{-\frac{\gamma}{2}}\Big)\Big(\sum_{j=1}^k \eta_j \phi_j^1 j^{-\frac{\gamma+\theta\beta}{2}}\Big) \leq c_\sigma 2^{\gamma} (k+1)^{-\gamma},
$$

with the constant c_σ *given by*

$$
c_\sigma = \begin{cases} (1-\sigma)^{-1} + 4(\theta\beta)^{-1} + 1, & \sigma < 1, \\ \zeta^{-1} + 8(\theta\beta)^{-1} + 1, & \sigma = 1, \\ 2(\sigma-1)^{-1} + 3\zeta^{-1} + 1, & \sigma > 1. \end{cases}
$$

The next result gives some basic estimates used in the proof of Theorem 2.2.

PROPOSITION A.2. *Under the induction hypothesis of Theorem 2.2 and (4.20), there hold*

$$\begin{aligned} a_{k+1} \leq& \Big((c_1(c_0\varrho + (c_R\varrho^{\frac12}+1)\|w\|) + c_\nu\|w\|)^2 + 2n(c_2\varrho + c_3\|w\|^2) \\ &+ 2nc_1^2\big(\varrho^{\frac12}+\|w\|\big)\big(c_0\varrho^{\frac12}+c_R\|w\|\big)\varrho^{\frac{\theta}{2}} + nc_1^2(c_0\varrho^{\frac12}+c_R\|w\|)^2\varrho^\theta\Big)(k+1)^{-\beta}, \\ b_{k+1} \leq& \Big((c_0c_1'\varrho + c_5'(c_R\varrho^{\frac12}+1)\|w\| + c_\nu\|w\|)^2 + 2n(c_2'\varrho + c_3\|w\|^2) \\ &+ 2n(c_3'\varrho^{\frac12}+c_5'\|w\|)(c_0c_3'\varrho^{\frac12}+c_5'c_R\|w\|)\varrho^{\frac{\theta}{2}} + n(c_0c_4'\varrho^{\frac12}+c_5'c_R\|w\|)^2\varrho^\theta\Big)(k+1)^{-\gamma}, \end{aligned}$$

where the constants c_1, c_2, c_3 and $c_1', \ldots, c_5'$ are given in the proof.

Proof. First, it follows directly from Lemma A.1, Lemma A.2, and Lemma A.3 and the assumptions $\|B\| \leq 1$ and $\eta_0 \leq 1$ that for any $\sigma \in [0, 1-\alpha)$,

$$\sum_{j=1}^k \eta_j\phi_j^{\frac12} j^{-\sigma} \leq \eta_0^{\frac12}(2^{-1}B(\tfrac12, 1-\alpha-\sigma)+1)k^{\frac{1-\alpha}{2}-\sigma}, \tag{A.11}$$

$$\sum_{j=1}^k \eta_j^2(\phi_j^{\frac12})^2 \leq \eta_0(|1-2\alpha|^{-1}+\alpha^{-1}+1) := c_3, \tag{A.12}$$

where we have abused the writing 0^{-1} for 1. Meanwhile, by Proposition A.1, we have

$$\sum_{j=1}^k \eta_j\phi_j^{\frac12} j^{-\frac{\gamma}{2}} \leq c_1(k+1)^{-\frac{\beta}{2}} \quad \text{and} \quad \sum_{j=1}^k \eta_j^2(\phi_j^{\frac12})^2 j^{-\gamma} \leq c_2(k+1)^{-\beta}, \tag{A.13}$$

with $c_1 = 2^{\frac{\beta}{2}}\eta_0^{\frac12}(2^{-1}B(\frac12,\zeta)+1)$, $\zeta = (\frac12-\nu)(1-\alpha)$ and $c_2 = 2^\beta\eta_0(\alpha^{-1}+2)$. By (A.11)-(A.13) and the monotonicity of the Beta function, and $k+1 \leq k^*$ (cf. (4.20)), we obtain

