Inverse Problems

PAPER • OPEN ACCESS

On the discrepancy principle for stochastic gradient descent

To cite this article: Tim Jahn and Bangti Jin 2020 *Inverse Problems* **36** 095009

OPEN ACCESS
IOP Publishing
Inverse Problems
Inverse Problems **36** (2020) 095009 (30pp)
https://doi.org/10.1088/1361-6420/abaa58

On the discrepancy principle for stochastic gradient descent

Tim Jahn[1] **and Bangti Jin**[2,3]

[1] Institute for Mathematics, Goethe-University Frankfurt, 60325 Frankfurt am Main, Germany
[2] Department of Computer Science, University College London, Gower Street, London WC1E 6BT, United Kingdom

E-mail: jahn@math.uni-frankfurt.de and b.jin@ucl.ac.uk

Received 30 April 2020, revised 29 June 2020
Accepted for publication 29 July 2020
Published 2 September 2020

Abstract
Stochastic gradient descent (SGD) is a promising numerical method for solving large-scale inverse problems. However, its theoretical properties remain largely underexplored in the lens of classical regularization theory. In this note, we study the classical discrepancy principle, one of the most popular *a posteriori* choice rules, as the stopping criterion for SGD, and prove the finite-iteration termination property and the convergence of the iterate in probability as the noise level tends to zero. The theoretical results are complemented with extensive numerical experiments.

Keywords: stochastic gradient descent, discrepancy principle, convergence

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

1. Introduction

In this work, we study the following finite-dimensional linear inverse problem:

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

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

[3] Author to whom any correspondence should be addressed.

$y^\dagger = Ax^\dagger$ (with the reference solution $x^\dagger$ being any exact solution)

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

where $\xi \in \mathbb{R}^n$ denotes the noise, with a noise level $\delta = \|\xi\|$. In the literature, a large number of numerical methods have been proposed for solving linear inverse problems accurately and efficiently (see, e.g., [4, 9, 13]).

When the size of problem (1.1) is massive, one attractive method is a simple stochastic gradient descent (SGD) [3, 19]. In its simplest form, it reads as follows: given an initial guess $x_1^\delta = x_1 \in \mathbb{R}^m$, let

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

where $\eta_k > 0$ is a decreasing stepsize, a_i is the ith row of the matrix A (as a column vector), $(\cdot, \cdot)$ denotes Euclidean inner product on $\mathbb{R}^m$, and the row index i_k at the kth SGD iteration is chosen uniformly (with replacement) from the set $\{1, \ldots, n\}$. It can be derived by applying stochastic gradient descent to the quadratic functional:

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

Distinctly, the method (1.2) operates only on one single data pair (a_{i_k}, y_{i_k}) each time, and thus it is directly scalable to the data size n of problem (1.1). This feature makes it especially attractive in the context of massive data. In fact, SGD and its variants (e.g., minibatch and accelerated) have been established as the workhorse behind many challenging training tasks in deep learning [2, 3], and they are also popular for image reconstruction in computed tomography [6, 18].

Despite the apparent simplicity of the method, the mathematical theory in the lens of classical regularization theory is far from complete. In the work [10], the regularizing property of SGD was proved for a polynomially decaying stepsize schedule, when the stopping index k is determined *a priori* in relation with the noise level δ. Further, a convergence rate in the mean squared norm between the iterate x_k^δ and the exact solution $x^\dagger$ was derived, under suitable source type condition on the ground truth $x^\dagger$. These results were recently extended to mildly nonlinear inverse problems, further assisted with suitable nonlinearity conditions of the forward map [11]. However, in these works, the convergence rate can only be achieved under a knowledge of the smoothness parameter of $x^\dagger$, which is usually not directly accessible in practice. Therefore, it is of enormous practical importance and theoretical interest to develop *a posteriori* stopping rules that do not require such a knowledge.

For deterministic iterative methods [13], e.g., Landweber method and Gauss–Newton method, one popular *a posteriori* stopping rule is the discrepancy principle, due to Morozov [17]. Specifically, with x_k^δ being the kth iterate constructed by an iterative regularization method, the principle determines the stopping index $k(\delta)$ by

$$k(\delta) := \min\left\{k \in \mathbb{N} : \|Ax_k^\delta - y^\delta\| \leqslant \tau\delta\right\}, \tag{1.3}$$

where the constant $\tau > 1$ is fixed. Note that for SGD, the stopping index $k(\delta)$ depends on the random iterate x_k^δ, and thus it is also a random variable, which poses the main challenge in the theoretical analysis. The use of the discrepancy principle to many deterministic iterative methods is well understood (see the monograph [13] and the references therein), but in the context of stochastic iterative methods, it has not been explored so far, to the best of our knowledge. The goal of this work is to study the basic properties of the discrepancy principle for SGD.

It is worth noting that a direct computation of the residual $\|Ax_k^\delta - y^\delta\|$ at every SGD iteration is demanding. However, one may compute it not at every SGD iteration but only with a given frequency (e.g., per epoch, see section 5), as done by the popular stochastic variance reduced gradient [12], for which residual evaluation is a part of gradient computation. Also there are efficient methods to compute the residual $\|Ax_k^\delta - y^\delta\|$ using randomized SVD [14], by exploiting the intrinsic low-rank nature of many practical inverse problems.

Now we specify the algorithmic parameters for SGD, and state the main results of the work. Throughout, we make the following assumption on the stepsizes and the regularity condition on the ground truth solution $x^\dagger$, i.e., the minimum-norm solution defined by

$$x^\dagger = \arg\min_{x:Ax=y^\dagger} \|x\|. \tag{1.4}$$

The stepsize schedule in (i) is commonly known as the polynomially decaying stepsize schedule, and (ii) is the classical power type source condition, where $B = n^{-1}(A^{\mathrm{T}}A)$ (with n being the data size, i.e., the number of rows in A), imposing a type of smoothness on the solution $x^\dagger$ (relative to the system matrix A and the initial guess x_1). In the analysis and computation below, x_1 is fixed at 0. Generally, in classical regularization theory for infinite-dimensional inverse problems, the source element w plays the role of a Lagrangian multiplier of the constrained problem in (1.4), whose existence is not ensured for an operator with a nonclosed range and has to be assumed [4, 9]. In the finite-dimensional case, the existence of a source element w for the case $p \leqslant \frac{1}{2}$ is ensured, but the norm of the source element w can be arbitrarily large.

Assumption 1.1. The following conditions hold.

(i) The stepsizes η_j satisfy $\eta_j = c_0 j^{-\alpha}$, with $\alpha \in (0,1)$ and $c_0 \leqslant (\max_{j=1,\ldots,n}\|a_j\|^2)^{-1}$.
(ii) There is a $p > 0$ and a $w \in \mathbb{R}^m$ such that $x^\dagger - x_1 = B^p w$

The first theorem gives a finite-iteration termination property of the discrepancy principle, where $\mathbb{P}$ is with respect to the filtration generated by the random index $(i_k)_{k=1}^\infty$. It can also be viewed as a partial result on the optimality. It implies in particular that for $p < \frac{1}{2}$, the data propagation error is of optimal order. The proof relies crucially on the observation that the variance component of the mean squared residual contributes only marginally for sufficiently large k.

