Electronic Research Archive

https://www.aimspress.com/journal/era

ERA, 34(10): 7779–7810.
DOI: 10.3934/era.2026334
Received: 31 May 2026
Revised: 23 August 2026
Accepted: 25 August 2026
Published: 09 September 2026

Research article

Block stochastic gradient descent for linear ill-posed problems in Hilbert spaces under the power-type source condition

Rong Wang[1], **De-Han Chen**[1,2,*], **Ting Cheng**[1,2] **and Daijun Jiang**[1,3]

[1] School of Mathematics and Statistics, Central China Normal University, Wuhan 430079, China
[2] Hubei Key Laboratory of Mathematical Sciences, Central China Normal University, Wuhan 430079, China
[3] Key Lab NAA–MOE, Central China Normal University, Wuhan 430079, China

* **Correspondence:** Email: dhchen@ccnu.edu.cn.

Abstract: In this paper, we consider linear ill-posed inverse problems in Hilbert spaces under the classical power-type source condition. We propose a block stochastic gradient descent (Block SGD) algorithm, which generalizes the classical SGD method by allowing updates on data blocks rather than single indices. Under the power-type source condition, we derive a novel mean squared error bound between the true solution and the regularized solution, in terms of the step size, the batch size, and the block number. In particular, our estimate shows that the stochastic error decreases as the batch size increases, which is confirmed by numerical experiments.

Keywords: stochastic gradient descent; regularization property; ill-posed problems; convergence rates; power-type source condition

1. Introduction

This paper studies linear ill-posed inverse problems in the sense of Hadamard. For such problems, the underlying operators are typically compact with non-closed range, so that the generalized inverse is unbounded and small data errors can be amplified arbitrarily much. Hence, without additional a priori information on the exact solution, no method, regularized or not, can yield an error estimate that is uniform over all admissible data and tends to zero with the noise level [1]. The classical remedy is a power-type source condition, which confines the exact solution to a relatively compact admissible set; it is this compactness of the admissible set that keeps the modulus of continuity finite [2] and thereby makes quantitative error estimates of optimal order available [1]. Source conditions of this power-type form are standard in the analysis of stochastic gradient descent (SGD)-type methods [3–6], and we adopt one of them as Assumption 2.2.

Consider the linear ill-posed problem of finding $x^\dagger \in \mathcal{X}$ such that

$$A_j x^\dagger = y_j \quad \forall j \in \{1, 2, \dots, l\}, \tag{1.1}$$

where $x^\dagger$ denotes the unknown true solution to be recovered and each $A_j : \mathcal{X} \to \mathcal{Y}_j$ is a bounded linear operator from a fixed Hilbert space $\mathcal{X}$ into a Hilbert space $\mathcal{Y}_j \subseteq \mathcal{Y}$ (viewed as mutually orthogonal subspaces of a common Hilbert space $\mathcal{Y}$). Let $Q_j : \mathcal{Y} \to \mathcal{Y}_j$ be the orthogonal projection onto $\mathcal{Y}_j$ and

$$A := \sum_{j=1}^{l} A_j. \tag{1.2}$$

The ill-posed problem can be reformulated as the following inverse problem:

$$y^\delta = Ax^\dagger + \xi, \tag{1.3}$$

where ξ represents the noise with noise level $\delta = \|\xi\|_{\mathcal{Y}}$, and $y^\dagger = Ax^\dagger$ is the exact (noise-free) data, and $y_j^\delta = Q_j y^\delta$ for $1 \le j \le l$. The ill-posed problem (1.1) arises in a broad range of applications including imaging and inverse problems with discrete data (see [3, 7]).

To solve such problems, stochastic gradient descent (SGD) and its variants—widely used and highly effective optimization methods in machine learning, especially for neural network training—have been adapted to the ill-posed setting (see [7–10]). In the block-structured setting introduced above, this paper proposes a natural variant, which at each iteration randomly selects one block of the data to update the approximation. More precisely, let r_i be the dimension of $\mathcal{Y}_i$. The Block SGD algorithm is defined as follows. At the k-th iteration, an index i_k is drawn uniformly at random from $\{1, 2, \dots, l\}$ and we choose a step size η_k. The approximate solution is then updated by

$$x_{k+1}^\delta = x_k^\delta - \frac{\eta_k}{r_{i_k}} A^* Q_{i_k} (Ax_k^\delta - y^\delta). \tag{1.4}$$

In the special case where $|r_i| \equiv 1$ (i.e., each block consists of a single index), the above algorithm simplifies to the classical SGD method (see [3]). If the dimension $|r_i|$ is fixed and the subspaces $\mathcal{Y}_i$ are randomly reformulated at each iteration, the algorithm becomes the so-called Batch SGD, where a random batch of indices is selected at each step. Thus, the Block SGD can be viewed as an interpolation between classical SGD and Batch SGD. On the one hand, it utilizes more information per iteration than classical SGD; on the other hand, unlike Batch SGD, it does not require random subspace reformulation at each step.

Remark 1.1. *Our Block SGD differs from classical SGD in that it updates using a block of data, whereas classical SGD uses only a single data point per iteration. Unlike mini-batch SGD, which uses equal-sized, resampled batches at every iteration to reduce gradient variance, our method uses a fixed random partition with possibly unequal block sizes to reduce computational cost per update. Compared with BCD (Block Coordinate Descent) methods [11], which exactly solve each selected block and are thus restricted to tensor-product operators, our Block SGD updates one random block via a single gradient step, avoiding costly solves and applying to broader Hilbert space operators.*

Since the studies of [3, 12], the regularization property of SGD-type algorithms has received a lot of attention and has been widely investigated and established. Specifically, for the standard SGD, a

polynomially decaying step size was studied in [3], while [12] focused on a constant step size. Later, the absence of saturation was proven in [13] if the initial step size of the schedule is sufficiently small and the forward operator satisfies some constructive assumption. For Batch SGD, [5] established the bounds for the root mean squared error in terms of the step size and the underlying smoothness in terms of general source conditions. Recently, the investigation of convergence of SGD-type algorithms was extended to Banach-space settings [12, 14, 15]. For more recent developments, we refer the reader to [4, 7, 16–19]. Zhang and Chen [20] introduced stochastic asymptotical regularization (SAR) for linear inverse problems in Hilbert spaces, extending classical asymptotical regularization by adding a Q-Wiener process. They proved order-optimal convergence rates under range-type source conditions and demonstrated that SAR enables uncertainty quantification of the estimated solution. Subsequently, Long and Zhang [21] extended SAR to nonlinear problems, establishing mean-square convergence and convergence rates under source conditions, and showing that SAR can escape local minima and identify multiple solutions. Although both SAR and Block SGD are stochastic regularization methods, their mechanisms are distinct. SAR obtains randomness from a continuous-time stochastic differential equation; our Block SGD, by contrast, introduces stochasticity through block-wise random data subsampling while retaining the classical discrete SGD structure. The two methods are therefore complementary. It is worth pointing out that all of these papers rely on the power-type source conditions (see [3–6]).

In this paper, under the power-type source condition (stated as Assumption 2.2 below), we establish a novel mean-squared-error bound between the true solution and the regularized solution x_{k+1}^{δ} for the Block SGD algorithm, in terms of the step size, the batch size, and the block number. In particular, our novel error estimate quantifies the effect of the batch size and block number, showing that larger batches improve the stochastic error. Thus, with a suitable batch size, Block SGD exhibits better numerical performance than the classical SGD. Our numerical experiments further support that the stochastic error decreases as the batch size increases, and Block SGD achieves better numerical performance in concrete examples.

The outline of this paper is as follows: Section 2 is devoted to the statement of the main results, whose proof is the main subject of Section 4. Several illustrative numerical examples are presented in Section 5, which compare the Block SGD with the classical SGD and Batch SGD, including a case in which the power-type source condition is not satisfied. Finally, the Appendix contains some technical lemmas needed for our analysis.

2. Main results

In this section, we present the two main results of this paper, which address the plain convergence and convergence rates of the Block SGD algorithm. Let us first give the precise definition of Block SGD, and some settings for it.

Let us denote by $\{v_1, \cdots, v_n\}$ any (fixed) orthogonal basis in $\mathcal{Y}$, and consider the one-dimensional projection

$$Q_i := v_i \oplus v_i \quad \forall 1 \leq i \leq n.$$

Let the j-th subspace $\mathcal{Y}_j$ be indexed by some sub-index set $\mathcal{J}_i \subset \{1, 2, \cdots, n\}$ such that

$$\mathcal{J}_i \cap \mathcal{J}_j = \phi \quad \forall i \neq j \quad \text{and} \quad \bigcup_{j=1}^{l} \mathcal{J}_j = \{1, 2, \cdots, n\}.$$

Let us define the size of $\mathcal{J}_i$, and the maximum and minimum size as follows:

$$r_i := |\mathcal{J}_i| \quad \forall\, 1 \leq i \leq l, \quad \bar{r} := \max_{1\leq i\leq l}\{r_i\}, \quad \underline{r} := \min_{1\leq i\leq l}\{r_i\}. \tag{2.1}$$

Throughout this paper, since n is usually very large, for the sake of simplicity, we may assume that

$$1 \leq \frac{\bar{r}}{\underline{r}} \leq 2. \tag{2.2}$$

For $1 \leq j \leq l$ and the index set $\mathcal{J}_j$, the corresponding orthogonal projection is defined by

$$Q_j := \sum_{i\in\mathcal{J}_j} Q_i.$$

Here, Q_i denotes the projection onto the one-dimensional subspace spanned by v_i, while Q_j denotes the projection onto the j-th block subspace $\mathcal{Y}_j := \text{span}\{v_i : i \in \mathcal{J}_j\}$. With all the settings above, we can now present the precise Block SGD algorithm under the classical power-type source condition (Assumption 2.2) as follows:

Algorithm 1 Block stochastic gradient descent (Block SGD) for inverse problems

Require: Initial guess $x_1 \in \mathcal{X}$, sub-indexes $\{i\}_{i=1}^{l}$, step sizes $\{\eta_k\}_{k\geq 1} \subset \mathbb{R}^+$, batch size $r_i = |\mathcal{J}_i|$, noisy data y^δ.

1: **for** $k = 1, 2, \ldots$ **do**
2: Draw i_k uniformly at random from $\{1, \ldots, l\}$.
3: The iterate x_{k+1}^δ is then updated according to (1.4), where Q_i is the orthogonal projection onto $\mathcal{Y}_{i_k}$.
4: **end for**

Before proceeding to the convergence analysis, we introduce the necessary probability framework [22]. Let $\{i_k\}_{k\geq 1}$ be the sequence of random block indices (chosen uniformly with replacement) used in the Block SGD update (1.4). We denote by

$$\mathcal{F}_k := \sigma(i_1, i_2, \ldots, i_k)$$

the σ-algebra generated by the random indices up to the k-th iteration, with the convention that $\mathcal{F}_0 = \{\emptyset, \Omega\}$. All expectations $\mathbb{E}[\cdot]$ and conditional probabilities $\mathbb{P}(\cdot\,|\,\mathcal{F}_k)$ considered in the sequel are with respect to this filtration. Consequently, the stochastic iterate x_k^δ is $\mathcal{F}_{k-1}$-measurable, i.e., it depends only on the randomness realized before the k-th step.

Our main results rely on the following two necessary standing assumptions: the step-size condition and the source condition.

Assumption 2.1 (Step-size condition)**.** *The step size is selected according to*

$$\eta_j = c_0 j^{-\alpha} \quad \text{for all } j \in \mathbb{N}.$$

Here, $\alpha \in (0,1)$ *and the constant* c_0 *satisfies* $c_0 \max\limits_{1\le j\le l} \|A_j\|^2_{L(\mathcal{X},\mathcal{Y})} \le 1$.

Assumption 2.2 (Source condition)**.** *There exist a constant* $p \ge 0$ *and a non-trivial element* $w \in \mathcal{X}$ *such that*

$$x^\dagger - x_1 = \bar{B}^p w, \tag{2.3}$$

where

$$\bar{B} := \frac{1}{l}\sum_{j=1}^{l} \frac{1}{r_j} A^* Q_j A$$

is a symmetric, positive semidefinite operator on $\mathcal{X}$*, and* $\bar{B}^p$ *denotes the fractional power defined by spectral decomposition.*

Remark 2.3. *We pause here to discuss the three main assumptions and their practical implications.*

(i) ***(Step-size condition)*** *The condition* $c_0 \max_{1\le j\le l} \|A_j\|^2 \le 1$ *is standard in stochastic approximation theory. It can always be met by scaling the operators or choosing* c_0 *sufficiently small. The polynomial decay* $\eta_j = c_0 j^{-\alpha}$ *is common for deriving convergence rates in the SGD literature [3, 5, 13].*

(ii) ***(Source condition and spectral structure)*** *The operator* $\bar{B}$ *is symmetric and positive semidefinite on* $\mathcal{X}$*, provided that each* Q_j *is orthogonal. In particular, for discrete pointwise measurements at* $\{\xi_i\}_{i=1}^N$*, the measurement space* $\mathcal{Y}$ *is isometric to* $\mathbb{R}^N$ *with the Euclidean inner product; hence each measurement corresponds to a coordinate direction, making them mutually orthogonal.*

Moreover, Lemma 4.3 shows that Assumption 2.2 is equivalent to the classical smoothness condition (4.11) *that has been widely used in regularization theory and SGD theory (see [3–5, 13, 18, 20]). Such a source condition holds trivially if* $x^\dagger - x_1 \in \mathcal{R}(A^*)$*, which corresponds to the smoothness of* $x^\dagger - x_1$ *with respect to the forward problem. Subsection 5.1 examines a concrete elliptic inverse problem in which the validity of this source condition is established.*

Power-type source conditions of this kind are standard in the regularization literature; see the classical monographs [1, 23].

Under these assumptions, we first obtain the convergence result.

Theorem 2.4. *Let Assumption 2.1 be fulfilled. If the stopping index* $k(\delta)$ *satisfies*

$$\lim_{\delta\to 0^+} k(\delta) = \infty \quad \text{and} \quad \lim_{\delta\to 0^+} k(\delta)^{\frac{1-\alpha}{2}}\delta = 0,$$

then the iterate $x^\delta_{k(\delta)}$ *satisfies the convergence*

$$\lim_{\delta\to 0^+} \mathbb{E}\|x^\delta_{k(\delta)} - x^\dagger\|^2_{\mathcal{X}} = 0.$$

Theorem 2.4 establishes a basic convergence result. Under the source condition (2.3), we can further obtain the following error estimate.