$$\begin{aligned} \sum_{j=1}^k \eta_j\phi_j^{\frac12}\big(c_0\varrho j^{-\frac{\beta+\gamma}{2}} + c_R\varrho^{\frac12}j^{-\frac{\beta}{2}}\delta + \delta\big) \leq& c_0c_1\varrho(k+1)^{-\frac{\beta}{2}} + (c_R\varrho^{\frac12}+1)c_1(k+1)^{\frac{1-\alpha}{2}}\delta \\ \leq& c_1\big(c_0\varrho + (c_R\varrho^{\frac12}+1)\|w\|\big)(k+1)^{-\frac{\beta}{2}}, \\ \sum_{j=1}^k \eta_j^2(\phi_j^{\frac12})^2(\varrho^{\frac12}j^{-\frac{\gamma}{2}}+\delta)^2 \leq& 2(c_2\varrho + c_3\|w\|^2)(k+1)^{-\beta}. \end{aligned}$$

Likewise, by the monotonicity of the Beta function, we deduce

$$\begin{aligned} &\Big(\sum_{i=1}^k \eta_i\phi_i^{\frac12}(\varrho^{\frac12}i^{-\frac{\gamma}{2}}+\delta)\Big)\Big(\sum_{j=1}^k \eta_j\phi_j^{\frac12}(c_0\varrho^{\frac12}j^{-\frac{\gamma}{2}}+c_R\delta)\varrho^{\frac{\theta}{2}}j^{-\frac{\theta\beta}{2}}\Big) \\ \leq& c_1^2(\varrho^{\frac12}+\|w\|)(c_0\varrho^{\frac12}+c_R\|w\|)\varrho^{\frac{\theta}{2}}(k+1)^{-\beta}, \\ &\sum_{j=1}^k \eta_j\phi_j^{\frac12}(c_0\varrho^{\frac12}j^{-\frac{\gamma}{2}}+c_R\delta)\varrho^{\frac{\theta}{2}}j^{-\frac{\theta\beta}{2}} \leq c_1(c_0\varrho^{\frac12}+c_R\|w\|)\varrho^{\frac{\theta}{2}}(k+1)^{-\frac{\beta}{2}}. \end{aligned}$$

The last four estimates give (4.21). Now we prove (4.23). By Proposition A.1, we have

$$\sum_{j=1}^k \eta_j\phi_j^1 j^{-\frac{\beta+\gamma}{2}} \leq c_1'(k+1)^{-\frac{\gamma}{2}}, \quad \sum_{j=1}^{[\frac{k}{2}]} \eta_j^2(\phi_j^r)^2 j^{-\gamma} + \sum_{j=[\frac{k}{2}]+1}^k \eta_j^2(\phi_j^{\frac12})^2 j^{-\gamma} \leq c_2'(k+1)^{-\gamma},$$

$$\Big(\sum_{i=1}^k \eta_i\phi_i^1 i^{-\frac{\gamma}{2}}\Big)\Big(\sum_{j=1}^k \eta_j\phi_j^1 j^{-\frac{\gamma+\theta\beta}{2}}\Big) \le c_3'^2(k+1)^{-\gamma}, \quad \sum_{j=1}^k \eta_j\phi_j^1 j^{-\frac{\gamma+\theta\beta}{2}} \le c_4'(k+1)^{-\frac{\gamma}{2}},$$

with $c_1' = 2^{\frac{\gamma}{2}}(\zeta^{-1}+2\beta^{-1}+1)$, $c_2' = 2^\gamma\eta_0^{2-2r}(3\alpha^{-1}+1)$, $c_3' = 2^{\frac{\gamma}{2}}(((\frac12-\nu-\theta\nu)(1-\alpha))^{-1}+4(\theta\beta)^{-1}+1)$ and $c_4' = 2^{\frac{\gamma}{2}}(\zeta^{-1}+2(\theta\beta)^{-1}+1)$. Further, by (A.9) and (A.10), for any $\sigma\in[0,\frac{\gamma}{2}]$,

$$k^{-\nu(1-\alpha)}\sum_{j=1}^k \eta_j\phi_j^1 j^{-\sigma} \le \zeta^{-1}+2(\nu(1-\alpha))^{-1}+1 := c_5'.$$