Theorem 1.1. *Let assumption 1.1 be fulfilled, and $k(\delta)$ be determined by the discrepancy principle (1.3). Then for all $0 < r < 1$ and $\tau > \tau^* > 1$, with $c = \left(\frac{\tau^*-1}{\sqrt{n}c_p}\right)^{-\frac{2}{(1-\alpha)(\min(2p,r)+1)}} + 2$, there holds*

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

with the constant $c_p = (\frac{(p+\frac{1}{2})(1-\alpha)}{c_0 e(2^{1-\alpha}-1)})^{p+\frac{1}{2}}\|w\|$.

The second contribution of this work is on the convergence in probability of the SGD iterate $x_{k(\delta)}^\delta$ with the stopping index $k(\delta)$ determined by (1.3). This result has one drawback. In the proof, we have to assume that the stopping index $k(\delta)$ is independent of the iterates $x_{k(\delta)}^\delta$. In practice, this can be achieved by running SGD twice with the same data (y^δ, δ): the first round is for the determination of $k(\delta)$, then the second (independent) round is stopped using $k(\delta)$. This increases the computational expense by a factor of 2. However, the numerical results in section 5 show that one can use the iterate from the first run without compromising the accuracy.

Theorem 1.2. *Let assumption 1.1 be fulfilled, and $k(\delta)$ be determined by the discrepancy principle (1.3). Then for all $\varepsilon > 0$ there holds*

$$\mathbb{P}\left(\|x_{k(\delta)}^{\delta} - x^{\dagger}\| \geqslant \varepsilon\right) \to 0 \quad \textit{as} \quad \delta \to 0^{+},$$

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

In sum, theorems 1.1 and 1.2 confirm that the discrepancy principle is a valid *a posteriori* stopping rule for SGD. However, they do not give a rate of convergence, which remains an open problem. Numerically, we observe that the convergence rate obtained by the discrepancy principle is nearly order-optimal for low-regularity solutions, as the *a priori* rule in the regime in [10], and the performance is competitive with the standard Landweber method. Thus, the method is especially attractive for finding a low-accuracy solution. However, for very smooth solutions (i.e., large p), it manifested an undesirable saturation phenomenon, due to the presence of the significant variance component (when compared with the approximation error), under the setting of assumption 1.1. The rest of the paper is organized as follows. In sections 2 and 3, we prove theorems 1.1 and 1.2, respectively. Several auxiliary results needed for the proof of theorem 1.1 are given in section 4. Finally, several numerical experiments are presented in section 5 to complement the theoretical analysis. We conclude with some useful notation. We denote the SGD iterate for exact data $y^\dagger$ by x_k, and that for noisy data y^δ by x_k^δ. The expectation $\mathbb{E}[\cdot]$ is with respect to the filtration $\mathcal{F}_k$, generated by the random indices $\{i_1, \ldots, i_k\}$.

2. The proof of theorem 1.1

In this section, we give the proof of theorem 1.1. First, we give several preliminary facts. By the construction in (1.2), since x_k^δ is measurable with respect to $\mathcal{F}_{k-1}$,

$$\begin{aligned}\mathbb{E}[x_{k+1}^\delta|\mathcal{F}_{k-1}] &= x_k^\delta - \eta_k n^{-1}\sum_{i=1}^{n}((a_i, x_k^\delta) - y_i^\delta)a_i \\ &= x_k^\delta - \eta_k n^{-1}(A^t A x_k^\delta - A^t y^\delta).\end{aligned}$$

Thus, by the law of total expectation, the sequence $(\mathbb{E}[x_k^\delta])_{k\in\mathbb{N}}$ satisfies the following recursion:

$$\mathbb{E}[x_{k+1}^\delta] = \mathbb{E}[x_k^\delta] - \eta_k(\bar{A}^t\bar{A}\mathbb{E}[x_k^\delta] - \bar{A}^t\bar{y}^\delta) \tag{2.1}$$

with $\bar{A} = n^{-\frac{1}{2}}A$ and $\bar{y}^\delta = n^{-\frac{1}{2}}y^\delta$. This is exactly the classical Landweber method [16] (but with diminishing stepsizes) applied to the rescaled linear system $\bar{A}x = \bar{y}^\delta$. For the Landweber method, the discrepancy principle (1.3), e.g., regularizing property and optimal convergence rates, has been thoroughly studied for both linear and nonlinear inverse problems (see, e.g., [4, chapter 6] and [13]). The key insight for the analysis below is the following empirical observation: for a suitably large k, typically the variance component $\mathbb{E}[\|A(x_k^\delta - \mathbb{E}[x_k^\delta])\|^2] \ll \delta^2$, as confirmed by the numerical experiments in section 5.2. This fact allows us to transfer the results for the Landweber method to SGD.

The proof of theorem 1.1 employs two preliminary results, whose lengthy proofs are deferred to section 4. The first result gives an upper bound of the following stopping index

$k^*(\delta)$, for any $\tau^* > 1$, defined by

$$k^*(\delta) := \min\{k \in \mathbb{N} \ : \ \|A\mathbb{E}[x_k^\delta] - y^\delta\| \leqslant \tau^*\delta\}. \tag{2.2}$$

Clearly, $k^*(\delta)$ is the stopping index by the classical discrepancy principle, when applied to the sequence $(\mathbb{E}[x_k^\delta])_{k\in\mathbb{N}}$, which is exactly the Landweber method, in view of the relation (2.1).

Proposition 2.1. *Let assumption 1.1 be fulfilled. Then for $k^*(\delta)$ defined in (2.2), there holds*

$$k^*(\delta) \leqslant \left(\frac{\tau^* - 1}{\sqrt{n}c_p}\delta\right)^{-\frac{2}{(1-\alpha)(2p+1)}} + 2, \tag{2.3}$$

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

The second result gives an upper bound on the variance component $\mathbb{E}[\|A(x_{\kappa(\delta)}^\delta - \mathbb{E}[x_{\kappa(\delta)}^\delta])\|^2]$ of the mean squared residual $\mathbb{E}[\|Ax_k - y^\delta\|^2]$. It indicates that the variance $\mathbb{E}[\|A(x_{k(\delta)}^\delta - \mathbb{E}[x_{k(\delta)}^\delta])\|^2]$ contributes only marginally to the mean squared residual $\mathbb{E}[\|Ax_{k(\delta)}^\delta - y^\delta\|^2]$, and consequently the squared residual $\|Ax_{k(\delta)}^\delta - y^\delta\|^2$ of individual realizations of SGD may be used instead for determining an appropriate stopping index.