Theorem 2.5. *Let Assumptions 2.1 and 2.2 be fulfilled, and let $\underline{r}$ and l be given by* (2.1)*. Then, we have*

$$\mathbb{E}[\|\bar{B}^s(x_{k+1}^\delta - \mathbb{E}[x_{k+1}^\delta])\|_{\mathcal{X}}^2] \le \frac{c}{\underline{r}^2} k^{-\min((1+2s)(1-\alpha),\, 3\alpha,\, 1-2s\alpha)} \ln k + \frac{c}{\underline{r}} \delta^2 k^{-\min((1-2s)\alpha,(2s+1)(1-\alpha))},$$

$$\mathbb{E}\|x_{k+1}^\delta - x^\dagger\|_{\mathcal{X}}^2 \le c k^{-2p(1-\alpha)} + \frac{c\delta^2}{l\underline{r}} k^{1-\alpha} + \frac{c}{\underline{r}^2} k^{-\min(3\alpha,1-\alpha)} \ln k + \frac{c}{\underline{r}} \delta^2 k^{-\min(\alpha,1-\alpha)},$$

where $c > 0$ depends on α, p, $\|w\|_{\mathcal{X}}$, and $\|A\|_{\mathrm{L}(\mathcal{X},\mathcal{Y})}$.

Remark 2.6. *Compared to the previous work in [3], our error estimate quantifies the effect of the batch size via $\underline{r}$ and l, showing that larger batches improve the stochastic error $\mathbb{E}[\|x_{k+1}^\delta - \mathbb{E}[x_{k+1}^\delta]\|_{\mathcal{X}}^2]$. Thus, the Block SGD may outperform standard SGD by leveraging the block updates to reduce stochastic error. The corresponding numerical experiments will be included in Section 5.*

Next, let us discuss the convergence result of Block SGD under a posteriori stopping rules. Of primary concern is the discrepancy principle, which determines the stopping index $k(\delta)$ via

$$k(\delta) := \min\{k \in \mathbb{N} \mid \|Ax_k^\delta - y^\delta\|_{\mathcal{X}} \le \tau\delta\}, \tag{2.4}$$

where $\tau > 1$ is a fixed constant. The result below shows that the discrepancy principle (2.4) is a valid a posteriori stopping rule for the Block SGD.

Theorem 2.7. *Let Assumptions 2.1 and 2.2 be fulfilled, and $k(\delta)$ is determined by the discrepancy principle* (2.4)*. Then, the following assertions are valid.*

(i) *For all $0 < r < 1$ and $\tau > \tau^* > 1$, with $c = \left(\frac{\tau^*-1}{\sqrt{l\bar{r}}\, c_p}\right)^{-\frac{2}{(1-\alpha)(\min\{2p,r\}+1)}} + 2$, we have*

$$\mathbb{P}\left(k(\delta) \le c\, \delta^{-\frac{2}{(1-\alpha)(\min\{2p,r\}+1)}} + 2\right) \to 1 \quad as \quad \delta \to 0^+,$$

where $c_p = \left(\frac{(p+\frac{1}{2})(1-\alpha)}{c_0 e\,(2^{1-\alpha}-1)}\right)^{p+\frac{1}{2}} \|w\|$.

(ii) *For all $\epsilon > 0$, it holds that*

$$\mathbb{P}(\|x_{k(\delta)}^\delta - x^\dagger\|_{\mathcal{X}} \ge \varepsilon) \to 0 \quad as \quad \delta \to 0^\dagger,$$

where $(x_k^\delta)_{k\in\mathbb{N}}$ are Block SGD iterates independent of $k(\delta)$ but using the same data (y^δ, δ).

Remark 2.8. *The present work focuses on linear inverse problems. A natural future direction is to extend Block SGD to nonlinear settings, e.g., by combining it with Gauss–Newton or Landweber-type linearization, or by adopting stochastic variance-reduced methods adapted to blocks. Recent advances in SGD for nonlinear inverse problems [17, 24] reveal several major challenges.*

First, the tangential cone and range invariance conditions must be reformulated for block-structured updates, requiring per-block verification and aggregation of estimates with constants that may depend on the block size and number of blocks. Second, to keep iterates within the smoothness region under random block updates, the step size should be scaled by the maximum block size (i.e., $\eta_k \lesssim 1/\bar{r}$), while still preserving variance reduction. Third, the bias and variance terms become coupled in the nonlinear setting [17]; variance reduction or local linearization can help decouple the recursions, enabling a block-adapted bias–variance analysis. These tailored implementations, along with adaptive stopping rules, constitute the core of our planned future investigation for nonlinear problems.

3. Technical results

In this section, we collect several technical lemmas that are used throughout the main body of the paper. These include elementary inequalities, estimates for certain sums and integrals arising from the step-size condition, and some basic properties of conditional expectations and source conditions in Hilbert spaces.

3.1. Some technical inequalities

We begin with a basic estimate for the step sizes $\eta_i = c_0 i^{-\alpha}$, which follows from standard integral comparisons.

Lemma 3.1. *For the choice $\eta_i = c_0 i^{-\alpha}$, $\alpha \in [0, 1)$, for any $1 \le i \le k$, it holds that*

$$(1-\alpha)^{-1} c_0((k+1)^{1-\alpha} - j^{1-\alpha}) \ge \sum_{i=j+1}^{k} \eta_i \ge (2^{1-\alpha} - 1)(1-\alpha)^{-1} c_0(k^{1-\alpha} - j^{1-\alpha}).$$

Proof. The right inequality has been proven in [3]. Since $\eta_i = c_0 i^{-\alpha}$, $f(x) = x^{-\alpha}$ is decreasing on $[1, \infty]$, by the monotonicity of integration, for $j \ge 2$, we have

$$\int_j^{j+1} x^{-\alpha} dx \le j^{-\alpha} \le \int_{j-1}^{j} x^{-\alpha} dx.$$

Summing the right-hand inequality for $j = 2, \cdots, k$ and treating the term $j = 1$ separately yields

$$\sum_{i=j+1}^{k} \eta_i = \sum_{i=j}^{k} c_0 i^{-\alpha} \le c_0 \sum_{i=j+1}^{k} \int_{i-1}^{i} x^{-\alpha} dx \le \frac{c_0}{1-\alpha}((k+1)^{1-\alpha} - j^{1-\alpha}).$$

□

The next lemma provides a simple but useful inequality relating the linear function $1 - x$ to an exponential bound on a bounded interval.

Lemma 3.2. *For every $a \in (0, 1)$, there is a constant $c > 0$, such that*

$$1 - x \ge e^{-cx} \quad \forall x \in (0, a).$$

Proof. Let $f(x) = e^{-cx}$ with $c = -\frac{\ln(1-a)}{a}$. Then, $f''(x) = c^2 e^{-cx} > 0$, so f is convex on $(0, a)$. By the convexity of f, it follows that the secant line connecting $(0, f(0))$ and $(a, f(a))$ lies above the graph. That is,

$$f(x) \le \left(1 - \frac{x}{a}\right) f(0) + \frac{x}{a} f(a) \quad \forall x \in (0, a).$$

Substituting $f(0) = 1$ and $f(a) = e^{-ca}$, and using $c = -\frac{\ln(1-a)}{a}$, we obtain

$$e^{-cx} \le \left(1 - \frac{x}{a}\right) + \frac{x}{a} e^{-ca} = 1 - \frac{x}{a}(1 - e^{-ca}) = 1 - x.$$

□

The following lemma provides a simple and useful bound for $\|\sum_{j=1}^{k}\eta_j\mathcal{P}_{j+1}^{k}(S)S^s\|_{\mathcal{L}(\mathcal{X})}$ in terms of the total step size $(\sum_{i=1}^{k}\eta_i)^{1-s}$, where $s\in[0,1]$.

Lemma 3.3. *Suppose that $s\in[0,1]$. Let $j,k\in\mathbb{N}$ with $j<k$, and let S be any symmetric positive semidefinite operator with step sizes $\eta_i\in(0,a\|S\|_{\mathcal{L}(\mathcal{X})}^{-1}]$ with some $a\in(0,1)$. Then, we have that*

$$\|\sum_{j=1}^{k}\eta_j\mathcal{P}_{j+1}^{k}(S)S^s\|_{\text{Ł}(\mathcal{X})}\le c_a(\sum_{i=1}^{k}\eta_i)^{1-s},$$

where $c_a>0$ depends only on a and s.

Proof. (Step 1) Let $\{E_\lambda\}$ be the spectral decomposition of S. Then, we have

$$\sum_{j=1}^{k}\eta_j\mathcal{P}_{j+1}^{k}(S)S^s=\int_0^{\|S\|_{\text{Ł}(\mathcal{X})}}f(\lambda)dE_\lambda,$$

where the function $f:(0,\|S\|_{\mathcal{L}(\mathcal{X})}]\to\mathbb{R}$ is given by

$$f(\lambda):=\lambda^s(\sum_{j=1}^{k}\eta_jP_j^k)\quad\forall\lambda\in(0,\|S\|_{\text{Ł}(\mathcal{X})}].$$

Here, we set

$$P_j^k(\lambda):=\prod_{i=j}^{k}(1-\eta_i\lambda),\quad 1\le j\le k,\quad\text{with the convention}\quad P_{k+1}^k\equiv 1.\tag{3.1}$$

In this step, we aim at showing that

$$\frac{f(\lambda)}{\lambda^s}=\frac{1}{\lambda}\Big(1-\prod_{i=1}^{k}(1-\eta_i\lambda)\Big)\qquad\forall\lambda\in(0,\|S\|_{\text{Ł}(\mathcal{X})}].\tag{3.2}$$

Indeed, for every $1\le j\le k$, a direct computation yields

$$P_j^k=(1-\eta_j\lambda)P_{j+1}^k=P_{j+1}^k-\eta_j\lambda P_{j+1}^k.$$

Rearranging this and dividing by λ gives

$$\eta_jP_{j+1}^k=\frac{1}{\lambda}(P_{j+1}^k-P_j^k).$$

Summing this equality from $j=1$ to k, we find that

$$\sum_{j=1}^{k}\eta_jP_{j+1}^k=\frac{1}{\lambda}\sum_{j=1}^{k}(P_{j+1}^k-P_j^k)=\frac{1}{\lambda}(P_{k+1}^k-P_1^k).$$

By the definition that $P_{k+1}^k=1$, it follows that (3.2) holds.

(Step 2) Now, we give an estimate of f as follows:

$$\sup_{\lambda\in(0,\|S\|_{\mathrm{L}(X)}]}|f(\lambda)|\le\sqrt{\frac{1}{a}\ln(\frac{1}{1-a})}\max_{x\in(0,\infty)}(x^{-1+s}(1-e^{-x}))(\sum_{i=1}^{k}\eta_i)^{1-s}.$$

Since $\eta_i\|S\|_{\mathrm{L}(X)}<a$ for all $1\le i\le k$, we infer from Lemma 3.2 that

$$\prod_{i=1}^{k}(1-\eta_i\lambda)\ge e^{-c_a\lambda\sum_{i=1}^{k}\eta_i}=e^{-c_a\lambda S_k}\quad\text{with}\quad S_k:=\sum_{i=1}^{k}\eta_i,$$

where $c_a=-\ln(1-a)/a$. This, together with the fact that $0\le\prod_{i=1}^{k}(1-\eta_i\lambda)\le 1$, implies

$$0\le f(\lambda)\le\lambda^{-(1-s)}(1-e^{-c\lambda S_k}).$$

Writing $x=c_a\lambda S_k$, the decay behavior of $t\mapsto t^{s-1}(1-e^{-t})$ at $t=0$ and $t=\infty$ shows that

$$|f(\lambda)|\le(c_aS_k)^{1-s}x^{-1+s}(1-e^{-x})=\max_{x\in(0,\infty)}(x^{-1+s}(1-e^{-x}))c_a^sS_k^{1-s}\quad\forall\lambda\in[0,\|S\|_{\mathrm{L}(X)}].$$

□

For the proof of Lemmas 3.4 and 3.5, let us refer to [3].

Lemma 3.4. *Let $\eta_j=c_0j^{-\alpha}$ with $\alpha\in(0,1)$, $\beta\in[0,1]$, and $r\ge 0$. Then, it holds that*

$$\sum_{j=1}^{[\frac{k}{2}]}\frac{\eta_j^2}{(\Sigma_{i=j+1}\eta_i)^r}j^{-\beta}\le c_{\alpha,\beta,r}k^{-r(1-\alpha)+\max(0,1-2\alpha-\beta)},\tag{3.3}$$

$$\sum_{j=[\frac{k}{2}]+1}^{k}\frac{\eta_j^2}{(\Sigma_{i=j+1}\eta_i)^r}j^{-\beta}\le c'_{\alpha,\beta,r}k^{-((2-r)\alpha+\beta)+\max(0,1-r)},\tag{3.4}$$

where

$$c_{\alpha,\beta,r}=c_0^{2-r}\begin{cases}2^r(2\alpha+\beta-1)^{-1}, & 2\alpha+\beta>1,\\ 2, & 2\alpha+\beta=1,\\ 2^{r-1+2\alpha+\beta}(1-2\alpha-\beta)^{-1}, & 2\alpha+\beta<1,\end{cases}$$

$$c'_{\alpha,\beta,r}=2^{2\alpha+\beta}c_0^{2-r}=\begin{cases}(r-1)^{-1}, & r>1,\\ 1, & r=1,\\ 2^{r-1}(1-r)^{-1}, & r<1.\end{cases}$$

Lemma 3.5. *Let $\eta_j=c_0j^{-\alpha}$, $\alpha\in(0,1)$. Suppose that a non-negative sequence $\{b_j\}_{j=1}^{\infty}$, $a_1\ge 0$ and c_i>0, $\{a_j\}_{j=1}^{\infty}$ satisfies*

$$a_{k+1}=c_1\sum_{j=1}^{k-1}\frac{\eta_j^2}{\sum_{i=j+1}^{k}\eta_i}a_j+c_2k^{-2\alpha}a_k+b_k.$$

If b_j is nondecreasing, then for some $c(\alpha,c_i)$ dependent of α amd c_i, it holds that

$$a_{k+1}\le c(\alpha,c_i)k^{-\min(\alpha,1-\alpha)}\ln k+2b_k.$$

3.2. Stochastic analysis

We now recall several fundamental properties of conditional expectations for Hilbert-space-valued random variables (see [22, 25]), which are used extensively in the stochastic error analysis.