With these estimates and (4.20), we deduce

$$\begin{aligned}
\sum_{j=1}^k \eta_j\phi_j^1\big(c_0\varrho j^{-\frac{\beta+\gamma}{2}} + c_R\varrho^{\frac12}j^{-\frac{\beta}{2}}\delta+\delta\big) \le& (c_0c_1'\varrho + c_5'(c_R\varrho^{\frac12}+1)\|w\|)(k+1)^{-\frac{\gamma}{2}},\\
\sum_{j=1}^k \eta_j^2(\phi_j^1)^2(\varrho^{\frac12}j^{-\frac{\gamma}{2}}+\delta)^2 \le& 2(c_2'\varrho + c_3\|w\|^2)(k+1)^{-\gamma},\\
\sum_{j=1}^k \eta_j\phi_j^1(c_0\varrho^{\frac12}j^{-\frac{\gamma}{2}}+c_R\delta)\varrho^{\frac{\theta}{2}}j^{-\frac{\theta\beta}{2}} \le& (c_0c_4'\varrho^{\frac12}+c_5'c_R\|w\|)\varrho^{\frac{\theta}{2}}(k+1)^{-\frac{\gamma}{2}},
\end{aligned}$$

where the second line is due to (A.12) and the inequality $\sum_{j=1}^k \eta_j^2(\phi_j^1)^2 \le \sum_{j=1}^k \eta_j^2(\phi_j^{\frac12})^2$ (since $\|B\|\le 1$). Last, repeating the argument in Proposition A.1 gives

$$\begin{aligned}
&\Big(\sum_{i=1}^k \eta_i\phi_i^1(\varrho^{\frac12}i^{-\frac{\gamma}{2}}+\delta)\Big)\Big(\sum_{j=1}^k \eta_j\phi_j^1(c_0\varrho^{\frac12}j^{-\frac{\gamma}{2}}+c_R\delta)\varrho^{\frac{\theta}{2}}j^{-\frac{\theta\beta}{2}}\Big)\\
\le& (c_3'\varrho^{\frac12}+c_5'\|w\|)(c_0c_3'\varrho^{\frac12}+c_5'c_R\|w\|)\varrho^{\frac{\theta}{2}}(k+1)^{-\gamma}.
\end{aligned}$$

Then combining the last four estimates yields the desired bound on b_{k+1}. ∎

REFERENCES

[1] L. Bottou, F. E. Curtis, and J. Nocedal, *Optimization methods for large-scale machine learning*, SIAM Rev., 60 (2018), pp. 223–311.

[2] K. Chen, Q. Li, and J.-G. Liu, *Online learning in optical tomography: a stochastic approach*, Inverse Problems, 34 (2018), pp. 075010, 26 pp.

[3] C. Clason and V. H. Nhu, *Bouligand–Landweber iteration for a non-smooth ill-posed problem*, Numer. Math., (2019), p. in press.

[4] A. Dieuleveut and F. Bach, *Nonparametric stochastic approximation with large step-sizes*, Ann. Statist., 44 (2016), pp. 1363–1399.

[5] H. W. Engl, M. Hanke, and A. Neubauer, *Regularization of Inverse Problems*, Kluwer, Dordrecht, 1996.

[6] C. Geiersbach and G. C. Pflug, *Projected stochastic gradients for convex constrained problems in Hilbert spaces*, SIAM J. Optim., 29 (2019), pp. 2079–2099.

[7] R. M. Gower, N. Loizou, X. Qian, A. Sailanbayev, E. Shulgin, and P. Richtárik, *SGD: general analysis and improved rates*, in Proceedings of the 36 th International Conference on Machine Learning, PMLR 97, K. Chaudhuri and R. Salakhutdinov, eds., Long Beach, California, 2019, pp. 5200–5209.

[8] M. Hanke, A. Neubauer, and O. Scherzer, *A convergence analysis of the Landweber iteration for nonlinear ill-posed problems*, Numer. Math., 72 (1995), pp. 21–37.

[9] G. T. Herman, A. Lent, and P. H. Lutz, *Relaxation method for image reconstruction*, Comm. ACM, 21 (1978), pp. 152–158.