Proposition 2.2. *Under assumption 1.1 with $\kappa(\delta) \geqslant \delta^{-\frac{2}{(1-\alpha)(\min(2p,r)+1)}}$ and $0 < r < 1$, there holds*

$$\mathbb{E}[\|A(x_{\kappa(\delta)}^\delta - \mathbb{E}[x_{\kappa(\delta)}^\delta])\|^2] = o(\delta^2), \quad as \quad \delta \to 0^+.$$

Now we can present the proof of theorem 1.1.

Proof. Set $1 < \tau^* < \tau$ and $\bar{k}(\delta) = [c\delta^{-\frac{2}{(1-\alpha)(\min(2p,r)+1)}}] + 2$ ($[\cdot]$ denotes taking the integral part of a real number), with $c = \left(\frac{\tau^*-1}{\sqrt{n}c_p}\right)^{-\frac{2}{(1-\alpha)(\min(2p,r)+1)}}$. By the definition of $k(\delta)$ in (1.3), the event $\mathcal{E} = \{k(\delta) \leqslant \bar{k}(\delta)\}$ is given by

$$\mathcal{E} = \{\exists i \in \{1, \ldots, \bar{k}(\delta)\} \quad \text{such that} \quad \|Ax_i^\delta - y^\delta\| \leqslant \tau\delta\}.$$

Thus, $\mathcal{E} \supset \{\|Ax_{\bar{k}(\delta)}^\delta - y^\delta\| \leqslant \tau\delta\}$. Consequently,

$$\begin{aligned}\mathbb{P}(k(\delta) \leqslant \bar{k}(\delta)) &\geqslant \mathbb{P}(\|Ax_{\bar{k}(\delta)}^\delta - y^\delta\| \leqslant \tau\delta) \\ &\geqslant \mathbb{P}\left(\|A(x_{\bar{k}(\delta)}^\delta - \mathbb{E}[x_{\bar{k}(\delta)}^\delta])\| \leqslant (\tau - \tau^*)\delta, \ \|A\mathbb{E}[x_{\bar{k}(\delta)}^\delta] - y^\delta\| \leqslant \tau^*\delta\right).\end{aligned}$$

By the choice of $\bar{k}(\delta)$, proposition 2.1 implies

$$\|A\mathbb{E}[x_{\bar{k}(\delta)}^\delta] - y^\delta\| \leqslant \tau^*\delta.$$

Consequently,

$$\begin{aligned}\mathbb{P}(k(\delta) \leqslant \bar{k}(\delta)) &\geqslant \mathbb{P}(\|A(x_{\bar{k}(\delta)}^\delta - \mathbb{E}[x_{\bar{k}(\delta)}^\delta])\| \leqslant (\tau - \tau^*)\delta) \\ &= 1 - \mathbb{P}(\|A(x_{\bar{k}(\delta)}^\delta - \mathbb{E}[x_{\bar{k}(\delta)}^\delta])\| > (\tau - \tau^*)\delta).\end{aligned}$$

Meanwhile, by Chebyshev's inequality [5, p 233], we have

$$\mathbb{P}(\|A(x_{\bar{k}(\delta)}^\delta - \mathbb{E}[x_{\bar{k}(\delta)}^\delta])\| > (\tau - \tau^*)\delta) \leqslant \frac{\mathbb{E}\|A(x_{\bar{k}(\delta)}^\delta - \mathbb{E}[x_{\bar{k}(\delta)}^\delta])\|^2}{(\tau - \tau^*)^2\delta^2}.$$

Therefore,

$$\mathbb{P}(k(\delta) \leqslant \bar{k}(\delta)) \geqslant 1 - \frac{\mathbb{E}\|A(x^\delta_{\bar{k}(\delta)} - \mathbb{E}[x^\delta_{\bar{k}(\delta)}])\|^2}{(\tau - \tau^*)^2\delta^2},$$

which together with proposition 2.2 directly implies

$$\mathbb{P}(k(\delta) \leqslant \bar{k}(\delta)) \to 1 \quad \text{as} \quad \delta \to 0^+.$$

This completes the proof of the theorem. □

Remark 2.1. The condition $r < 1$ is related to an apparent saturation phenomenon with SGD: for any $p > \frac{1}{2}$, the SGD iterate x_k^δ with *a priori* stopping can only achieve a convergence rate comparable with that for $p = \frac{1}{2}$ in the setting of assumption 1.1, at least for the current analysis [10]. It remains unclear whether this is an intrinsic drawback of SGD or due to limitations of the proof technique.

Remark 2.2. In practice, we prefer computing the residual with a frequency $\omega n \in \mathbb{N}$:

$$k_\omega(\delta) := \min\{\omega nk \ : \ k \in \mathbb{N}\ ,\ \|Ax^\delta_{\omega nk} - y^\delta\| \leqslant \tau\delta\}.$$

Since one of the numbers $[c\delta^{-\frac{2}{(1-\alpha)(\min(2p,r)+1)}}] + 2, \ldots, [c\delta^{-\frac{2}{(1-\alpha)(\min(2p,r)+1)}}] + \omega n + 1$ is of the form ωnk, with $k \in \mathbb{N}$, there holds

$$\mathbb{P}\left(k_\omega(\delta) \leqslant c\delta^{-\frac{2}{(1-\alpha)(\min(2p,r)+1)}} + \omega n + 1\right) \to 1 \quad \text{as} \quad \delta \to 0^+.$$

That is, the upper bound on the stopping index remains largely valid for a variant of the discrepancy principle (1.3) evaluated with a given frequency.

Remark 2.3. The finite-iteration termination property in theorem 1.1 relies heavily on the assumption $\alpha < 1$ in the definition of the stepsize schedule. Without this condition, theorem 1.1 (and thus also the convergence in probability) generally do not hold. Indeed, if $\mathrm{rank}(A) \geqslant 2$, $y^\dagger \neq 0$ and $\alpha > 1$, then there holds

$$\liminf_{\delta \to 0+} \mathbb{P}(k(\delta) = \infty) > 0. \tag{2.4}$$

To prove this assertion, let $k^* \in \mathbb{N}$ be such that $\eta_k\|A\|^2 \leqslant \frac{1}{2}$ for all $k \geqslant k^*$. Since $\mathrm{rank}(A) \geqslant 2$ and $y^\dagger \neq 0$, there exists an index $j \in \{1, \ldots, n\}$ such that $y^\dagger \notin \mathrm{span}(Aa_j)$. In view of the fact $Ax_k\chi_{\{i_1 = \ldots = i_{k^*-1} = j\}} \in \mathrm{span}(Aa_j)$, for $k \in \{1, \ldots, k^*\}$, there exists an $\eta > 0$ with

$$\mathbb{P}\left(\|Ax_k - y^\dagger\| \geqslant \eta, \quad \forall k \leqslant k^*\right) \geqslant \mathbb{P}\left(i_1 = \ldots = i_{k^*-1} = j\right) > 0.$$