Lemma 3.6 (Properties of the conditional expectation for Hilbert-space-valued random variables)**.** *Let* $(\Omega, \mathcal{F}, \mathbf{P})$ *be a probability space and* $(H, (\cdot,\cdot))$ *a separable Hilbert space. Let* $X : \Omega \to H$ *be strongly measurable and satisfy* $\mathbb{E}\|X\|_H < \infty$*. Let* $\mathcal{G} \subset \mathcal{F}$ *be a sub* σ*-algebra. Then the conditional expectation* $\mathbb{E}[X \mid \mathcal{G}]$ *exists, is* $\mathcal{G}$ *measurable, i.e.,* $\mathbb{E}[\|\mathbb{E}[X \mid \mathcal{G}]\|] < \infty$*, and satisfies:*

- ***(Linearity)*** *For any* $\lambda \in \mathbb{R}$ *and any* $Y : \Omega \to H$ *with* $\mathbb{E}[\|Y\|] < \infty$*,*

$$\mathbb{E}[\lambda X + Y \mid \mathcal{G}] = \lambda \mathbb{E}[X \mid \mathcal{G}] + \mathbb{E}[Y \mid \mathcal{G}] \quad a.s.$$

- ***(Taking out what is known)*** *If* $Z : \Omega \to \mathbb{R}$ *is* $\mathcal{G}$ *measurable and bounded (or such that* ZX *is integrable), then* $\mathbb{E}[ZX \mid \mathcal{G}] = Z\,\mathbb{E}[X \mid \mathcal{G}]$*.*

- ***(Tower property)*** *If* $\mathcal{H} \subset \mathcal{G}$ *is another sub* σ*-algebra, then we have*

$$\mathbb{E}\big[\mathbb{E}[X \mid \mathcal{G}] \mid \mathcal{H}\big] = \mathbb{E}[X \mid \mathcal{H}].$$

- ***(Conditional variance)*** *Assume* $\mathbb{E}[\|X\|^2] < \infty$*. Then, we have*

$$\mathrm{Var}(X \mid \mathcal{G}) := \mathbb{E}[\|X - \mathbb{E}[X \mid \mathcal{G}]\|^2 \mid \mathcal{G}] = \mathbb{E}[\|X\|^2 \mid \mathcal{G}] - \big\|\mathbb{E}[X \mid \mathcal{G}]\big\|^2.$$

- ***(Independence)*** *If* X *is independent of* $\mathcal{G}$*, then* $\mathbb{E}[X \mid \mathcal{G}] = \mathbb{E}[X]$*.*

4. Proof of main results

This section is devoted to proving Theorems 2.4 and 2.5. To this end, we decompose the mean squared error (MSE) into three components: approximation error, propagation error, and stochastic error. More precisely, by combining the bias-variance decomposition with the triangle inequality, we obtain

$$\mathbb{E}\|x_{k+1}^\delta - x^\dagger\|_X^2 \le 2\underbrace{\|\mathbb{E}[x_{k+1}] - x^\dagger\|_X^2}_{\text{approximation error}} + 2\underbrace{\|\mathbb{E}[x_{k+1} - x_{k+1}^\delta]\|_X^2}_{\text{propagation error}} + \underbrace{\mathbb{E}\|\mathbb{E}[x_{k+1}^\delta] - x_{k+1}^\delta\|_X^2}_{\text{stochastic error}}, \tag{4.1}$$

where x_k is the random iterate for exact data $y^\dagger$. This reduces the estimation of the MSE to the separate estimation of these three parts. Lemmas 4.4 and 4.5 provide bounds for the approximation error and the propagation error, respectively, while Lemma 4.6 bounds the stochastic error. Lemma 4.7 supplies the necessary estimate for proving convergence of the MSE, after which the proof of the main result is complete. Compared with the argument in [3], our approach yields novel improvements on the propagation error and the stochastic error. Throughout this section, $c > 0$ denotes a generic constant depending on α, p, $\|w\|_\mathcal{X}$, and $\|A\|_{\mathcal{L}(\mathcal{X},\mathcal{Y})}$.

4.1. Approximation and propagation errors

For our subsequent analysis, we first introduce auxiliary iterations. Define

$$e_k^\delta := x_k^\delta - x^\dagger \quad \text{and} \quad e_k := x_k - x^\dagger$$

as the errors for the SGD iterates x_k^δ (corresponding to noisy data y^δ) and x_k (corresponding to exact data $y^\dagger$), respectively. These errors satisfy the following recursions:

$$e_{k+1} = e_k - \frac{\eta_k}{r_{i_k}} A^* Q_{i_k} A e_k, \tag{4.2}$$

$$e_{k+1}^\delta = e_k^\delta - \frac{\eta_k}{r_{i_k}} A^* Q_{i_k} A e_k^\delta + \frac{\eta_k}{r_{i_k}} A^* Q_{i_k} \xi. \tag{4.3}$$

Next, by a direct computation, we observe that

$$\bar{B} = \mathbb{E}\Big[\frac{1}{r_i} A^* Q_i A\Big] \quad \text{and} \quad \bar{A}^* = \mathbb{E}\Big[\frac{1}{r_i} A^* Q_i\Big],$$

where i follows the independent and identically distributed (i.i.d.) uniform sampling and

$$\bar{A} := \frac{1}{l} \sum_{i=1}^{l} \frac{1}{r_i} Q_i A.$$

Clearly, the identity $\bar{B} = \bar{A}^* A$ holds, and thus it is a symmetric positive semidefinite (SPD) operator. In addition, we have the following observation:

Lemma 4.1. *There exists a bounded operator $\bar{U} \in Ł(\mathcal{Y}, \mathcal{X})$ such that*

$$\bar{A}^* = \bar{B}^{\frac{1}{2}} \bar{U} \quad \text{with} \quad \|\bar{U}\|_{Ł(\mathcal{Y},\mathcal{X})} \le \frac{1}{\sqrt{l\underline{r}}}. \tag{4.4}$$

In addition, we have

$$A^* = \bar{B}^{\frac{1}{2}} V \quad \text{with} \quad \|V\|_{Ł(\mathcal{Y},\mathcal{X})} \le \sqrt{l\overline{r}}. \tag{4.5}$$

Proof. Consider the two operators $C \in Ł(\mathcal{X}, \mathcal{Y})$ and $\widetilde{Q} \in Ł(\mathcal{Y})$, given by

$$\widetilde{Q} := \frac{1}{\sqrt{l}} \sum_{j=1}^{l} \frac{1}{\sqrt{r_j}} Q_j \quad \text{and} \quad C := \widetilde{Q} A.$$

Obviously, $\widetilde{Q}$ is an SPD operator, and

$$\|\widetilde{Q}\|_{Ł(\mathcal{Y})} \le \frac{1}{\sqrt{l\underline{r}}} \quad \text{and} \quad \|Q^{-1}\|_{Ł(\mathcal{Y})} \le \sqrt{l\overline{r}}.$$

A direct computation yields

$$\bar{B} = C^* C \quad \text{and} \quad \bar{A}^* = (\widetilde{Q} A)^* \widetilde{Q} = C^* \widetilde{Q}, \tag{4.6}$$

which implies that

$$\|\bar{B}\|_{Ł(\mathcal{X})} \leq \|A\|_{Ł(\mathcal{X},\mathcal{Y})}\|A^*\|_{Ł(\mathcal{Y},\mathcal{X})}\|\widetilde{Q}^2\|_{Ł(X)} \leq \frac{1}{l\underline{r}}\|A\|^2_{Ł(\mathcal{X},\mathcal{Y})}. \tag{4.7}$$

Then, by polar decomposition (see [26]), it follows that there exists a unitary operator $U \in Ł(\mathcal{X},\mathcal{Y})$ such that

$$C^* = \bar{B}^{\frac{1}{2}}U^* \quad \text{and thus} \quad A^*Q^* = \bar{B}^{\frac{1}{2}}U^*, \tag{4.8}$$

which, together with the second identity of (4.6), implies that (4.4) holds with

$$\bar{U} := U^*\widetilde{Q} \quad \text{with} \quad \|\bar{U}\|_{Ł(\mathcal{Y},\mathcal{X})} \leq \|U^*\|_{Ł(\mathcal{Y},\mathcal{X})}\|\widetilde{Q}\|_{Ł(\mathcal{Y})} \leq \frac{1}{\sqrt{l\underline{r}}}.$$

On the other hand, by (4.8), it follows that

$$A^* = \bar{B}^{\frac{1}{2}}U^*\widetilde{Q}^{-1} \tag{4.9}$$

and thus

$$\|U^*\widetilde{Q}^{-1}\|_{Ł(\mathcal{Y},\mathcal{X})} \leq \sqrt{l\bar{r}}. \tag{4.10}$$

□

In the sequel, the following operator plays a central role:

$$\mathcal{P}_j^k(\bar{B}) := \begin{cases} \prod_{i=j}^k (I - \eta_i\bar{B}), & j \leq k, \\ I, & j = k+1. \end{cases}$$

To estimate the operator $\mathcal{P}_j^k(\bar{B})$, we will utilize the following lemma (see [8, Lemma 15]).

Lemma 4.2. *Let $j, k \in \mathbb{N}$ with $j < k$, and let S be any symmetric positive semidefinite operator in $\mathcal{X}$ with step sizes $\eta_i \in (0, \|S\|^{-1}_{Ł(\mathcal{X})}]$. Then for any $p > 0$,*

$$\Big\|\prod_{i=j}^{k}(I - \eta_i S)S^p\Big\|_{Ł(\mathcal{X})} \leq \frac{p^p}{e^p \cdot (\sum_{i=j}^k \eta_i)^p}.$$

Using this auxiliary lemma, let us begin by bounding the approximation error. As a preparation, we present the following lemma, which establishes the equivalence between the source condition (2.3) and a new condition governed by $B := A^*A$:

$$\exists\, w' \quad \text{such that} \quad x^\dagger - x_1 = B^p w'. \tag{4.11}$$

Lemma 4.3. *The source condition* (2.3) *holds if and only if* (4.11) *holds. Furthermore, under the assumption that $p \in (0, 1/2]$, there exists a constant $C > 0$, independent of w, such that*

$$C^{-1}\|w'\|_{\mathcal{X}} \leq \|w\|_{\mathcal{X}} \leq C\|w'\|_{\mathcal{X}}.$$

Proof. Since A^*Q_iA is symmetric positive definite (SPD) for each $1 \le i \le l$, and $r_i \ge \underline{r} > 0$, we have $\frac{1}{r_i} \le \frac{1}{\underline{r}}$. By the order-preserving property of positive operators, it follows that

$$\bar{B} = \frac{1}{l}\sum_{i=1}^{l}\frac{1}{r_i}A^*Q_iA \;\le\; \frac{1}{l}\sum_{i=1}^{l}\frac{1}{\underline{r}}A^*Q_iA = \frac{1}{l\underline{r}}A^*\Big(\sum_{i=1}^{l}Q_i\Big)A = \frac{n}{l\underline{r}}B,$$

where the inequality $\le$ is understood in the partial order: for two self-adjoint operators X and Y, $X \le Y$ means $Y - X$ is positive semidefinite (see [27]). Similarly, we have

$$\bar{B} \le \frac{n}{l\bar{r}}B \quad \text{and thus} \quad \frac{n}{l\underline{r}}B \ge \bar{B} \ge \frac{n}{l\bar{r}}B, \tag{4.12}$$

which implies $\bar{B}$ and B have identical kernels. On the other hand, by spectral mapping, it follows that $\mathrm{Ker}(B^p) = \mathrm{Ker}(B)$ and $\mathrm{Ker}(\bar{B}^p) = \mathrm{Ker}(\bar{B})$. Therefore, we can conclude that $\mathrm{Ker}(B^p) = \mathrm{Ker}(B^p)$, which, together with the closed range theorem, implies that they share the same range. This guarantees the equivalence of the existence of w' and w. On the other hand, by the Heinz inequality and the order relations (4.12), it follows that

$$(\frac{n}{l\underline{r}})^s B^s \ge \bar{B}^s \ge (\frac{n}{l\bar{r}})^s B^s \quad \forall s \in [0,1], \tag{4.13}$$

which is equivalent to

$$(\frac{n}{l\underline{r}})^s \|B^{s/2}x\|_{\mathcal{X}}^2 \ge \|\bar{B}^{s/2}x\|_{\mathcal{X}}^2 \ge (\frac{n}{l\bar{r}})^s \|B^{s/2}x\|_{\mathcal{X}}^2 \quad \forall x \in \mathcal{X}. \tag{4.14}$$

On the other hand, the conditions (2.3) and (4.11) imply that

$$\|B^p w'\|_{\mathcal{X}}^2 = \|\bar{B}^p w\|_{\mathcal{X}}^2, \tag{4.15}$$

which, together with (4.14), completes the proof.

□

Lemma 4.4. *Let Assumptions 2.1 and 2.2 be fulfilled. It holds that*

$$\|\bar{B}^s\mathbb{E}[x_{k+1} - x^\dagger]\|_{\mathcal{X}} \le c_p k^{-(p+s)(1-\alpha)},$$

where $c_p := \left(\frac{(p+s)(1-\alpha)}{e(2^{1-\alpha}-1)c_0}\right)^{p+s}\|w\|_{\mathcal{X}}$.

Proof. Since e_k is $\mathcal{F}_{k-1}$-measurable, taking conditional expectation in (4.2) with respect to $\mathcal{F}_{k-1}$ gives

$$\mathbb{E}[e_{k+1} \mid \mathcal{F}_{k-1}] = (I - \eta_k\bar{B})e_k \quad \forall k \ge 1.$$

By subsequently taking the full expectation, we obtain the recursion for the mean error:

$$\mathbb{E}[e_{k+1}] = (I - \eta_k\bar{B})\mathbb{E}[e_k] \quad \forall k \ge 1.$$

Applying the above recursion gives

$$\mathbb{E}[e_{k+1}] = \prod_{i=1}^{k}(I - \eta_i\bar{B})e_1 = \mathcal{P}_1^k(\bar{B})e_1 \qquad \forall k \in \mathbb{N}. \tag{4.16}$$

From the Assumption 2.2, and Lemmas 4.2 and 3.1, we have

$$\begin{aligned}\|\bar{B}^s\mathbb{E}e_{k+1}\|_{\mathcal{X}} &= \|\bar{B}^s\mathcal{P}_1^k(\bar{B})\bar{B}^p w\|_{\mathcal{X}} \leq \|\bar{B}^{p+s}\mathcal{P}_1^k(\bar{B})\|_{\mathrm{L}(\mathcal{X})}\|w\|_{\mathcal{X}} \\ &\leq \frac{(p+s)^{p+s}e^{-p+s}}{(\sum_{j=1}^k \eta_j)^{p+s}}\|w\|_{\mathcal{X}} \leq \Big(\frac{(p+s)(1-\alpha)}{e(2^{1-\alpha}-1)c_0}\Big)^{p+s}\|w\|_{\mathcal{X}}k^{-p(1-\alpha)} = c_p k^{-p(1-\alpha)}.\end{aligned}$$

□

In the next step, we derive a bound for the propagation error, which constitutes the second component in the bias-variance decomposition, which relies on the following novel estimate.

Lemma 4.5. *Let $s \in [0, 1/2]$ and let Assumption 2.1 be fulfilled. Then it holds that*

$$\|\bar{B}^s\mathbb{E}[x_{k+1} - x_{k+1}^\delta]\|_{\mathcal{X}} \leq c\frac{\delta}{\sqrt{l\underline{r}}}k^{(1-\alpha)(\frac{1}{2}-s)},$$

where c depends only on s, α, and c_0.