[10] G. T. Herman and L. B. Meyer, *Algebraic reconstruction techniques can be made computationally efficient*, IEEE Trans. Medical Imag., 12 (1993), pp. 600–609.

[11] K. Ito and B. Jin, *A new approach to nonlinear constrained Tikhonov regularization*, Inverse Problems, 27 (2011), pp. 105005, 23 pp.

[12] ———, *Inverse Problems: Tikhonov Theory and Algorithms*, World Scientific, Hackensack, NJ, 2015.

[13] Y. Jiao, B. Jin, and X. Lu, *Preasymptotic convergence of randomized Kaczmarz method*, Inverse Problems, 33 (2017), pp. 125012, 21 pp.

[14] B. Jin and X. Lu, *On the regularizing property of stochastic gradient descent*, Inverse Problems, 35 (2019), pp. 015004, 27 pp.

[15] B. Kaltenbacher, A. Neubauer, and O. Scherzer, *Iterative Regularization Methods for Nonlinear Ill-posed Problems*, Walter de Gruyter, Berlin, 2008.

[16] D. P. Kingma and J. Ba, *Adam: a method for stochastic optimization*, in Proceedings of the 3rd International Conference on Learning Representations (ICLR), 2015.

[17] H. J. Kushner and G. G. Yin, *Stochastic Approximation and Recursive Algorithms and Applications*, Springer-Verlag, New York, second ed., 2003.

[18] L. Landweber, *An iteration formula for Fredholm integral equations of the first kind*, Amer. J. Math., 73 (1951), pp. 615–624.

[19] J. Lin and L. Rosasco, *Optimal rates for multi-pass stochastic gradient methods*, J. Mach. Learn. Res., 18 (2017), pp. 1–47.

[20] A. K. Louis, *Inverse und Schlecht Gestellte Probleme*, B. G. Teubner, Stuttgart, 1989.

[21] S. F. McCormick and G. H. Rodrigue, *A uniform approach to gradient methods for linear operator equations*, J. Math. Anal. Appl., 49 (1975), pp. 275–285.

[22] H. Robbins and S. Monro, *A stochastic approximation method*, Ann. Math. Stat., 22 (1951), pp. 400–407.

[23] O. Scherzer, M. Grasmair, H. Grossauer, M. Haltmeier, and F. Lenzen, *Variational Methods in Imaging*, Springer, New York, 2009.

[24] T. Schuster, B. Kaltenbacher, B. Hofmann, and K. S. Kazimierski, *Regularization Methods in Banach Spaces*, Walter de Gruyter, Berlin, 2012.

[25] T. Strohmer and R. Vershynin, *A randomized Kaczmarz algorithm with exponential convergence*, J. Fourier Anal. Appl., 15 (2009), pp. 262–278.

[26] I. Sutskever, J. Martens, G. Dahl, and G. E. Hinton, *On the importance of initialization and momentum in deep learning*, in Proceedings of the 30th International Conference on Machine Learning (ICML-13), S. Dasgupta and D. Mcallester, eds., Atlanta, GA, 2013, pp. 1139–1147.

[27] Y. S. Tan and R. Vershynin, *Phase retrieval via randomized Kaczmarz: theoretical guarantees*, Inf. Inference, 8 (2019), pp. 97–123.

[28] 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.

[29] V. V. Vasin, *Iterative methods for solving ill-posed problems with a priori information in Hilbert spaces*, Zh. Vychisl. Mat. i Mat. Fiz., 28 (1988), pp. 971–980, 1117.

[30] G. M. Vaĭnikko and A. Y. Veretennikov, *Iteration Procedures in Ill-posed Problems*, "Nauka", Moscow, 1986.

[31] Y. Ying and M. Pontil, *Online gradient descent learning algorithms*, Found. Comput. Math., 8 (2008), pp. 561–596.

[32] T. Zhang, *Solving large scale linear prediction problems using stochastic gradient descent algorithms*, in Proceedings of of the Twenty First International Conference on Machine Learning, C. Brodley, ed., Banff, Alberta, Canada, 2004, pp. 919–926.