Meanwhile for $k > k^*$, similar to (3.2) below, there holds

$$\begin{aligned}\|Ax_k - y^\dagger\| &\geqslant \|Ax_{k-1} - y^\dagger\| - \eta_{k-1}|(Ax_{k-1} - y^\dagger, e_{i_{k-1}})|\|AA^t e_{i_{k-1}}\| \\ &\geqslant \ldots \geqslant \|Ax_{k^*} - y^\dagger\| \prod_{i=k^*}^{k-1}(1 - \eta_i\|A\|^2).\end{aligned}$$

Using the elementary inequalities $1 + x \leqslant \mathrm{e}^x$ for all $x \in \mathbb{R}$ and $1 + x \geqslant \mathrm{e}^{x - x^2}$ for all $x \in [-\frac{1}{2}, 0]$ and the estimate (3.2) below, we deduce

$$\begin{aligned}\|Ax_k^\delta - y^\delta\| &\geqslant \|Ax_k - y^\dagger\| - \|A(x_k - x_k^\delta) - (y^\dagger - y^\delta)\| \\ &\geqslant \|Ax_{k^*} - y^\dagger\| \prod_{i=k^*}^{k-1}(1 - \|A\|^2\eta_i) - \delta \prod_{i=1}^{k-1}(1 + \|A\|^2\eta_i) \\ &\geqslant \|Ax_{k^*} - y^\dagger\| \exp\left(-c_0\|A\|^2 \sum_{i=k^*}^{k-1} i^{-\alpha} - c_0^2\|A\|^4 \sum_{i=k^*}^{k-1} i^{-2\alpha}\right) \\ &\quad - \delta \ \exp\left(c_0\|A\|^2 \sum_{i=1}^{k-1} i^{-\alpha}\right) \geqslant c'\|Ax_{k^*} - y^\dagger\| - c''\delta,\end{aligned}$$

with

$$c' := \mathrm{e}^{-c_0\|A\|^2 \sum_{i=1}^\infty i^{-\alpha} - c_0^2\|A\|^4 \sum_{i=1}^\infty i^{-2\alpha}} > 0 \quad \text{and} \quad c'' := \mathrm{e}^{c_0\|A\|^2 \sum_{i=1}^\infty i^{-\alpha}} < \infty.$$

So for small enough $\delta > 0$, there holds

$$\|Ax_k^\delta - y^\delta\| \chi_{\{\|Ax_i - y^\dagger\| \geqslant \eta, \quad \forall i \leqslant k^*\}} \geqslant c'\eta - c''\delta > \tau\delta.$$

Consequently,

$$\liminf_{\delta > 0} \mathbb{P}(k(\delta) = \infty) \geqslant \mathbb{P}\left(\|Ax_i - y^\dagger\| \geqslant \eta, \quad \forall i \leqslant k^*\right) > 0.$$

This shows the assertion (2.4).

3. The proof of theorem 1.2

In this section, we prove theorem 1.2. It employs the following proposition, which states that potential early stopping actually does not cause any problem.

Proposition 3.1. *For all $\varepsilon > 0$, there is a sequence $(k_\delta^-)_\delta$ with $k_\delta^- \to \infty$ for $\delta \to 0^+$, such that*

$$\|x_{k(\delta)}^\delta - x^\dagger\| \chi_{\{k(\delta) \leqslant k_\delta^-\}} \leqslant \varepsilon$$

for $\delta > 0$ small enough.

Proof. It suffices to show that for all $K \in \mathbb{N}$

$$\|x_{k(\delta)} - x^\dagger\| \chi_{\{k(\delta) \leqslant K\}} \to 0 \quad \text{as} \quad \delta \to 0^+. \tag{3.1}$$

In order to show this, we need the following two estimates for the iterated noise:

$$\|A(x_k^\delta - x_k) - (y^\delta - y^\dagger)\| \leqslant \delta \prod_{j=1}^{k-1}(1 + \eta_j\|A\|^2), \tag{3.2}$$

$$\|x_k^\delta - x_k\| \leqslant \delta\|A\| \sum_{j=1}^{k-1} \eta_j \prod_{i=1}^{j-1}(1 + \eta_i\|A\|^2), \tag{3.3}$$

with the conventions $\sum_{j=1}^{0} = 0$ and $\prod_{j=1}^{0} = 1$. We prove the estimates (3.2) and (3.3) by mathematical induction. Note that $a_i = A^t e_i$. For the estimate (3.2), by the triangle inequality and the defining relation (1.2) of SGD iteration,

$$\begin{aligned}&\|A(x_{k+1}^\delta - x_{k+1}) - (y^\delta - y^\dagger)\|\\ &\quad\leqslant \|A(x_k^\delta - x_k) - (y^\delta - y^\dagger)\| + \eta_k \|\left(A(x_k^\delta - x_k) - (y^\delta - y^\dagger), e_{i_k}\right) AA^t e_{i_k}\|\\ &\quad\leqslant \|A(x_k^\delta - x_k) - (y^\delta - y^\dagger)\| \left(1 + \eta_k \|A\|^2\right),\end{aligned}$$

and since $x_1 = x_1^\delta$, $\|A(x_1^\delta - x_1) - (y^\delta - y^\dagger)\| = \|y^\delta - y^\dagger\| \leqslant \delta$. For the estimate (3.3), we have $\|x_1^\delta - x_1\| = 0$ and

$$\|x_{k+1}^\delta - x_{k+1}\| \leqslant \|x_k^\delta - x_k\| + \eta_k \|A\| \|A(x_k^\delta - x_k) - (y^\delta - y^\dagger)\|,$$

so the claim follows using the estimate (3.2). Now, for each fixed K, since there are only finitely many different realizations of the first K SGD iterates, there is a (deterministic) $\eta > 0$, which depends on K, such that

$$\min_{k=1,\ldots,K} \left(\|Ax_k - y^\dagger\| - \eta\right) \chi_{\{\|Ax_k - y^\dagger\|>0\}} \geqslant 0, \tag{3.4}$$

where without loss of generality, we have assumed $y^\dagger \neq 0$. Therefore, using estimates (3.2) and (3.4),

$$\begin{aligned}&\|Ax_k^\delta - y^\delta\| \chi_{\{\|Ax_k - y^\dagger\|>0\}}\\ &\quad\geqslant \|Ax_k - y^\dagger\| \chi_{\{\|Ax_k - y^\dagger\|>0\}} - \|A(x_k - x_k^\delta) - (y^\dagger - y^\delta)\| \chi_{\{\|Ax_k - y^\dagger\|>0\}}\\ &\quad\geqslant \left(\eta - \delta \prod_{j=1}^{k-1} (1 + \eta_j \|A\|^2)\right) \chi_{\|Ax_k - y^\dagger\|>0\}} > \tau\delta \chi_{\{\|Ax_k - y^\dagger\|>0\}},\end{aligned}$$

for any $\delta < \frac{\eta}{\tau + \prod_{j=1}^{K-1}(1+\eta_j\|A\|^2)}$. Then by the definition of the discrepancy principle in (1.3), this implies

$$\{k(\delta) \leqslant K\} \subset \{\|Ax_{k(\delta)} - y^\dagger\| = 0\}$$

for $\delta > 0$ small enough. Meanwhile, since by construction $x_{k(\delta)} \in \mathcal{R}(A^t) = \mathcal{N}(A)^\perp$, $\|Ax_{k(\delta)} - y^\dagger\| = 0$ implies $x_{k(\delta)} = x^\dagger$, the minimum norm solution. The proof of (3.1) is concluded by

$$\begin{aligned}\|x_{k(\delta)}^\delta - x^\dagger\| \chi_{\{k(\delta)\leqslant K\}} &= \|x_{k(\delta)}^\delta - x_{k(\delta)}\| \chi_{\{k(\delta)\leqslant K\}}\\ &\leqslant \delta \|A\| \sum_{j=1}^{K-1} \eta_j \prod_{i=1}^{j-1} (1 + \eta_i \|A\|^2) \to 0\end{aligned}$$

for $\delta \to 0^+$, where we have used estimate (3.3). This completes the proof of the proposition. □

Now we can state the proof of theorem 1.2.