Proof. Let us write the propagation error at the k-th step:

$$\nu_k := \mathbb{E}[x_k^\delta - x_k] \quad \forall k \geq 1. \tag{4.17}$$

By the recursions (4.2) and (4.3), it follows that

$$x_{k+1}^\delta - x_{k+1} = x_k^\delta - x_k - \frac{\eta_k}{r_{i_k}}A^*Q_{i_k}A(x_k^\delta - x_k) + \frac{\eta_k}{r_{i_k}}A^*Q_{i_k}\xi. \tag{4.18}$$

Taking the conditional expectation on both sides of (4.18) and using the independence of i_k from $\mathcal{F}_{k-1}$, we obtain

$$\begin{aligned}\mathbb{E}[x_{k+1}^\delta - x_{k+1} \mid \mathcal{F}_{k-1}] &= x_k^\delta - x_k - \eta_k\mathbb{E}\left[\frac{A^*Q_{i_k}A}{r_{i_k}}\right](x_k^\delta - x_k) + \eta_k\mathbb{E}\left[\frac{A^*Q_{i_k}}{r_{i_k}}\xi\right] \\ &= x_k^\delta - x_k - \eta_k\bar{B}(x_k^\delta - x_k) + \eta_k\bar{A}^*\xi.\end{aligned}$$

Taking the full expectation of this identity and recalling (4.17), we obtain

$$\nu_{k+1} = \nu_k - \eta_k\bar{B}\nu_k + \eta_k\bar{A}^*\xi \quad \forall k \geq 1.$$

Applying the recursion repeatedly yields

$$\nu_{k+1} = \sum_{j=1}^k \eta_j\mathcal{P}_j^k(\bar{B})\bar{A}^*\xi \quad \forall k \geq 1. \tag{4.19}$$

Applying Lemma 4.1 to (4.19) and using the triangle inequality, we have

$$\begin{aligned}\|\bar{B}^s\nu_{k+1}\|_{\mathcal{X}} &\leq \left\|\sum_{j=1}^k \eta_j\mathcal{P}_j^k(\bar{B})\bar{B}^{\frac{1}{2}+s}\bar{U}\xi\right\|_{\mathcal{X}} \leq \|\sum_{j=1}^k \eta_j\mathcal{P}_j^k(\bar{B})\bar{B}^{\frac{1}{2}+s}\|_{\mathrm{L}(\mathcal{X})}\|\bar{U}\|_{\mathrm{L}(\mathcal{Y},\mathcal{X})}\|\xi\|_{\mathcal{Y}} \\ &\leq \frac{\delta}{\sqrt{l\underline{r}}}\|\sum_{j=1}^k \eta_j\mathcal{P}_j^k(\bar{B})\bar{B}^{\frac{1}{2}+s}\|_{\mathrm{L}(\mathcal{X})}.\end{aligned}$$

Then, an application of Lemma 3.3 yields

$$\|v_{k+1}\|_{\mathcal{X}} \leq c\frac{\delta}{\sqrt{l\underline{r}}}(\sum_{j=1}^{k}\eta_j)^{\frac{1}{2}-s} \leq c\frac{\delta}{\sqrt{l\underline{r}}}k^{(1-\alpha)(\frac{1}{2}-s)},$$

which completes the proof.

□

4.2. Stochastic error

The previous two lemmas together control the deterministic part of the error. It remains to analyze the stochastic error, which is the subject of Lemmas 4.6 and 4.7.

Lemma 4.6. *Let Assumption 2.2 be fulfilled. For the SGD iteration (1.2), it holds that*

$$\mathbb{E}[\|\bar{B}^s(x_{k+1}^\delta - \mathbb{E}[x_{k+1}^\delta])\|_{\mathcal{X}}^2] \leq \frac{c}{\underline{r}^2}k^{-\min((1+2s)(1-\alpha),\,3\alpha,\,1-2s\alpha)}\ln k + \frac{c}{\underline{r}}\delta^2 k^{-\min((1-2s)\alpha,(2s+1)(1-\alpha))}, \tag{4.20}$$

where c depends on p, α, and $\|A\|_{\mathrm{L}(\mathcal{X},\mathcal{Y})}^2$.

Proof. (Step 1) Let us write

$$z_k := x_k^\delta - \mathbb{E}[x_k^\delta]. \tag{4.21}$$

The aim of this step is to establish a recursive formula of z_k. Since x_k^δ is $\mathcal{F}_{k-1}$-measurable and i_k is independent of $\mathcal{F}_{k-1}$, we can infer from the definition of the iterate x_k^δ that

$$\begin{aligned}\mathbb{E}[x_{k+1}^\delta|\mathcal{F}_{k-1}] &= x_k^\delta - \eta_k(\mathbb{E}[\frac{1}{r_{i_k}}A^*Q_{i_k}A]x_k^\delta - \mathbb{E}[\frac{1}{r_{i_k}}A^*Q_{i_k}]y^\delta)\\ &= x_k^\delta - \eta_k(\bar{B}x_k^\delta - \bar{A}^*y^\delta).\end{aligned}$$

Taking the full expectation then yields

$$\mathbb{E}[x_{k+1}^\delta] = \mathbb{E}[x_k^\delta] - \eta_k(\bar{B}\mathbb{E}[x_k^\delta] - \bar{A}^*y^\delta).$$

Together with (1.4), this implies that z_k satisfies the recursive formula

$$\begin{cases} z_{k+1} = z_k - \eta_k\left[\left(\frac{A^*Q_{i_k}A}{r_{i_k}}x_k^\delta - \frac{A^*Q_{i_k}}{r_{i_k}}y^\delta\right) - \left(\bar{B}\mathbb{E}[x_k^\delta] - \bar{A}^*y^\delta\right)\right],\\ z_1 = 0.\end{cases}$$

Upon defining the iterative noise

$$d_k := (\bar{B}x_k^\delta - \bar{A}^*y^\delta) - (\frac{A^*Q_{i_k}A}{r_{i_k}}x_k^\delta - \frac{A^*Q_{i_k}}{r_{i_k}}y^\delta), \tag{4.22}$$

we can find that z_k satisfies

$$z_{k+1} = (I - \eta_k\bar{B})z_k + \eta_k d_k. \tag{4.23}$$

(Step 2) In this step, we will estimate a recursive formula of $V_{k+1} := \mathbb{E}\|z_{k+1}\|_{\mathcal{X}}^2$. To this end, we first investigate the property of iterative noise $\{d_k\}_{k\geq 1}$. Let us first show it is a *martingale difference* with

respect to the filtration $\{\mathcal{F}_k\}_{k\geq 1}$. Indeed, since x_j^δ is measurable with respect to $\mathcal{F}_{j-1}$ and i_j is independent of x_j^δ, the property of conditional expectation yields

$$\begin{aligned}\mathbb{E}[d_j|\mathcal{F}_{j-1}] &= (\bar{B}x_j^\delta - \bar{A}^*y^\delta) - \mathbb{E}[(\frac{A^*Q_{i_j}A}{r_{i_j}}x_j^\delta - \frac{A^*Q_{i_j}}{r_{i_j}}y^\delta) \mid \mathcal{F}_{j-1}]\\ &= (\bar{B}x_j^\delta - \bar{A}^*y^\delta) - (\mathbb{E}[\frac{A^*Q_{i_j}A}{r_{i_j}}]x_j^\delta - \mathbb{E}[\frac{A^*Q_{i_j}}{r_{i_j}}y^\delta])\\ &= (\bar{B}x_j^\delta - \bar{A}^*y^\delta) - (\bar{B}x_j^\delta - \bar{A}^*y^\delta) = 0.\end{aligned}$$

Let $j_1, j_2 \in \mathbb{N}$ with $j_1 \neq j_2$ and let $A_1, A_2 \in \text{Ł}(\mathcal{X})$. Without loss of generality, let us assume that $j_1 < j_2$. Since d_{j_1} is measurable with respect to $\mathcal{F}_{j_2-1}$, we have that

$$\mathbb{E}[(A_1d_{j_1}, A_2d_{j_2})_\mathcal{X}|\mathcal{F}_{l-1}] = (A_1d_{j_1}, \mathbb{E}[A_2d_{j_2}|\mathcal{F}_{j_2-1}])_\mathcal{X} = 0.$$

Then taking the full expectation yields

$$\mathbb{E}[(A_1d_{j_1}, A_2d_{j_2})_\mathcal{X}] = 0, \quad \forall j_1 \neq j_2. \tag{4.24}$$

Applying the recursion (4.23) repeatedly gives

$$z_{k+1} = \sum_{j=1}^{k} \eta_j \mathcal{P}_{j+1}^k(\bar{B})d_j. \tag{4.25}$$

Then, it follows from (4.24) that

$$\mathbb{E}\|\bar{B}^s z_{k+1}\|_\mathcal{X}^2 = \sum_{j=1}^{k} \eta_j^2 \mathbb{E}\|\bar{B}^s\mathcal{P}_{j+1}^k(\bar{B})d_j\|_\mathcal{X}^2. \tag{4.26}$$

By Lemma 4.1, we have

$$\|\bar{B}^s\mathcal{P}_{j+1}^k(\bar{B})A^*\|_{\text{Ł}(\mathcal{Y},\mathcal{X})}^2 = \|\bar{B}^s\mathcal{P}_{j+1}^k(\bar{B})\bar{B}^{\frac{1}{2}}V\|_{\text{Ł}(\mathcal{Y},\mathcal{X})}^2 \leq \bar{lr}\|\mathcal{P}_{j+1}^k(\bar{B})\bar{B}^{\frac{1}{2}+s}\|_{\text{Ł}(\mathcal{X})}^2. \tag{4.27}$$

In view of the identity

$$d_j = A^*N_j \quad \text{with} \quad N_j := (\bar{A}x_j^\delta - \bar{y}^\delta) - Q_{i_j}(\frac{A}{r_{i_j}}x_j^\delta - \frac{1}{r_{i_j}}y^\delta), \quad \bar{y}^\delta := \frac{1}{l}\sum_{i=1}^{l}\frac{1}{r_i}Q_iy^\delta.$$

and using (4.27), we can estimate

$$\begin{aligned}\mathbb{E}\|\bar{B}^s\mathcal{P}_{j+1}^k(\bar{B})d_j\|_\mathcal{X}^2 &= \mathbb{E}\|\bar{B}^s\mathcal{P}_{j+1}^k(\bar{B})A^*N_j\|_\mathcal{X}^2 \leq \|\bar{B}^s\mathcal{P}_{j+1}^k(\bar{B})A^*\|_{\text{Ł}(\mathcal{Y},\mathcal{X})}^2\mathbb{E}\|N_j\|_\mathcal{Y}^2\\ &\leq \bar{lr}\|\mathcal{P}_{j+1}^k(\bar{B})\bar{B}^{\frac{1}{2}+s}\|_{\text{Ł}(\mathcal{X})}^2\mathbb{E}\|N_j\|_\mathcal{Y}^2.\end{aligned} \tag{4.28}$$

Since x_j^δ is measurable to $\mathcal{F}_{j-1}$ and i_j is independent of $\mathcal{F}_{j-1}$, we have that

$$
\begin{aligned}
\mathbb{E}[\|N_j\|_{\mathcal{Y}}^2|\mathcal{F}_{j-1}] &= \mathbb{E}\left[\|\bar{A}x_j^\delta - \bar{y}^\delta - Q_{i_j}(\frac{A}{r_{i_j}}x_j^\delta - \frac{1}{r_{i_j}}y^\delta)\|_{\mathcal{Y}}^2|\mathcal{F}_{j-1}\right] \\
&= \mathbb{E}[\|\bar{A}x_j^\delta - \bar{y}^\delta\|_{\mathcal{Y}}^2|\mathcal{F}_{j-1}] - 2\mathbb{E}[(\bar{A}x_j^\delta - \bar{y}^\delta, Q_{i_j}(\frac{A}{r_{i_j}}x_j^\delta - \frac{1}{r_{i_j}}y^\delta))|\mathcal{F}_{j-1}] \\
&\quad + \mathbb{E}[\|Q_{i_j}(\frac{A}{r_{i_j}}x_j^\delta - \frac{1}{r_{i_j}}y^\delta)\|_{\mathcal{Y}}^2|\mathcal{F}_{j-1}] \\
&= \mathbb{E}[\|Q_{i_j}(\frac{A}{r_{i_j}}x_j^\delta - \frac{1}{r_{i_j}}y^\delta)\|_{\mathcal{Y}}^2|\mathcal{F}_{j-1}] - \|\bar{A}x_j^\delta - \bar{y}^\delta\|_{\mathcal{Y}}^2 \\
&\le \frac{1}{l}\|\sum_{i=1}^{l}\frac{1}{r_i}Q_i\|^2\|Ax_j^\delta - y^\delta\|_{\mathcal{Y}}^2 \le \frac{1}{l\underline{r}^2}\|Ax_j^\delta - y^\delta\|_{\mathcal{Y}}^2.
\end{aligned} \tag{4.29}
$$

Thus, by taking the full expectation, we obtain

$$
\mathbb{E}[\|N_j\|_{\mathcal{Y}}^2] \le \frac{1}{l\underline{r}^2}h_j, \quad h_j := \mathbb{E}\|Ax_j^\delta - y^\delta\|_{\mathcal{Y}}^2. \tag{4.30}
$$

Applying (4.30) to (4.28), and applying the resulting inequality to (4.27), we finally arrive at

$$
\mathbb{E}\|\bar{B}^s z_{k+1}\|_{\mathcal{X}}^2 \le \frac{1}{\underline{r}}\sum_{j=1}^{k}\eta_j^2\|\mathcal{P}_{j+1}^k(\bar{B})\bar{B}^{\frac{1}{2}+s}\|_{\mathrm{L}(\mathcal{X})}^2 h_j. \tag{4.31}
$$

(Step 3) In this step, we will give an estimate of $\{h_j\}_{j=1}^k$. By bias-variance decomposition and the triangle inequality, we have

$$
\begin{aligned}
h_{k+1} &= \mathbb{E}\|Ax_{k+1}^\delta - y^\delta\|_{\mathcal{Y}}^2 \\
&= \|A\mathbb{E}x_{k+1}^\delta - A\mathbb{E}x_{k+1} + A\mathbb{E}x_{k+1} - Ax^\dagger - \xi\|_{\mathcal{Y}}^2 + \mathbb{E}\|A\mathbb{E}x_{k+1}^\delta - Ax_{k+1}^\delta\|_{\mathcal{Y}}^2 \\
&\le 4\|A(\mathbb{E}x_{k+1} - x^\dagger)\|_{\mathcal{Y}}^2 + 4\|A\mathbb{E}[x_{k+1}^\delta - x_{k+1}]\|_{\mathcal{Y}}^2 + \mathbb{E}\|A\mathbb{E}x_{k+1}^\delta - Ax_{k+1}^\delta\|_{\mathcal{Y}}^2 + 2\|\xi\|_{\mathcal{Y}}^2 \\
&:= 4I_1 + 4I_2 + I_3 + I_4.
\end{aligned}
$$

By Lemma 4.4, it follows that

$$
I_1 = \|A\mathbb{E}e_{k+1}\|_{\mathcal{Y}}^2 \le c_p k^{-(2p+1)(1-\alpha)}. \tag{4.32}
$$

Meanwhile, by (2.2), and Lemmas 4.1 and 4.5 with $s = \frac{1}{2}$, it follows that

$$
I_2 = \|V^*\bar{B}^{\frac{1}{2}}\mathbb{E}[x_{k+1}^\delta - x_{k+1}]\|_{\mathcal{Y}}^2 \le \bar{lr}\|\bar{B}^{\frac{1}{2}}\mathbb{E}[x_{k+1}^\delta - x_{k+1}]\|_{\mathcal{X}}^2 \le c\delta^2. \tag{4.33}
$$

Next, by Lemma 4.1, and the reasoning leading to (4.26) and (4.28), we can find that

$$
\begin{aligned}
I_3 &\le \bar{lr}\mathbb{E}\|\bar{B}^{\frac{1}{2}}(\mathbb{E}[x_{k+1}^\delta] - x_{k+1}^\delta)\|_{\mathcal{X}}^2 \\
&= \bar{lr}\sum_{j=1}^{k}\eta_j^2\mathbb{E}\|\bar{B}^{\frac{1}{2}}\mathcal{P}_{j+1}^k(\bar{B})d_j\|_{\mathcal{X}}^2 \le \bar{lr}\sum_{j=1}^{k}\eta_j^2\mathbb{E}\|\bar{B}^{\frac{1}{2}}\mathcal{P}_{j+1}^k(\bar{B})d_j\|_{\mathcal{X}}^2 \\
&\le l\sum_{j=1}^{k}\eta_j^2\|\mathcal{P}_{j+1}^k(\bar{B})\bar{B}\|_{\mathrm{L}(\mathcal{X})}^2\mathbb{E}\|Ax_j^\delta - y^\delta\|_{\mathcal{Y}}^2,
\end{aligned}
$$

where (4.30) is used in the last inequality. Using the fact that $P_{k+1}^k(\bar{B}) = \mathrm{id}$, then

$$\begin{aligned} I_3 &\le l\sum_{j=1}^{k-1}\eta_j^2\|\bar{B}^{\frac{1}{2}}\bar{B}^{\frac{1}{2}}\mathcal{P}_{j+1}^k(\bar{B})\|_{\mathrm{L}(\mathcal{X})}^2 h_j + l\eta_k^2\|\bar{B}\|_{\mathrm{L}(\mathcal{X})}^2 h_k \\ &\underbrace{\le}_{(4.7)} \frac{1}{\underline{r}}\sum_{j=1}^{k-1}\eta_j^2\|A\|_{\mathrm{L}(\mathcal{X},\mathcal{Y})}^2\|\bar{B}^{\frac{1}{2}}\mathcal{P}_{j+1}^k(\bar{B})\|_{\mathrm{L}(\mathcal{X})}^2 h_j + \frac{1}{l\underline{r}^2}\eta_k^2\|A\|_{\mathrm{L}(\mathcal{X},\mathcal{Y})}^4 h_k \\ &\le \frac{\|A\|_{\mathrm{L}(\mathcal{X},\mathcal{Y})}^2}{2e\cdot\underline{r}}\sum_{j=1}^{k-1}\frac{\eta_j^2}{\sum_{i=j+1}\eta_i}h_j + \frac{1}{l\underline{r}^2}c_0^2\|A\|_{\mathrm{L}(\mathcal{X},\mathcal{Y})}^4 k^{-2\alpha}h_k \\ &= \frac{c}{\underline{r}}\sum_{j=1}^{k-1}\frac{\eta_j^2}{\sum_{i=j+1}\eta_i}h_j + \frac{c}{l\underline{r}^2}k^{-2\alpha}h_k, \end{aligned} \tag{4.34}$$

where Lemma 3.3 is used in the second-to-last estimate. Combining the estimates (4.32)–(4.34) yields

$$h_{k+1} \le \frac{c}{\underline{r}}\sum_{j=1}^{k-1}\frac{\eta_j^2}{\sum_{i=j+1}\eta_i}h_j + \frac{c}{l\underline{r}^2}k^{-2\alpha}h_k + c_p k^{-(2p+1)(1-\alpha)} + c\delta^2. \tag{4.35}$$

Applying Lemmas 3.4 and 3.5, and using $l\underline{r} \ge n/2$, we can obtain

$$\begin{aligned} h_{k+1} &\le \max_{1\le j\le k}\{h_j\}\Big(\frac{c}{\underline{r}}\sum_{j=1}^{k-1}\frac{\eta_j^2}{\sum_{i=j+1}^k\eta_i} + \frac{c}{l\underline{r}^2}k^{-2\alpha}\Big) + c_p k^{-(2p+1)(1-\alpha)} + c\delta^2 \\ &\le \max_{1\le j\le k}\{h_j\}\frac{c}{\underline{r}}k^{-\min(\alpha,1-\alpha)}\ln k + c_p k^{-(2p+1)(1-\alpha)} + c\delta^2 \\ &\le \frac{c}{\underline{r}}k^{-\beta}\ln k + c\delta^2 \quad \text{with} \quad \beta := \min(\alpha,(1-\alpha)). \end{aligned} \tag{4.36}$$

(Step 4) It remains to give the estimate $\mathbb{E}\|z_k\|_{\mathcal{X}}^2$. Applying Lemma 4.2 to (4.31), as well as $\mathcal{P}_{k+1}^k(\bar{B}) = \mathrm{id}$, yields

$$\begin{aligned} \mathbb{E}\|\bar{B}^s z_{k+1}\|_{\mathcal{X}}^2 &\le \frac{\overline{r}}{\underline{r}^2}\sum_{j=1}^{k-1}\eta_j^2\|\mathcal{P}_{j+1}^k(\bar{B})\bar{B}^{\frac{1}{2}+s}\|_{\mathrm{L}(\mathcal{X})}^2 h_j + \frac{\overline{r}}{\underline{r}^2}\eta_k^2\|\bar{B}^{\frac{1}{2}+s}\|_{\mathrm{L}(\mathcal{X})}^2 h_k \\ &\le \frac{c}{\underline{r}}\sum_{j=1}^{k-1}\frac{\eta_j^2}{\big(\sum_{i=j+1}^k\eta_i\big)^{1+2s}}h_j + \frac{c}{\underline{r}}k^{-2\alpha}\|\bar{B}^{\frac{1}{2}+s}\|_{\mathrm{L}(\mathcal{X})}^2 h_k \\ &:= J_1 + J_2. \end{aligned}$$

In the sequel, we will estimate J_1 and J_2 separately. Using (4.36) and Lemma 3.4, we first bound J_1 as follows:

$$\begin{aligned} J_1 &\le \frac{c}{\underline{r}}\sum_{j=1}^{k-1}\frac{\eta_j^2}{\big(\sum_{i=j+1}^k\eta_i\big)^{2s+1}}\Big(\frac{c}{\underline{r}}(j-1)^{-\beta}\ln k + c\delta^2\Big) \\ &\le \frac{c}{\underline{r}^2}(k^{-(1+2s)(1-\alpha)+\max(0,1-2\alpha-\beta)} + k^{-((1-2s)\alpha+\beta)})\ln k + \frac{c}{\underline{r}}\delta^2(k^{-(1+2s)(1-\alpha)+\max(0,1-2\alpha)} + k^{-(1-2s)\alpha}) \\ &\le \frac{c}{\underline{r}^2}k^{-\min((1+2s)(1-\alpha),1-2s\alpha)}\ln k + \frac{c}{\underline{r}}\delta^2 k^{-\min((1-2s)\alpha,(2s+1)(1-\alpha))}. \end{aligned}$$

Using (4.36) and Lemma 4.1 again,

$$\begin{aligned} J_2 &= \frac{c}{\underline{r}}\eta_k^2\|\bar{B}^{\frac{1}{2}+s}\|_{\mathrm{L}(\mathcal{X})}^2 h_k \le \frac{c}{\underline{r}}\eta_k^2\frac{\|A\|_{\mathrm{L}(\mathcal{X},\mathcal{Y})}^{4s+2}}{(l\underline{r})^{2s+1}}h_k \\ &\le \frac{c}{l^{2s+1}\underline{r}^{2s+3}}\|A\|_{\mathrm{L}(\mathcal{X},\mathcal{Y})}^{4s+2}k^{-2\alpha-\beta}\ln k + \frac{c\|A\|_{\mathrm{L}(\mathcal{X},\mathcal{Y})}^{4s+2}}{l^{2s+1}\underline{r}^{2s+2}}\delta^2 k^{-2\alpha} \\ &\le \frac{c}{n^{2s+1}\underline{r}^2}k^{-2\alpha-\beta}\ln k + \frac{c}{n^{2s+1}\underline{r}}\delta^2 k^{-2\alpha}. \end{aligned}$$

Summing up the estimates for J_1 and J_2 implies that

$$\begin{aligned} J_1 + J_2 &\le \frac{c}{\underline{r}^2}k^{-\min((1+2s)(1-\alpha),\,1-2s\alpha)}\ln k + \frac{c}{\underline{r}}\delta^2 k^{-\min((1-2s)\alpha,\,(2s+1)(1-\alpha))} + \frac{c}{n^{2s+1}\underline{r}^2}k^{-2\alpha-\beta}\ln k + \frac{c}{n^{2s+1}\underline{r}}\delta^2 k^{-2\alpha} \\ &\le \frac{c}{\underline{r}^2}k^{-\min((1+2s)(1-\alpha),\,3\alpha,\,1-2s\alpha)}\ln k + \frac{c}{\underline{r}}\delta^2 k^{-\min((1-2s)\alpha,(2s+1)(1-\alpha))}. \end{aligned}$$

□

Lemma 4.7 below establishes the convergence of the Block SGD iterate x_k (for exact data) to $x^\dagger$. In view of (4.16), the proof follows along the same lines as that in [3], and we therefore omit the details.

Lemma 4.7. *Let Assumption 2.1 be fulfilled. Then the SGD iterate x_k converges to the minimum norm solution $x^\dagger$ as $k(\delta) \to \infty$, i.e.,*

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

With the above preparations at hand, we are now ready to prove Theorem 2.4.

Proof of Theorem 2.4. Similar to the bias decomposition 4.1 and making use of Lemmas 4.5 and 4.6, we can achieve

$$\begin{aligned} \mathbb{E}\|x_{k_{(\delta)}}^\delta - x^\dagger\|_{\mathcal{X}}^2 &\le 2\|\mathbb{E}[x_{k_{(\delta)}}] - x^\dagger\|_{\mathcal{X}}^2 + 2\|\mathbb{E}[x_{k_{(\delta)}} - x_{k_{(\delta)}}^\delta]\|_{\mathcal{X}}^2 + \mathbb{E}\|\mathbb{E}[x_{k_{(\delta)}}^\delta] - x_{k_{(\delta)}}^\delta\|_{\mathcal{X}}^2 \\ &\le 2\|\mathbb{E}[x_{k_{(\delta)}}] - x^\dagger\|_{\mathcal{X}}^2 + c\frac{\delta^2}{l\underline{r}}k_{(\delta)}^{1-\alpha} + \frac{c}{\underline{r}^2}k^{-\min(3\alpha,1-\alpha)}\ln k + \frac{c}{\underline{r}}\delta^2 k^{-\min(\alpha,1-\alpha)}, \end{aligned}$$

where we have used the property that $c^{-1}(k(\delta)-1) \le k(\delta) \le c(k(\delta)-1)$ when $k(\delta)$ is large enough. By Lemma 4.7 and the limit $k(\delta) \to \infty$ from (2.4), we have

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

On the other hand, the choice of $k(\delta)$ in (2.4) ensures that

$$\lim_{\delta\to 0^+}\left[c\frac{\delta^2}{l\underline{r}}k_{(\delta)}^{1-\alpha} + \frac{c}{\underline{r}^2}k^{-\min(3\alpha,1-\alpha)}\ln k + \frac{c}{\underline{r}}\delta^2 k^{-\min(\alpha,1-\alpha)}\right] = 0.$$

□

Finally, we are in a position to give the proof of Theorem 2.5 on the error estimate of Block SGD.

Proof of Theorem 2.5. Applying Lemmas 4.4–4.6 to the bias decomposition 4.1, we obtain

$$\begin{aligned}\mathbb{E}\|x_{k+1}^{\delta}-x^{\dagger}\|_{\mathcal{X}}^2 &\le 2\|\mathbb{E}[x_{k+1}]-x^{\dagger}\|_{\mathcal{X}}^2+2\|\mathbb{E}[x_{k+1}-x_{k+1}^{\delta}]\|_{\mathcal{X}}^2+\mathbb{E}\|\mathbb{E}[x_{k+1}^{\delta}]-x_{k+1}^{\delta}\|_{\mathcal{X}}^2\\ &\le ck^{-2p(1-\alpha)}+\frac{c\delta^2}{l\underline{r}}k^{1-\alpha}+\frac{c}{\underline{r}^2}k^{-\min(3\alpha,1-\alpha)}\ln k+\frac{c}{\underline{r}}\delta^2k^{-\min(\alpha,1-\alpha)}.\end{aligned}$$

This completes the proof. □

4.3. Convergence rate under the discrepancy principle

Since our approach follows the argument of [6], we shall confine ourselves to a short proof.