Proof of Theorem 1.2. Fix $\varepsilon > 0$. Proposition 3.1 and theorem 1.1 guarantee the existence of two sequences $(k_\delta^-)_\delta$, $(k_\delta^+)_\delta$, with $k_\delta^- \leqslant k_\delta^+ \leqslant c\delta^{-\frac{2}{(1-\alpha)(\min(2p,r)+1)}}$, $k_\delta^- \to \infty$ for $\delta \to 0^+$ and

$$\|x_{k(\delta)}^\delta - x^\dagger\| \chi_{\{k(\delta)\leqslant k_\delta^-\}} \leqslant \varepsilon \quad \text{for } \delta \text{ small enough}$$

and

$$\mathbb{P}\left(k(\delta) \leqslant k_\delta^+\right) \to 1 \quad \text{for} \quad \delta \to 0^+.$$

Consequently, for $\delta > 0$ small enough, there holds

$$\begin{aligned}
&\mathbb{P}(\|x_{k(\delta)}^\delta - x^\dagger\| > \varepsilon)\\
&\quad = \mathbb{P}(\|x_{k(\delta)}^\delta - x^\dagger\| > \varepsilon, k(\delta) \leqslant k_\delta^-) + \mathbb{P}(\|x_{k(\delta)}^\delta - x^\dagger\| > \varepsilon, k(\delta) > k_\delta^-)\\
&\quad = \mathbb{P}(\|x_{k(\delta)} - x^\dagger\| > \varepsilon, k(\delta) > k_\delta^-)\\
&\quad = \mathbb{P}(\|x_{k(\delta)} - x^\dagger\| > \varepsilon, k_\delta^- < k(\delta) \leqslant k_\delta^+) + \mathbb{P}(\|x_{k(\delta)} - x^\dagger\| > \varepsilon, k(\delta) > k_\delta^+)\\
&\quad \leqslant \mathbb{P}(\|x_{k(\delta)} - x^\dagger\| > \varepsilon, k_\delta^- < k(\delta) \leqslant k_\delta^+) + \mathbb{P}(k(\delta) > k_\delta^+).
\end{aligned}$$

In view of theorem 1.1, it remains to show that

$$\mathbb{P}(\|x_{k(\delta)} - x^\dagger\| > \varepsilon, k_\delta^- < k(\delta) \leqslant k_\delta^+) \to 0 \quad \text{for} \quad \delta \to 0^+.$$

To this end, let $\Omega_\delta := \{k_\delta^- \leqslant k(\delta) \leqslant k_\delta^+\}$ and we split the error into three parts in a customary way: approximation error, data propagation error and stochastic error. Specifically, by the triangle inequality, there are constants c_1 and c_2 such that

$$\begin{aligned}
\|x_{k(\delta)}^\delta - x^\dagger\|\chi_{\Omega_\delta} &= \sum_{k=k_\delta^-}^{k_\delta^+} \|x_k^\delta - x^\dagger\|\chi_{\{k(\delta)=k\}}\\
&\leqslant \sum_{k=k_\delta^-}^{k_\delta^+} \left(\|\mathbb{E}[x_k] - x^\dagger\| + \|\mathbb{E}[x_k] - \mathbb{E}[x_k^\delta]\| + \|x_k^\delta - \mathbb{E}[x_k^\delta]\|\right)\chi_{\{k(\delta)=k\}}\\
&\leqslant \sum_{k=k_\delta^-}^{k_\delta^+} \left(c_1(k-1)^{-(1-\alpha)p} + c_2\delta(k-1)^{\frac{1-\alpha}{2}} + \|x_k^\delta - \mathbb{E}[x_k^\delta]\|\right)\chi_{\{k(\delta)=k\}}\\
&\leqslant c_1\left(k_\delta^- - 1\right)^{-(1-\alpha)p} + c_2\delta\left(k_\delta^+ - 1\right)^{\frac{1-\alpha}{2}} + \sum_{k=k_\delta^-}^{k_\delta^+} \|x_k^\delta - \mathbb{E}[x_k^\delta]\|\chi_{\{k(\delta)=k\}},
\end{aligned}$$

where we have used [10, theorem 3.2] and lemma 4.1 below in the third line. The first two terms clearly tend to 0 for $\delta \to 0^+$ (since $k_\delta^- \to \infty$, and $\delta(k_\delta^+)^{\frac{1-\alpha}{2}} \to 0$, in view of theorem 1.1). By Markov's inequality [5, p 242] and the independence assumption between $k(\delta)$ and $x_{k(\delta)}^\delta$,

$$\begin{aligned}
\mathbb{P}\left(\sum_{k=k_\delta^-}^{k_\delta^+} \|x_k^\delta - \mathbb{E}[x_k^\delta]\|\chi_{\{k(\delta)=k\}} > \varepsilon'\right) &\leqslant \frac{\sum_{k=k_\delta^-}^{k_\delta^+} \mathbb{E}\left[\|x_k^\delta - \mathbb{E}[x_k^\delta]\|\chi_{\{k(\delta)=k\}}\right]}{\varepsilon'}\\
&= \frac{\sum_{k=k_\delta^-}^{k_\delta^+} \mathbb{E}\left[\|x_k^\delta - \mathbb{E}[x_k^\delta]\|\right]\mathbb{P}\left(k(\delta) = k\right)}{\varepsilon'}.
\end{aligned}$$