Proof. (i) Let us fix $0<r<1$ and $\tau^*>1$. Consider the following stopping index $k^*(\delta)$, defined by

$$k^*(\delta):=\min\{k\in\mathbb{N}\mid \|A\mathbb{E}x_k^{\delta}-y^{\delta}\|_{\mathcal{Y}}\le\tau^*\delta\}. \tag{4.37}$$

By Lemma 4.4 with $s=\frac{1}{2}$, and using the definition of $k^*(\delta)$, we can deduce that

$$k^*(\delta)\le\bar{k}(\delta):=\Big\lceil\Big(\frac{\tau^*-1}{\sqrt{l\,\underline{r}}\,c_p}\delta\Big)^{-\frac{2}{(1-\alpha)(\min\{2p,r\}+1)}}\Big\rceil+2, \tag{4.38}$$

with $c_p=\left(\frac{(p+\frac{1}{2})(1-\alpha)}{c_0e\,(2^{1-\alpha}-1)}\right)^{p+\frac{1}{2}}\|w\|_{\mathcal{X}}$ (see [6, Proposition 2.1]). On the other hand, making use of Lemma 4.6 with $s=\frac{1}{2}$, we can further infer for $\beta<\min\{2(1-\alpha),(2p+1)(1-\alpha)+\alpha\}$ and $\gamma\in\min\{\alpha,1-\alpha\}$ that

$$\mathbb{E}\|A(x_{\kappa(\delta)}^{\delta}-\mathbb{E}[x_{\kappa(\delta)}^{\delta}])\|_{\mathcal{Y}}\le nc(\kappa(\delta)^{-\beta}+\delta^2\kappa(\delta)^{-\gamma}). \tag{4.39}$$

In particular, selecting $\gamma>0$ and $\beta>(1-\alpha)(\min\{2p,r\}+1)$ yields

$$\mathbb{E}\|A(x_{\kappa(\delta)}^{\delta}-\mathbb{E}[x_{\kappa(\delta)}^{\delta}])\|_{\mathcal{Y}}=o(\delta^2),\quad \delta\to0^+,\quad \text{if}\quad \kappa(\delta)\ge\bar{k}(\delta). \tag{4.40}$$

By the definition of $\kappa(\delta)$ and the triangle inequality, we have the inclusions

$$\begin{aligned}&\Big\{\|A(x_{\bar{\kappa}(\delta)}^{\delta}-\mathbb{E}[x_{\bar{\kappa}(\delta)}^{\delta}])\|_{\mathcal{Y}}\le(\tau-\tau^*)\delta,\ \|A\mathbb{E}[x_{\bar{\kappa}(\delta)}^{\delta}]-y^{\delta}\|_{\mathcal{Y}}\le\tau^*\delta\Big\}\\ &\subset\Big\{\|Ax_{\bar{\kappa}(\delta)}^{\delta}-y^{\delta}\|_{\mathcal{Y}}\le\tau\delta\Big\}\subset\Big\{\kappa(\delta)\le\bar{\kappa}(\delta)\Big\},\end{aligned}$$

which, together with the fact that $\mathbb{P}(\|A\mathbb{E}[x_{\bar{\kappa}(\delta)}^{\delta}]-y^{\delta}\|_{\mathcal{Y}}\le\tau^*\delta)=1$ due to (4.38), implies that

$$\mathbb{P}\Big(\kappa(\delta)\le\bar{\kappa}(\delta)\Big)\ge\mathbb{P}\Big(\|A(x_{\bar{\kappa}(\delta)}^{\delta}-\mathbb{E}[x_{\bar{\kappa}(\delta)}^{\delta}])\|_{\mathcal{Y}}\le(\tau-\tau^*)\delta\Big). \tag{4.41}$$

On the other hand, by Chebyshev's inequality, it follows that

$$\begin{aligned}\mathbb{P}\Big(\|A(x_{\bar{\kappa}(\delta)}^{\delta}-\mathbb{E}[x_{\bar{\kappa}(\delta)}^{\delta}])\|_{\mathcal{Y}}\le(\tau-\tau^*)\delta\Big)&=1-\mathbb{P}\Big(\|A(x_{\bar{\kappa}(\delta)}^{\delta}-\mathbb{E}[x_{\bar{\kappa}(\delta)}^{\delta}])\|_{\mathcal{Y}}>(\tau-\tau^*)\delta\Big)\\ &\ge1-(\tau-\tau^*)^{-2}\delta^{-2}\mathbb{E}\|A(x_{\bar{\kappa}(\delta)}^{\delta}-\mathbb{E}[x_{\bar{\kappa}(\delta)}^{\delta}])\|_{\mathcal{Y}}^2.\end{aligned} \tag{4.42}$$

Combining (4.40)–(4.42), we arrive at

$$\mathbb{P}\Big(\kappa(\delta)\le\bar{\kappa}(\delta)\Big)\ge1-(\tau-\tau^*)^{-2}\delta^{-2}\mathbb{E}\|A(x_{\bar{\kappa}(\delta)}^{\delta}-\mathbb{E}[x_{\bar{\kappa}(\delta)}^{\delta}])\|_{\mathcal{Y}}^2\to1$$

as $\delta \to 0^+$. This proves the first assertion.
(ii) It is not difficult to show that

$$\|A(x_k^\delta - x_k) - (y^\delta - y^\dagger)\| \le \delta \prod_{j=1}^{k-1}(1 + \frac{\eta_j}{\underline{r}}\|A\|^2_{\mathrm{L}(\mathcal{X},\mathcal{Y})}),$$

$$\|x_k^\delta - x_k\| \le \delta\|A\|_{\mathrm{L}(\mathcal{X},\mathcal{Y})} \sum_{j=1}^{k-1}\eta_j \prod_{j=1}^{k-1}(1 + \frac{\eta_j}{\underline{r}}\|A\|^2_{\mathrm{L}(\mathcal{X},\mathcal{Y})}),$$

where the conventions $\prod_{j=1}^{0} = 1$ and $\sum_{j=1}^{0} = 0$ are adopted. Fix $\varepsilon > 0$ and assume that δ is sufficiently small. Taking advantage of these two inequalities and following the arguments in [6, Proposition 3.1], we can deduce that there exists a sequence of $\{k_\delta^-\}_\delta$ such that

$$\|x_{k(\delta)}^\delta - x^\dagger\|_{\mathcal{X}}\chi_{\{k(\delta)\le k_\delta^-\}} \le \varepsilon \quad \text{and} \quad \lim_{\delta\to 0^+} k_\delta^- = +\infty. \tag{4.43}$$

The previous assertion (i) implies that there exists a sequence $\{k_\delta^+\}$ such that

$$k_\delta^- \le k_\delta^+ \le c\delta^{-\frac{2}{(1-\alpha)(\min\{2p,r\}+1)}} \quad \text{and} \quad \lim_{\delta\to 0^+} \mathbb{P}(\kappa(\delta) \le k_\delta^+) = 1.$$

Consequently, it holds that

$$\begin{aligned}
&\mathbb{P}(\|x_{k(\delta)}^\delta - x^\dagger\|_{\mathcal{X}} \ge \varepsilon)\\
&= \underbrace{\mathbb{P}(\|x_{k(\delta)}^\delta - x^\dagger\|_{\mathcal{X}} \ge \varepsilon,\ \kappa(\delta) \le \kappa_\delta^-)}_{=0 \text{ by } (4.43)} + \mathbb{P}(\|x_{k(\delta)}^\delta - x^\dagger\|_{\mathcal{X}} \ge \varepsilon,\ \kappa(\delta) > \kappa_\delta^-)\\
&= \mathbb{P}(\|x_{k(\delta)}^\delta - x^\dagger\|_{\mathcal{X}} \ge \varepsilon,\ k_\delta^+ \ge \kappa(\delta) > \kappa_\delta^-) + \mathbb{P}(\|x_{k(\delta)}^\delta - x^\dagger\|_{\mathcal{X}} \ge \varepsilon,\ \kappa(\delta) > \kappa_\delta^+).
\end{aligned}$$

By the first assertion, it remains to show that

$$\lim_{\delta\to 0^+} \mathbb{P}(\|x_{k(\delta)}^\delta - x^\dagger\|_{\mathcal{X}} \ge \varepsilon,\ k_\delta^+ \ge \kappa(\delta) > \kappa_\delta^-) = 0,$$

which can be verified in the same manner as in [6]. This completes the proof. □

Remark 4.8. *We emphasize that the deterministic discrepancy principle* (4.38) *appearing in the proof serves only as an auxiliary device to facilitate the probabilistic analysis, which follows the same technical route as Jahn and Jin used in [6]. The stopping rule we advocate is the random discrepancy principle defined in* (2.4)*, for which we have established convergence in probability. This is a weaker mode of convergence than the almost-sure convergence obtained in [28] for a related but technically distinct random discrepancy principle. Proving almost-sure convergence for our setting remains an open and challenging problem, which we leave for future work.*

5. Numerical experiment

5.1. Theoretical analysis

This section is devoted to the application of Block SGD to an elliptic inverse source problem. Let Ω be a bounded domain that is either polygonal in $\mathbb{R}^2$ or polyhedral in $\mathbb{R}^3$. We first discuss the spatial discretization of Ω. Let $\{\mathcal{T}_h\}_{h>0}$ be a quasi-uniform family of simplicial triangulations of Ω with mesh size h, and define the standard finite element space

$$S_h := \left\{ u_h \in C_0(\overline{\Omega}) \;\middle|\; u_h|_T \text{ is linear for all } T \in \mathcal{T}_h \right\}.$$

We consider the following finite-element approximation of an elliptic equation:

$$\int_\Omega \nabla y^\dagger \cdot \nabla \phi_h \, d\xi = \int_\Omega x^\dagger \phi_h \, d\xi \qquad \forall \phi_h \in S_h, \tag{5.1}$$

where $x^\dagger \in S_h$ is an unknown source term. Note that the Eq (5.1) corresponds to the finite-element discretization of the elliptic boundary value problem

$$-\Delta \widetilde{y} = \widetilde{x}^\dagger \quad \text{in } \Omega, \qquad \widetilde{y} = 0 \quad \text{on } \partial\Omega.$$

Now, we define the operators $P_h : L^2(\Omega) \to S_h$, $\Delta_h : S_h \to S_h$ $R_h : H_0^1(\Omega) \to S_h$ by

$$\begin{aligned}
(P_h u, \phi_h) &= (u, \phi_h) \quad \forall\, u \in L^2(\Omega)\, \phi_h \in S_h, \\
(-\Delta_h u_h, \phi_h) &= (\nabla u_h, \nabla \phi_h) \quad \forall\, u_h, \phi_h \in S_h, \\
(\nabla R_h u, \nabla \phi_h) &= (\nabla u, \nabla \phi_h) \quad \forall\, u \in H_0^1(\Omega)\, \phi_h \in S_h.
\end{aligned}$$

Then, the standard finite element theory implies that

$$y^\dagger = \mathcal{I} R_h \widetilde{y} \quad \text{and} \quad x^\dagger = P_h \widetilde{x}^\dagger, \tag{5.2}$$

where $\mathcal{I} : L_h^2(\Omega) \to \mathbb{R}^n$ is the natural mapping that sends a linear finite element function to its values at the nodes, and n is the total number of nodes. with the following approximation property:

$$\begin{aligned}
\|y - R_h \widetilde{y}\|_{L^2(\Omega)} + h\|y - R_h \widetilde{y}\|_{H^1(\Omega)} &\le Ch^2 |\widetilde{y}|_{H^2(\Omega)}, \\
\|y - P_h \widetilde{y}\|_{L^2(\Omega)} + h\|y - P_h \widetilde{y}\|_{H^1(\Omega)} &\le Ch^2 \|\widetilde{x}^\dagger\|_{H^2(\Omega)}.
\end{aligned}$$

The noisy data y^δ is generated from the numerical data $y^\dagger$ as

$$y_i^\delta = y^\dagger + \delta \max_{1 \le j \le n} (|y_j|^\dagger) \xi_i, \quad i = 1, \cdots, n,$$

where $\delta > 0$ is the relative noise level and the random variables ξ_i are standard Gaussian. We want to recover the approximated unknown source $x^\dagger$ in finite element space from the noisy data y_i^δ. The inverse source problem can be formulated as operator equation (1.3) by setting

$$\mathcal{X} = L_h^2(\Omega), \quad \mathcal{Y} = \mathbb{R}^n, \quad A = \mathcal{I} \circ (-\Delta_h)^{-1}. \tag{5.3}$$

To this end, we will apply Block SGD (Algorithm 1) to the discrete inverse source problem. To this end, let us assume that the initial guess $x_1 = 0$. Let us further assume that

$$\widetilde{x}^\dagger = (-\Delta)^{-1} \widetilde{w} \quad \text{with} \quad \widetilde{w} \in L^2(\Omega). \tag{5.4}$$

Lemma 5.1. *Assume that the source condition* (5.4) *is satisfied. Then there exists an element $w_h \in S_h$ for which*

$$x^\dagger = (-\Delta_h)^{-1} w_h,$$

and the norm $\|w\|_{\mathcal{X}}$ is independent of the mesh parameter $h > 0$.

Proof. The classical theory of elliptic equations implies that $\widetilde{x}^\dagger \in H^2(\Omega) \cap H_0^1(\Omega)$. By the argument as in Lemma 4.7 [29], we can show that

$$\|\Delta_h P_h \widetilde{x}^\dagger\|_{L^2(\Omega)} \leq C\|\widetilde{x}^\dagger\|_{H^2(\Omega)}.$$

Combined with the identity $x^\dagger = P_h\widetilde{x}^\dagger$ due to (5.2), this completes the proof. □

Because $\mathcal{I}$ is a bijection between finite-dimensional spaces, its adjoint $\mathcal{I}^*$ is also bijective. Together, Lemmas 4.3 and 5.1 show that Assumption 2.2 holds. In view of Lemma 5.1, we can apply Block SGD to the discrete inverse source problem above. In addition, Theorem 2.5 is applicable to the setting (5.3).

5.2. *Numerical simulation*

In this subsection, we will present some illustrative numerical examples. In the numerical results, let us set

$$\Omega = (0,1)\times(0,1), \quad \widetilde{x}(\xi) = \xi_1(1-\xi_1)\xi_2(1-\xi_2), \quad n = m = 361.$$

The numerical experiments below focus on the relative mean squared error re_k and the stochastic error se_k, i.e.,

$$re_k := \frac{\sqrt{\mathbb{E}\|x^\dagger - x_k^\delta\|_{\mathcal{X}}^2}}{\|x^\dagger\|_{\mathcal{X}}} \quad \text{and} \quad se_k := \mathbb{E}\|x_k^\delta - \mathbb{E}[x_k^\delta]\|_{\mathcal{X}}^2.$$

Figure 1 gives an illustration of the true source, the recovered one, and their difference in a particular experiment, in which the corresponding parameters are chosen as $k = 1500$, $\alpha = 0.01$, $\delta = 0.001$ and $l = 60$.

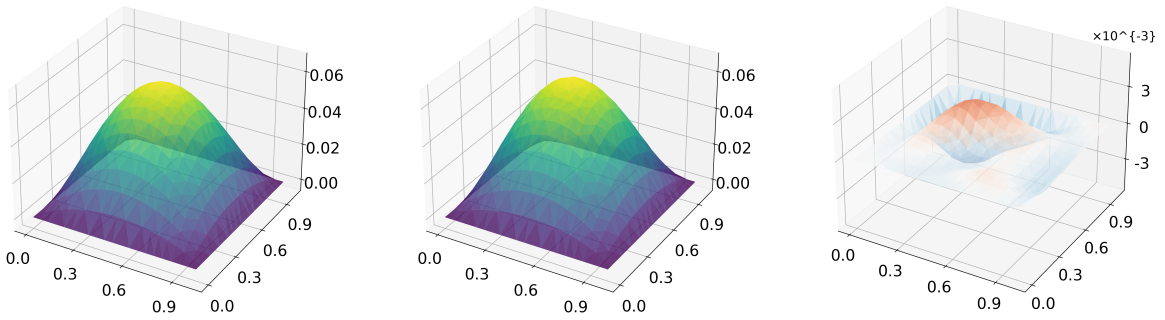
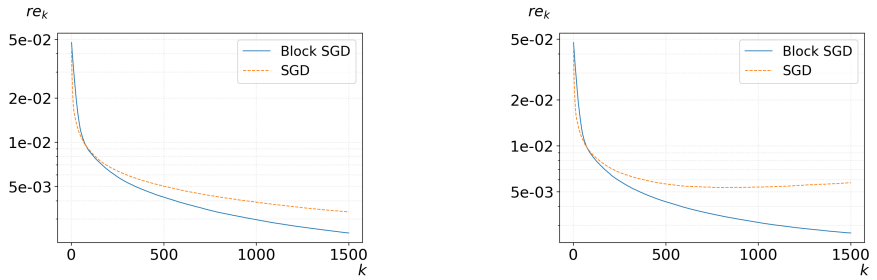
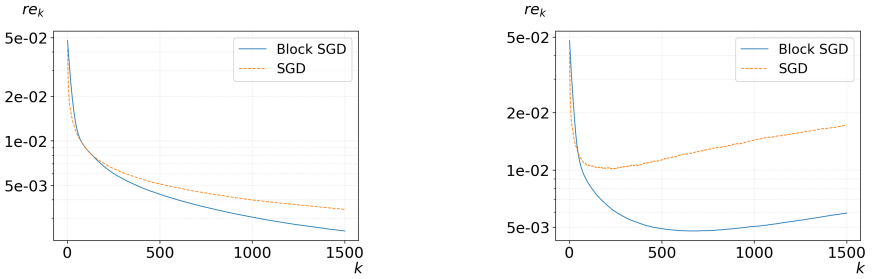


Figure 1. True source (left), the reconstruction source (middle), and their difference.

As shown in Figure 2, compared with the classical SGD method, the Block SGD exhibits better numerical performance at different noise levels, as illustrated by the relative mean squared error re_k with $\alpha = 0.01, l = 60$, and $\delta = 0, 0.001, 0.01, 0.05$. Further, let us compare the relative errors of the Block SGD and Batch SGD at different noise levels and learning rates. Note that they share the same block number $l = 22$ and batch size $r = 16$. As shown in Figure 4, the Block SGD also exhibits better numerical performance for $\alpha = 0.5$ at different noise levels, while their performance is similar for $\alpha = 0.01$.

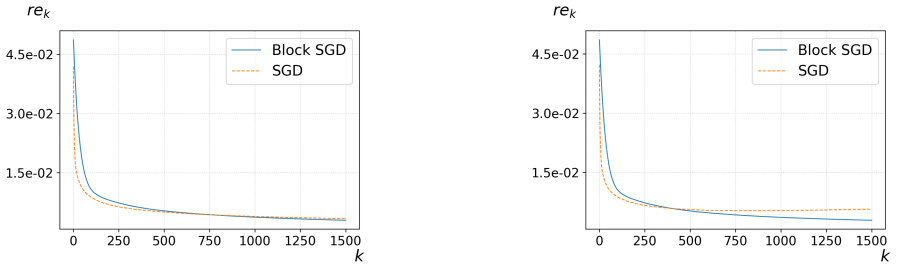


($\alpha = 0.01$, $\delta = 0$, $l = 60$) ($\alpha = 0.01$, $\delta = 0.001$, $l = 60$)

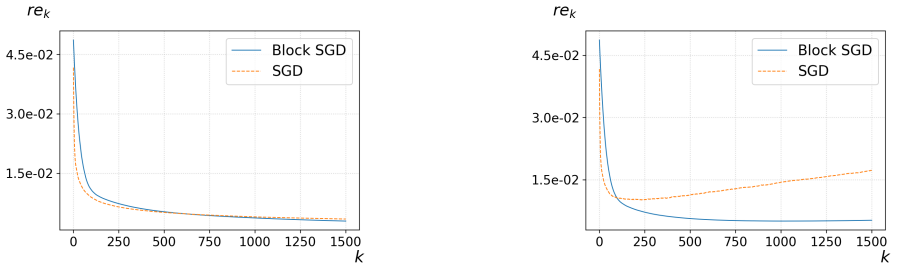


($\alpha = 0.01$, $\delta = 0.01$, $l = 60$) ($\alpha = 0.01$, $\delta = 0.05$, $l = 60$)

Figure 2. Convergence comparison of the relative mean squared error re_k over iteration k for Block SGD and SGD under different noise levels δ.



($\alpha = 0.01$, $\delta = 0$, $l = 22$) ($\alpha = 0.01$, $\delta = 0.001$, $l = 22$)

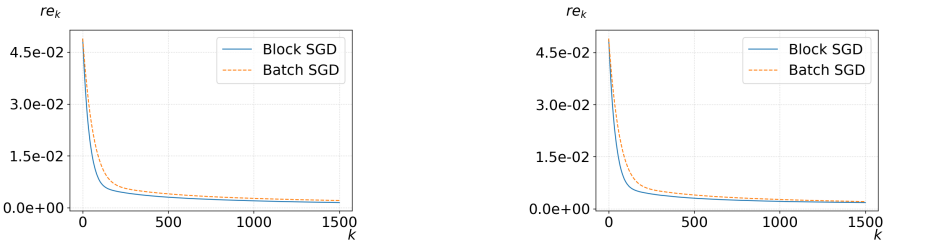


($\alpha = 0.01$, $\delta = 0.01$, $l = 22$) ($\alpha = 0.01$, $\delta = 0.05$, $l = 22$)

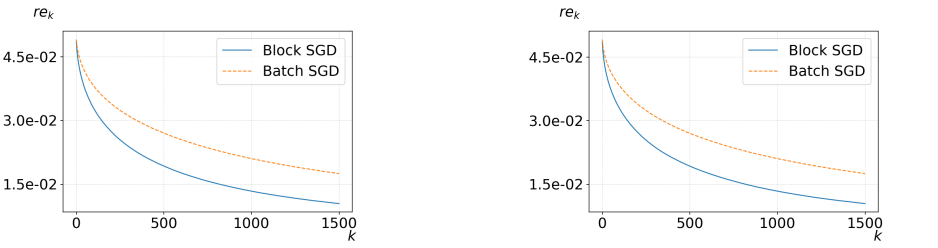
Figure 3. Convergence comparison of the relative mean squared error re_k over iteration k for Block SGD and SGD under different noise levels δ.

Figure 4 shows a comparison of the relative mean squared error re_k between Block SGD and SGD with block size $l = 22$, parameter $\alpha = 0.01$, and $\delta \in \{0, 0.01, 0.001, 0.05\}$. It can be observed that

for different values of δ, as the number of iterations increases, Block SGD consistently outperforms standard SGD. In particular, when $\delta = 0.05$, the improvement of Block SGD over SGD is much more pronounced.

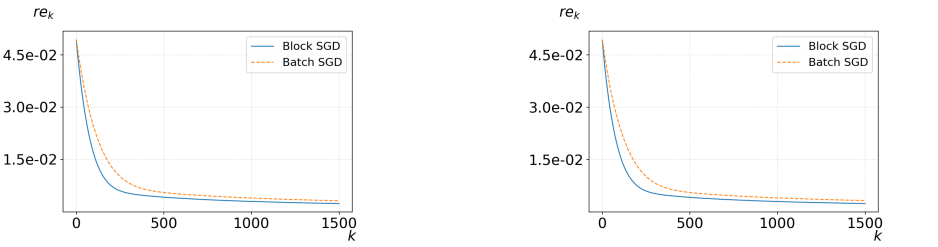


($\alpha = 0.01$, $\delta = 0001$, $l = 22$, $\underline{r} = 16$) ($\alpha = 0.01$, $\delta = 0.01$, $l = 22$, $\underline{r} = 16$)

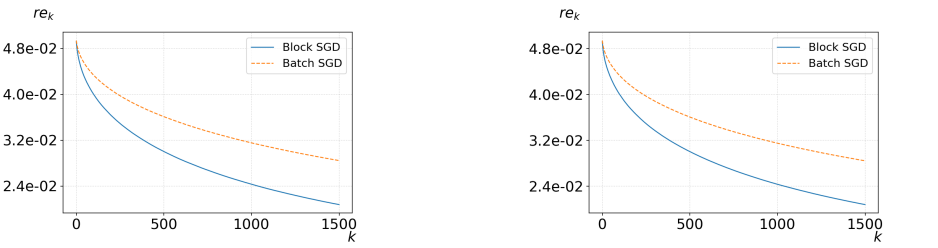


($\alpha = 0.5$, $\delta = 0.001$, $l = 22$, $\underline{r} = 16$) ($\alpha = 0.5$, $\delta = 0.01$, $l = 22$, $\underline{r} = 16$)

Figure 4. Convergence comparison of the relative mean squared error re_k over iteration k for Block SGD and SGD under different noise levels δ.



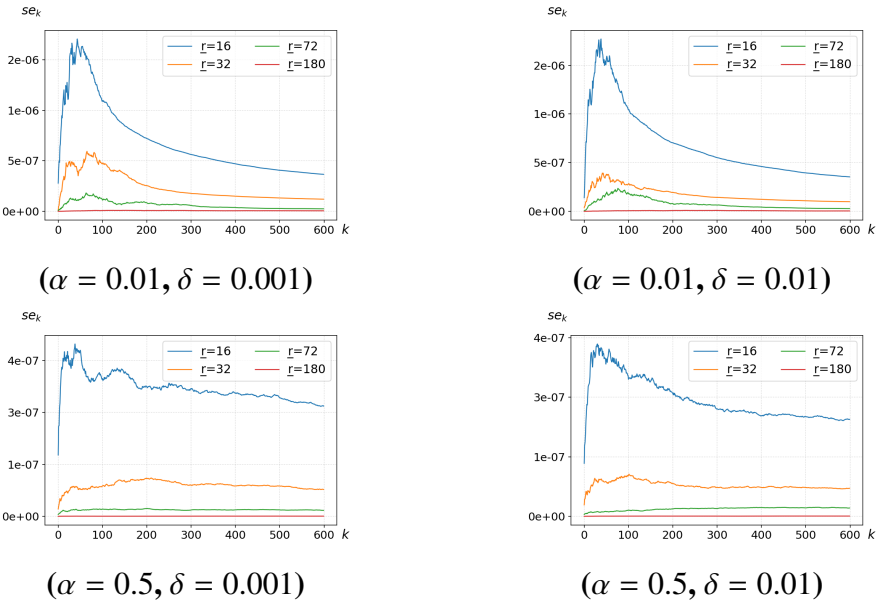
($\alpha = 0.01$, $\delta = 0001$, $l = 11$, $\underline{r} = 32$) ($\alpha = 0.01$, $\delta = 0.01$, $l = 11$, $\underline{r} = 32$)



($\alpha = 0.5$, $\delta = 0.001$, $l = 11$, $\underline{r} = 32$) ($\alpha = 0.5$, $\delta = 0.01$, $l = 11$, $\underline{r} = 32$)

Figure 5. Convergence comparison of the relative mean squared error re_k over iteration k for Block SGD and SGD under different noise levels δ.

The figure above shows the relative errors of Batch SGD and Block SGD with block number $l = 11$ and batch size $\bar{r} = 32$, for parameters $\alpha = 0.01, 0.5$ and $\delta = 0.01, 0.001$. It is observed that Block SGD consistently performs better than Batch SGD. Moreover, the performance gain of Block SGD over Batch SGD is more pronounced when the value of $\alpha = 0.5$.



($\alpha = 0.01, \delta = 0.001$) ($\alpha = 0.01, \delta = 0.01$)

($\alpha = 0.5, \delta = 0.001$) ($\alpha = 0.5, \delta = 0.01$)

Figure 6. Comparison of the stochastic error se_k over iteration k for Block SGD for different minimum batch sizes $\underline{r}$, noise levels δ, and α.

To explain the superiority of the Block SGD over the standard SGD, let us present our last experiment, which focuses on the comparison of the stochastic error se_k over iteration k. As illustrated in Figure 6, for any fixed α, δ, and for sufficiently large k, the stochastic error se_k decreases consistently as the batch size $\underline{r}$ increases. This confirms the theoretical results from Lemma 4.6. In summary, Block SGD outperforms standard SGD by leveraging block updates to reduce stochastic error and demonstrating superior numerical stability in practical experiments.

5.3. Inverse source problem in a wave equation

To illustrate the behavior of the methods when the assumptions of the theory are not satisfied, we consider an example in which the true source is deliberately constructed so that the power-type source condition of Assumption 2.2 fails. Since the error estimates of this paper are derived under this power-type source condition, they do not apply to this example; the numerical results below serve as an illustration of what happens outside the regime covered by the theory.

In this subsection, we apply SGD and Block SGD to an inverse source problem arising in a wave equation. More precisely, we consider the following initial-boundary value problem:

$$\begin{cases} \partial_t^2 u(\xi,t) - \Delta u(\xi,t) = \mu(t)\,\widetilde{x}^\dagger(\xi), & (\xi,t) \in \Omega \times [0,1], \\ u(\xi,0) = 0, \quad \partial_t u(\xi,0) = 0, & \xi \in \Omega, \\ u(\xi,t) = 0, & (\xi,t) \in \partial\Omega \times [0,1], \end{cases} \tag{5.5}$$

where

$$\mu(t) := 1 + t^2 \quad \text{and} \quad \Omega := \{\xi \in \mathbb{R}^2 : |\xi| < 1\},$$

and $\widetilde{x}^\dagger$ is the unknown true source. Let $u^\dagger$ denote the solution of (5.5) corresponding to the true source $\widetilde{x}^\dagger$. Our goal is to recover the space component $\widetilde{x}^\dagger$ from the heat flux $\partial u/\partial n$ on the boundary $\partial\Omega \times (0,1)$. For every $x \in H_0^1(\Omega)$, problem (5.5) with source term $\widetilde{x}^\dagger = x$ admits a unique solution u, and the resulting solution map induces a bounded linear operator $\mathcal{A} : H_0^1(\Omega) \to L^2(\partial\Omega \times (0,1))$ defined by $\mathcal{A}x := \frac{\partial u}{\partial n}\Big|_{\partial\Omega\times(0,1)}$. For the numerical implementation, we introduce two finite-dimensional subspaces $\mathcal{X} \subset H_0^1(\Omega)$ and $\mathcal{Y} \subset L^2(\partial\Omega \times (0,1))$ of dimensions m and n, respectively. Following [30], we choose orthonormal bases $\{\phi_i\}_{i=1}^m$ of $\mathcal{X}$ and $\{\psi_i\}_{i=1}^n$ of $\mathcal{Y}$. We then define the forward operator $A \in \mathscr{L}(\mathcal{X}, \mathcal{Y})$ by $A = \mathcal{P}_{\mathcal{Y}} \circ \mathcal{A}\big|_{\mathcal{X}}$, where $\mathcal{P}_{\mathcal{Y}} \in \mathscr{L}(L^2(\partial\Omega \times (0,1)), \mathcal{Y})$ is the orthogonal projection onto $\mathcal{Y}$.

We deliberately choose $m > n$, so that A is non-injective, and select $x^\dagger \in \mathcal{X}$ such that the orthogonal projection of $x^\dagger$ onto $\ker(A)$ is nontrivial, which ensures that

$$x^\dagger - x_1 \notin \operatorname{rg}(A^*) = \operatorname{rg}(B), \qquad \text{with } x_1 := 0.$$

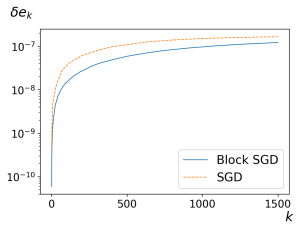
We claim that the source condition (2.3) does not hold in this case. Indeed, if (2.3) holds for some $p > 1/2$, then $x^\dagger - x_1 = \bar{B}^{1/2}\bar{B}^{p-1/2}w$, which implies $x^\dagger - x_1 \in \operatorname{rg}(\bar{B}^{1/2})$; the case $p \le 1/2$ is trivial. Hence $x^\dagger - x_1 \in \operatorname{rg}(\bar{B}^{\min\{1/2,p\}})$. By Lemma 4.3, we conclude that $x^\dagger - x_1 \in \operatorname{rg}(B^{\min\{1/2,p\}}) = \operatorname{rg}(B)$, which contradicts the inclusion $x^\dagger - x_1 \notin \operatorname{rg}(B)$. In the experiments below, we set $n = 105$ and $m = 210$.