Now Jensen's inequality and proposition 4.1 below (with $s=0$, $\gamma < \min(\alpha, 1-\alpha)$ and $\beta < 1-\alpha$) give

$$\begin{aligned}
&\mathbb{P}\left(\sum_{k=k_\delta^-}^{k_\delta^+} \|x_k^\delta - \mathbb{E}[x_k^\delta]\|\chi_{\{k(\delta)=k\}} > \varepsilon'\right)\\
&\quad\leqslant \frac{\sum_{k=k_\delta^-}^{k_\delta^+} \sqrt{\mathbb{E}\left[\|x_k^\delta - \mathbb{E}[x_k^\delta]\|^2\right]}\,\mathbb{P}\left(k(\delta)=k\right)}{\varepsilon'}\\
&\quad\leqslant \frac{\sqrt{c((k_\delta^-)^{-\beta} + \delta^2(k_\delta^-)^{-\gamma})}\sum_{k=k_\delta^-}^{k_\delta^+} \mathbb{P}\left(k(\delta)=k\right)}{\varepsilon'}\\
&\quad= \frac{\sqrt{c((k_\delta^-)^{-\beta} + \delta(k_\delta^-)^{-\gamma})}\,\mathbb{P}\left(\Omega_\delta\right)}{\varepsilon'} \to 0
\end{aligned}$$

as $\delta \to 0^+$. Thus it follows that

$$\mathbb{P}\left(\|x_k(\delta) - x^\dagger\| > \varepsilon, k_\delta^- < k(\delta) \leqslant k_\delta^+\right) \to 0$$

as $\delta \to 0^+$. This completes the proof of the theorem. □

Remark 3.1. Clearly, with $k_\omega(\delta)$ given as in remark 2.2, there holds $\mathbb{P}\left(\|x_{k_\omega(\delta)} - x^\dagger\| \geqslant \varepsilon\right) \to 0$ for $\delta \to 0^+$. That is, the convergence remains valid for the variant of the discrepancy principle (1.3) evaluated with a frequency.

4. The proofs of propositions 2.1 and 2.2

In this part, we prove propositions 2.1 and 2.2, which are used in the proof of the theorems 1.1 and 1.2. We shall use the following result from [10, theorem 3.1] frequently. Note that $\|B^{\frac{1}{2}}(x_k - x^\dagger)\| = \|Ax_k - y^\dagger\|/\sqrt{n}$.

Lemma 4.1. *Let assumption 1.1 be fulfilled, then for* $s \in \{0, \frac{1}{2}\}$ *and* $c_{p,s} := \left(\frac{(p+s)(1-\alpha)}{c_0 e(2^{1-\alpha}-1)}\right)^{p+s}\|w\|$, *there holds*

$$\|B^s(x_{k+1} - x^\dagger)\| \leqslant c_{p,s}k^{-(p+s)(1-\alpha)}.$$

4.1. The proof of proposition 2.1

Proof. We may assume $k^* > 2$. By the definition of $k^*(\delta)$ and the triangle inequality

$$\begin{aligned}
\tau^*\delta &\leqslant \|A\mathbb{E}[x_{k^*-1}^\delta] - y^\delta\|\\
&\leqslant \|A\mathbb{E}[x_{k^*-1}] - y^\dagger\| + \|A\mathbb{E}[x_{k^*-1}^\delta - x_{k^*-1}] + (y^\dagger - y^\delta)\|.
\end{aligned}$$

By lemma 4.1, the term $\|A\mathbb{E}[x_{k^*-1}] - y^\dagger\|$ is bounded by

$$\|A\mathbb{E}[x_{k^*-1}] - y^\dagger\| \leqslant c_p(k^*-2)^{-(p+\frac{1}{2})(1-\alpha)}, \quad \text{with} \quad c_p = \sqrt{n}c_{p,\frac{1}{2}}. \tag{4.1}$$

Next we claim

$$\|A\mathbb{E}[x^\delta_{k^*-1} - x_{k^*-1}] + (y^\dagger - y^\delta)\| \leqslant \delta. \tag{4.2}$$

Combining (4.1) with (4.2) immediately implies the desired assertion. It remains to show the claim (4.2). To this end, we employ the filter of the Landweber method. The relation (2.1) implies that $\mathbb{E}[x^\delta_k]$ satisfies the following recursion

$$A\mathbb{E}[x^\delta_{k+1}] - y^\delta = \left(I - \frac{\eta_k}{n}AA^t\right)\left(A\mathbb{E}[x^\delta_k] - y^\delta\right).$$

Using this yields

$$A\mathbb{E}[x^\delta_k] - y^\delta = \prod_{j=1}^{k-1}\left(I - \frac{\eta_j}{n}AA^t\right)\left(Ax_1 - y^\delta\right), \tag{4.3}$$

and consequently, by the choice of c_0,

$$\|A\mathbb{E}[x^\delta_k - x_k] + (y^\dagger - y^\delta)\| = \left\|\prod_{j=1}^{k-1}\left(I - \frac{\eta_j}{n}AA^t\right)\left(y^\dagger - y^\delta\right)\right\| \leqslant \delta. \tag{4.4}$$

This completes the proof of the proposition. □

4.2. Proof of proposition 2.2

The proof of proposition 2.2 employs several technical estimates [10].

Lemma 4.2. *For any $j < k$, and any symmetric and positive semidefinite operator S and stepsizes $\eta_j \in (0, \|S\|^{-1}]$ and $p \geqslant 0$, there holds*

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

Next we recall two useful estimates taken from [10].

Lemma 4.3. *For $\eta_j = \eta_0 j^{-\alpha}$ with $\alpha \in (0,1)$, $\beta \in [0,1]$ and $r \geqslant 0$, there hold*

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

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

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

$$c_{\alpha,\beta,r} = 2^r\eta_0^{2-r}\begin{cases}\dfrac{2\alpha+\beta}{2\alpha+\beta-1}, & 2\alpha+\beta > 1,\\ 2, & 2\alpha+\beta = 1, \quad \textit{and}\\ \dfrac{2^{2\alpha+\beta-1}}{1-2\alpha-\beta}, & 2\alpha+\beta < 1,\end{cases}$$

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

The next result gives an important recursion between the variance estimate.

Lemma 4.4. *Let assumption 1.1 be fulfilled. Then for the SGD iterate* x_k^δ, *with* $\phi_j^s = \|B^{\frac{1}{2}+s}\Pi_{j+1}^k(B)\|$, *there holds*

$$\begin{aligned}&\mathbb{E}[\|B^s(x_{k+1}^\delta - \mathbb{E}[x_{k+1}^\delta])\|^2]\\ &\quad\leqslant \sum_{j=1}^k \eta_j^2(\phi_j^s)^2\left(c_s\mathbb{E}[\|B^s\left(x_j^\delta - \mathbb{E}[x_j^\delta]\right)\|^2] + 2c_p j^{-2(1-\alpha)(p+\frac{1}{2})} + 2\delta^2\right),\end{aligned}$$

with $s \in \{0, \frac{1}{2}\}$ *and* c_s, c_p *given below.*

Proof. By [10, theorem 3.3] and the bias variance decomposition, the left-hand side (LHS) is bounded by

$$\begin{aligned}\mathrm{LHS} &\leqslant \sum_{j=1}^k \eta_j^2(\phi_j^s)^2\mathbb{E}\left[\|Ax_j^\delta - y^\delta\|^2\right.\\ &= \sum_{j=1}^k \eta_j^2(\phi_j^s)^2\left(\mathbb{E}[\|A\left(x_j^\delta - \mathbb{E}[x_j^\delta]\right)\|^2] + \|A\mathbb{E}[x_j^\delta] - y^\delta\|^2\right).\end{aligned}$$

Now by the triangle inequality and (4.4),

$$\begin{aligned}\mathrm{LHS} \leqslant \sum_{j=1}^k \eta_j^2(\phi_j^s)^2 &\Bigg(\mathbb{E}[\|A\left(x_j^\delta - \mathbb{E}[x_j^\delta]\right)\|^2] + \Big(\|A\mathbb{E}[x_j] - y^\dagger\|\\ &+ \|A\left(\mathbb{E}[x_j^\delta] - \mathbb{E}[x_j]\right) - \left(y^\delta - y^\dagger\right)\|\Big)^2\Bigg)\end{aligned}$$

Since $\|A\mathbb{E}[x_1] - y^\dagger\| = \|y^\dagger\|$, and

$$\|A\mathbb{E}[x_j] - y^\dagger\| \leqslant \sqrt{n}c_{p,\frac{1}{2}}(j-1)^{-(p+\frac{1}{2})(1-\alpha)} \leqslant \sqrt{n}c_{p,\frac{1}{2}}2^{(p+\frac{1}{2})(1-\alpha)}j^{-(p+\frac{1}{2})(1-\alpha)}$$

for $j \geqslant 2$ by lemma 4.1. Thus, with $c_p := \left(\max\{\|y^\dagger\|, \sqrt{n}c_{p,\frac{1}{2}}2^{(p+\frac{1}{2})(1-\alpha)}\}\right)^2$,

$$\begin{aligned}\mathrm{LHS} \leqslant \sum_{j=1}^k \eta_j^2(\phi_j^s)^2 &\left(n^{2s}\|A\|^{4(\frac{1}{2}-s)}\mathbb{E}[\|B^s\left(x_j^\delta - \mathbb{E}[x_j^\delta]\right)\|^2]\right.\\ &\left. + 2c_p j^{-2(1-\alpha)(p+\frac{1}{2})} + 2\delta^2\right)\end{aligned}$$

which completes the proof of the lemma with $c_s = n^{2s}\|A\|^{4(\frac{1}{2}-s)}$. □

The next result gives a sharp estimate on $\mathbb{E}[\|B^s(x_k^\delta - \mathbb{E}[x_k^\delta])\|^2]$.

Proposition 4.1. *Let assumption 1.1 be fulfilled. Then for the SGD iterate x_k^δ, the mean squared error $\mathbb{E}[\|B^s(x_k^\delta - \mathbb{E}[x_k^\delta])\|^2]$ with $s \in \{0, \frac{1}{2}\}$ satisfies*

$$\mathbb{E}[\|B^s(x_k^\delta - \mathbb{E}[x_k^\delta])\|^2] \leqslant c(\alpha, p, n, s, \beta, \gamma)(k^{-\beta} + \delta^2 k^{-\gamma})$$

for $\beta < \min((1+2s)(1-\alpha), (1+2p)(1-\alpha)+\alpha)$ and $\gamma < \min(\alpha, 1-\alpha)$.

Proof. Lemma 4.4 implies that the weighted mean squares error $d_j^s = \mathbb{E}[\|B^s(x_k^\delta - x^\dagger)\|^2]$ satisfies the following recursion

$$d_{k+1}^s \leqslant \sum_{j=1}^k \eta_j^2 (\phi_j^s)^2 \left(c_s d_j^s + 2c_p j^{-2(1-\alpha)(p+s)} + 2\delta^2\right) \tag{4.5}$$

Now we prove the desired assertion by mathematical induction (with $\beta = (2p+1)(1-\alpha)$):

$$d_k^s \leqslant c(k^{-\beta} + \delta^2 k^{-\gamma}),$$

where the constant $c \geqslant 1$ is to be determined. This assertion holds trivially for all finite k, up to k^*, provided that c is sufficiently large. Now suppose the assertion holds for $k \geqslant k^*$, and we prove the assertion for $k+1$. Indeed, it follows from the recursion (4.5), the induction hypothesis and since $\beta < 2(1-\alpha)(p+\frac{1}{2})$, that

$$\begin{aligned} d_{k+1}^s &\leqslant \sum_{j=1}^k \eta_j^2 (\phi_j^s)^2 (c_s c(j^{-\beta} + j^{-\gamma}\delta^2) + 2c_p j^{2(1-\alpha)(p+s)} + 2\delta^2) \\ &\leqslant c_s c \sum_{j=1}^k \eta_j^2 (\phi_j^s)^2 j^{-\beta} + (c_s c + 2)\delta^2 \sum_{j=1}^k \eta_j^2 (\phi_j^s)^2 \\ &\quad + 2c_p \sum_{j=1}^k \eta_j^2 (\phi_j^s)^2 j^{-2(1-\alpha)(p+\frac{1}{2})} \\ &\leqslant (c_s c + 2c_p) \sum_{j=1}^k \eta_j^2 (\phi_j^s)^2 j^{-\beta'} + (c_s c + 2)\delta^2 \sum_{j=1}^k \eta_j^2 (\phi_j^s)^2. \end{aligned}$$

with $\beta' = \min(\beta, (1+2p)(1-\alpha))$. Without loss of generality, we may assume that $\beta' \geqslant 1 - 2\alpha$. By lemmas 4.2 and 4.3, the first sum is bounded by

$$\begin{aligned} \sum_{j=1}^k \eta_j^2 (\phi_j^s)^2 j^{-\beta'} &\leqslant \mathrm{e}^{-2} c_{\alpha,\beta',1+2s} k^{-(1+2s)(1-\alpha)+\max(0,1-2\alpha-\beta')} \\ &\quad + \mathrm{e}^{-1} c'_{\alpha,\beta',1} \|B\| k^{-(\alpha+\beta')} \ \ln\ k + c_0^2 \|B\|^2 k^{-(2\alpha+\beta')}. \end{aligned} \tag{4.6}$$

Since $\beta' + \alpha > \beta$ and $\max(0, 1-2\alpha-\beta') = 0$, thus,

$$\begin{aligned} \sum_{j=1}^k \eta_j^2 \phi_j^2 j^{-\beta'} &\leqslant \Big(\mathrm{e}^{-2} c_{\alpha,\beta',1+2s} k^{-(1+2s)(1-\alpha)+\beta} \ \ln\ k \\ &\quad + \mathrm{e}^{-1} c'_{\alpha,\beta',1} \|B\| k^{-(\alpha+\beta')+\beta} \ \ln\ k + c_0^2 \|B\|^2 k^{-\alpha}\Big) k^{-\beta}. \end{aligned}$$