In the numerical examples below, let re_k denote the relative error of the k-th iterate, normalized so that $re_0 = 1$. Since re_k remains indistinguishable from 1 throughout, we plot the relative error reduction

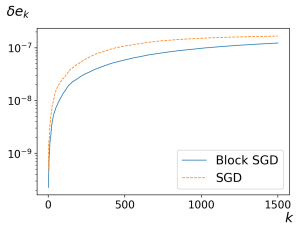
$$\delta e_k := 1 - re_k, \tag{5.6}$$

which makes visible the microscopic changes that a linear scale hides. Figures 7 and 8 compare δe_k for Block SGD and SGD with block sizes $l = 30$ and $l = 100$, respectively, at $\alpha = 0.01$ and noise levels $\delta \in \{0, 10^{-3}, 10^{-2}, 5 \times 10^{-2}\}$. Figures 9 and 10 compare δe_k for Block SGD and Batch SGD with $(l, \underline{r}) = (50, 2)$ and $(20, 5)$, respectively, at $(\alpha, \delta) = (0.01, 10^{-3}), (0.01, 10^{-2}), (0.5, 10^{-3}), (0.5, 10^{-2})$.

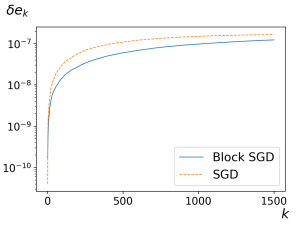
Figures 7 to 10 show that none of the methods—SGD, Batch SGD, or Block SGD—converge. In every case, δe_k rises rapidly at first and then saturates at a level of order 10^{-7}, in agreement with the failure of the source condition. The stagnation of the relative error indicates that the source condition is not merely a technical convenience but is essential for the quantitative error estimates established in this paper. To the best of our knowledge, a rigorous theoretical explanation of this stagnation phenomenon, where the classical source condition fails, remains open for stochastic iterative methods. This may be an interesting direction for future research.



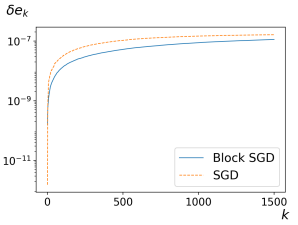
($\alpha = 0.01$, $\delta = 0$, $l = 30$)



($\alpha = 0.01$, $\delta = 0.001$, $l = 30$)

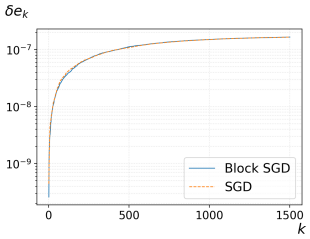


($\alpha = 0.01$, $\delta = 0.01$, $l = 30$)

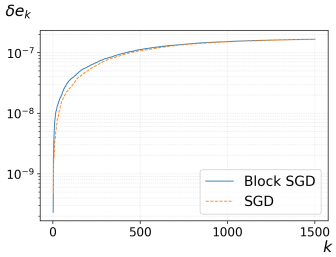


($\alpha = 0.01$, $\delta = 0.05$, $l = 30$)

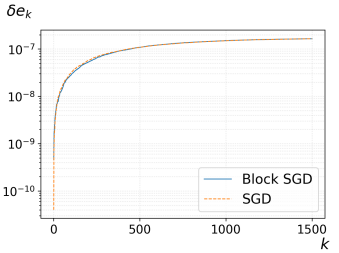
Figure 7. Convergence comparison of the relative error reduction δe_k over iteration k for Block SGD and SGD under different noise levels δ.



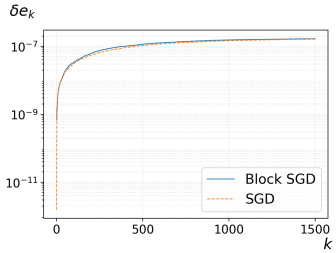
($\alpha = 0.01$, $\delta = 0$, $l = 100$)



($\alpha = 0.01$, $\delta = 0.001$, $l = 100$)

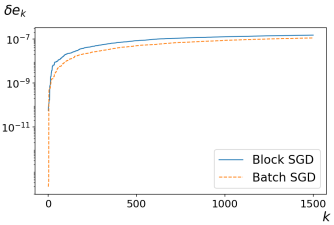


($\alpha = 0.01$, $\delta = 0.01$, $l = 100$)

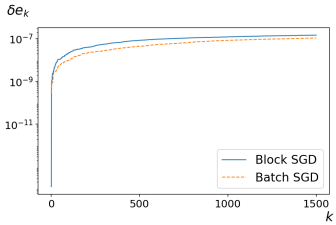


($\alpha = 0.01$, $\delta = 0.05$, $l = 100$)

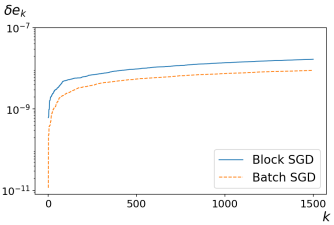
Figure 8. Convergence comparison of the relative error reduction δe_k over iteration k for Block SGD and SGD under different noise levels δ.



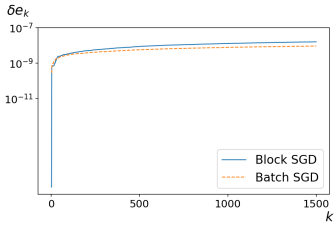
($\alpha = 0.01$, $\delta = 0001$, $l = 50$, $\underline{r} = 2$)



($\alpha = 0.01$, $\delta = 0.01$, $l = 50$, $\underline{r} = 2$)

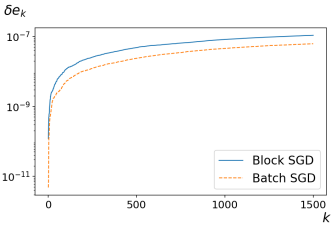


($\alpha = 0.5$, $\delta = 0.001$, $l = 50$, $\underline{r} = 2$)

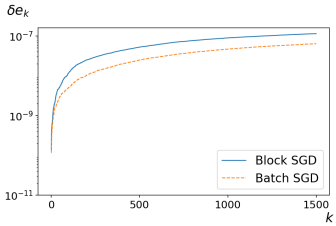


($\alpha = 0.5$, $\delta = 0.01$, $l = 50$, $\underline{r} = 2$)

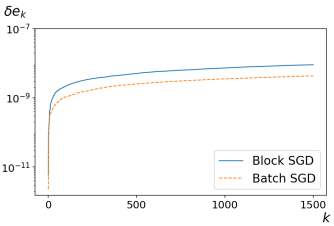
Figure 9. Convergence comparison of the relative error reduction δe_k over iteration k for the Block SGD and the Batch SGD under different noise levels δ.



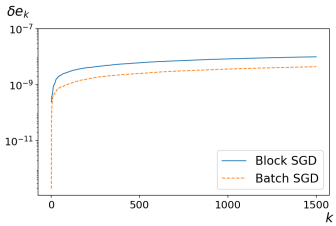
($\alpha = 0.01$, $\delta = 0.001$, $l = 20$, $\underline{r} = 5$)



($\alpha = 0.01$, $\delta = 0.01$, $l = 20$, $\underline{r} = 5$)



($\alpha = 0.5$, $\delta = 0.001$, $l = 20$, $\underline{r} = 5$)



($\alpha = 0.5$, $\delta = 0.01$, $l = 20$, $\underline{r} = 5$)

Figure 10. Convergence comparison of the relative error reduction δe_k over iteration k for the Block SGD and the Batch SGD under different noise levels δ.

Use of AI tools declaration

AI tools were used solely to assist with language polishing, including grammar correction and improvement of academic expression. All research design, analysis, and conclusions were independently completed by the authors.

Acknowledgments

The work of De-Han Chen was financially supported by Deutsche Forschungsgemeinschaft (DFG) research Grant YO 159/5-1 (No. 513566305). The work of Ting Cheng was financially supported by the National Natural Science Foundation of China (Nos. 12271196 and 12271197). The work of Daijun Jiang was financially supported by the National Natural Science Foundation of China (Grant Nos. 12271197 and 11871240), and the Fundamental Research Funds for the Central Universities (No. XJ2026000101).

Conflict of interest

The authors declare there are no conflicts of interest.

References

1. H. W. Engl, M. Hanke, A. Neubauer, *Regularization of inverse problems*, Springer Dordrecht, 2020. Available from: https://link.springer.com/book/9780792341574.
2. B. Hofmann, P. Mathé, M. Schieck, Modulus of continuity for conditionally stable ill-posed problems in Hilbert space, *J. Inverse Ill-Posed Probl.*, **16** (2008), 567–585. https://doi.org/10.1515/JIIP.2008.030
3. B. Jin, X. Lu, On the regularization property of stochastic gradient descent, *Inverse Probl.*, **35** (2019), 015004. https://doi.org/10.1088/1361-6420/aaea2a
4. B. Jin, Z. Zhou, On the convergence of stochastic variance reduced gradient for linear inverse problems, *Inverse Probl.*, **42** (2026), 045006. https://doi.org/10.1088/1361-6420/ae5845
5. S. Lu, P. Mathé, Stochastic gradient descent for linear inverse problems in Hilbert spaces, *Math. Comput.*, **91** (2022), 1763–1788. https://doi.org/10.1090/mcom/3714
6. T. Jahn, B. Jin, On the discrepancy principle for stochastic gradient descent, *Inverse Probl.*, **36** (2020), 095009. https://doi.org/10.1088/1361-6420/abaa58
7. M. J. Ehrhardt, Ž. Kereta, J. Liang, J. Tang, A guide to stochastic optimisation for large-scale inverse problems, *Inverse Probl.*, **41** (2025), 053001. http://doi.org/10.1088/1361-6420/adc0b7
8. J. Lin, L. Rosasco, Optimal rates for multi-pass stochastic gradient methods, *J. Mach. Learn. Res.*, **18** (2017), 1–47. Available from: http://jmlr.org/papers/v18/17-176.html.
9. S. Kassing, S. Weissmann, L. Döring, Controlling the flow: Stability and convergence for stochastic gradient descent with decaying regularization, in *Advances in Neural Information Processing Systems*, **38** (2026), 83893–83932. https://doi.org/10.52202/085713-2530

10. Z. Zhou, On the convergence of a data-driven regularized stochastic gradient descent for nonlinear ill-posed problems, *SIAM J. Imag. Sci.*, **18** (2025), 388–448. https://doi.org/10.1137/24M1647394

11. A. Beck, L. Tetruashvili, On the convergence of block coordinate descent type methods, *SIAM J. Optim.*, **23** (2013), 2037–2060. https://doi.org/10.1137/120887679

12. Q. Jin, X. Lu, L. Zhang, Stochastic mirror descent method for linear ill-posed problems in Banach spaces, *Inverse Probl.*, **39** (2023), 065010. https://doi.org/10.1088/1361-6420/accd8e

13. B. Jin, Z. Zhou, J. Zou, On the saturation phenomenon of stochastic gradient descent for linear inverse problems, *SIAM/ASA J. Uncertainty Quantif.*, **9** (2021), 1553–1588. https://doi.org/10.1137/20M1374456

14. R. Gu, Z. Fu, B. Han, H. Fu, Stochastic gradient descent method with convex penalty for ill-posed problems in Banach spaces, *Inverse Probl.*, **41** (2025), 055003. https://doi.org/10.1088/1361-6420/adc94d

15. B. Jin, Ž. Kereta, On the convergence of stochastic gradient descent for linear inverse problems in Banach spaces, *SIAM J. Imag. Sci.*, **16** (2023), 671–705. https://doi.org/10.1137/22M1518542

16. J. Huang, Q. Jin, X. Lu, L. Zhang, On early stopping of stochastic mirror descent method for ill-posed inverse problems, *Numer. Math.*, **157** (2025), 539–571. https://doi.org/10.1007/s00211-025-01458-7

17. B. Jin, Z. Zhou, J. Zou, On the convergence of stochastic gradient descent for nonlinear ill-posed problems, *SIAM J. Optim.*, **30** (2020), 1421–1450. https://doi.org/10.1137/19M1271798

18. Q. Jin, L. Chen, Stochastic variance reduced gradient method for linear ill-posed inverse problems, *Inverse Probl.*, **41** (2025), 055014. https://doi.org/10.1088/1361-6420/add04b

19. Q. Jin, Y. Liu, Convergence analysis of a stochastic heavy-ball method for linear ill-posed problems, *J. Comput. Appl. Math.*, **470** (2025), 116702. https://doi.org/10.1016/j.cam.2025.116702

20. Y. Zhang, C. Chen, Stochastic asymptotical regularization for linear inverse problems, *Inverse Probl.*, **39** (2023), 015007. https://doi.org/10.1088/1361-6420/aca70f

21. H. Long, Y. Zhang, Stochastic asymptotical regularization for nonlinear ill-posed problems, *Inverse Probl.*, **41** (2025), 065017. https://doi.org/10.1088/1361-6420/addffc

22. A. Klenke, *Probability Theory: A Comprehensive Course*, Springer, London, 2008. https://doi.org/10.1007/978-1-84800-048-3

23. K. Ito, B. Jin, *Inverse Problems: Tikhonov Theory and Algorithms*, World Scientific, 2014.

24. B. Jin, Z. Kereta, Y. Xia, Stochastic Gradient Descent for Nonlinear Inverse Problems in Banach Spaces, *SIAM J. Imag. Sci.*, **19** (2026), 1244–1271. https://doi.org/10.1137/25M1822289

25. T. Hytönen, J. van Neerven, M. Veraar, L. Weis, *Analysis in Banach spaces, Volume I: Martingales and Littlewood-Paley Theory*, Springer Cham, 2016. https://doi.org/10.1007/978-3-319-48520-1

26. K. Schmüdgen, *Unbounded Self-adjoint Operators on Hilbert Space*, Springer Dordrecht, 2012. https://doi.org/10.1007/978-94-007-4753-1

27. B. Simon, *Loewner's Theorem on Monotone Matrix Functions*, Springer Cham, 2019. https://doi.org/10.1007/978-3-030-22422-6

28. Y. Zhang, C. Chen, Stochastic linear regularization methods: random discrepancy principle and applications, *Inverse Probl.*, **40** (2024), 025007. https://doi.org/10.1088/1361-6420/ad149e

29. M. Geissert, Applications of discrete maximal L_p regularity for finite element operators, *Numer. Math.*, **108** (2007), 121–149. https://doi.org/10.1007/s00211-007-0110-1

30. R. Wang, D. H. Chen, T. Cheng, D. Jiang, A new strategy for convergence rates of non-stationary asymptotical regularization method for linear inverse problems, *Inverse Probl.*, **41** (2025), 105007. https://doi.org/10.1088/1361-6420/ae0e4b

AIMS Press