Meanwhile, with $-(1+2s)(1-\alpha)+\max(0,1-2\alpha)=-\min((1+2s)(1-\alpha),\alpha+2s(1-\alpha))$, we obtain

$$\begin{aligned}\sum_{j=1}^{k}\eta_j^2(\phi_j)^2 &\leqslant \mathrm{e}^{-2}c_{\alpha,0,2}k^{-\min((1+2s)(1-\alpha),\alpha+2s(1-\alpha))}\\ &\qquad + \mathrm{e}^{-1}c'_{\alpha,0,1}\|B\|k^{-\alpha}\ \ln\ k + c_0^2\|B\|^2k^{-2\alpha}\\ &\leqslant \left(\mathrm{e}^{-2}c_{\alpha,0,1+2s}k^{-\min((1-\alpha),\alpha)+\gamma}\right.\\ &\qquad \left. + \mathrm{e}^{-1}c'_{\alpha,0,1}\|B\|k^{-\alpha+\gamma}\ \ln\ k + c_0^2\|B\|^2k^{-2\alpha+\gamma}\right)k^{-\gamma}\end{aligned}$$

Combining the preceding estimates yields

$$\begin{aligned}d_{k+1} \leqslant{}& (cc_s+2c_p)\left(\mathrm{e}^{-2}c_{\alpha,\beta',1+2s}k^{-(1+2s)(1-\alpha)+\beta}\ \ln\ k\right.\\ &\left. + \mathrm{e}^{-1}c'_{\alpha,\beta',1}\|B\|k^{-(\alpha+\beta')+\beta}\ \ln\ k + c_0^2\|B\|^2k^{-\alpha}\right)k^{-\beta}\\ &+ (c_sc+2)\delta^2\left(\mathrm{e}^{-2}c_{\alpha,0,1+2s}k^{-\min((1-\alpha),\alpha)+\gamma}\right.\\ &\left. + \mathrm{e}^{-1}c'_{\alpha,0,1}\|B\|k^{-\alpha+\gamma}\ \ln\ k + c_0^2\|B\|^2k^{-2\alpha+\gamma}\right)k^{-\gamma}.\end{aligned}$$

Since by assumption, $\beta<(1+2s)(1-\alpha)$, $\beta<\alpha+\beta'$ and $\gamma<\min(\alpha,1-\alpha)$, there exists k^* such that for all $k\geqslant k^*$

$$\begin{aligned}(c_s+2c_p)\left(\mathrm{e}^{-2}c_{\alpha,\beta',1+2s}k^{-(1+2s)(1-\alpha)+\beta}\ \ln\ k+\mathrm{e}^{-1}c'_{\alpha,\beta',1}\|B\|k^{-(\alpha+\beta')+\beta}\ \ln\ k+c_0^2\|B\|^2k^{-2\alpha}\right)<\frac{1}{4},\\ (c_s+2)\delta^2\left(\mathrm{e}^{-2}c_{\alpha,0,1+2s}k^{-\min((1-\alpha),\alpha)+\gamma}+\mathrm{e}^{-1}c'_{\alpha,0,1}\|B\|k^{-\alpha+\gamma}\ \ln\ k+c_0^2\|B\|^2k^{-2\alpha+\gamma}\right)<\frac{1}{4}.\end{aligned}$$

Thus, with this choice of k^* and $k\geqslant k^*$,

$$\begin{aligned}d_{k+1} &\leqslant \frac{c}{4}\left(k^{-\beta}+\delta^2k^{-\gamma}\right) \leqslant c\frac{(1+k^{-1})^{\beta}}{4}\left((k+1)^{-\beta}+\delta^2(k+1)^{-\gamma}\right)\\ &< c\left((k+1)^{-\beta}+\delta^2(k+1)^{-\gamma}\right)\end{aligned}$$

and we obtain the desired assertion. □

Remark 4.1. The n factor in the estimate is due to the variance inflation of using stochastic gradients in place of gradient in SGD. This factor can be reduced by suitable variance reduction techniques, e.g., mini-batching and stochastic variance reduced gradient [12]. Note that with [10, theorems 3.1 and 3.2] and $s=0$, proposition 4.1 gives an improved (regarding the exponents) *a priori* bound for the mean squared error $\mathbb{E}[\|x_k^\delta-x^\dagger\|^2]$.

Last, using lemma 4.4 and proposition 4.1, we can prove proposition 2.2.

Proof of Proposition 2.2. Using lemma 4.4 and proposition 4.1 with $s=\frac{1}{2}$ and $c=c(\alpha,p,n,s,\beta,\gamma)$, we deduce

$$\mathbb{E}[\|A(x^\delta_{\kappa(\delta)}-\mathbb{E}[x^\delta_{\kappa(\delta)}])\|^2]\leqslant nc\left(\kappa(\delta)^{-\beta}+\delta^2\kappa(\delta)^{-\gamma}\right).$$

We choose $\gamma>0$. If $p<\frac{1}{2}$ and $r>2p$, then we can choose $\beta>(1-\alpha)(2p+1)$, so with the choice $\kappa(\delta)=\delta^{-\frac{2}{(1-\alpha)(2p+1)}}$, the claim follows. Otherwise, if $p\geqslant\frac{1}{2}$, then we can choose $\beta>$

$(1-\alpha)(r+1)$, so with the choice $\kappa(\delta)=\delta^{-\frac{2}{(1-\alpha)(r+1)}}$ the claim again follows. This completes the proof of the proposition. □

5. Numerical experiments and discussions

Now we provide numerical experiments to complement the theoretical analysis. Three model examples, i.e., `phillips` (mildly ill-posed, smooth), `gravity` (severely ill-posed, medium smooth) and `shaw` (severely ill-posed, nonsmooth), are taken from the open source `MATLAB` package Regutools [7], available at http://people.compute.dtu.dk/pcha/Regutools/ (last accessed on April 14, 2020). The problems cover a variety of setting, e.g., different solution smoothness and degree of ill-posedness. These examples are discretizations of Fredholm/Volterra integral equations of the first kind, by means of either the Galerkin approximation with piecewise constant basis functions or quadrature rules. All the examples are discretized into a linear system of size $n=m=1000$. In addition, we generate a synthetic example, termed `smoothed-phillips`, whose exact solution $x^\dagger$ is first generated by $\bar{x}^\dagger=A^tAA^t\bar{y}^\dagger$ and then normalized to have unit maximum, i.e., $x^\dagger=\bar{x}^\dagger/\|\bar{x}^\dagger\|_{\ell^\infty}$, where A is the system matrix and $\bar{y}^\dagger$ the exact data from `phillips`, and the corresponding exact data is formed by $y^\dagger=Ax^\dagger$. By its very construction, the solution $x^\dagger$ satisfies assumption 1.1(ii) with an exponent $p>2$, and thus it is very smooth in some sense. Throughout, the noisy data y^δ is generated according to

$$y_i^\delta:=y_i^\dagger+\delta\max_j(|y_j^\dagger|)\xi_i,\quad i=1,\ldots,n,$$

where the i.i.d. random variables ξ_i follow the standard Gaussian distribution (with zero mean and unit variance), and $\delta>0$ denotes the relative noise level (by slightly abusing the notation). The parameter c_0 in the stepsize schedule in assumption 1.1(i) is set to $(\max_i\|a_i\|^2)^{-1}$, the exponent α is taken from the set $\{0.1,0.3,0.5\}$, and unless otherwise stated, the stopping criterion is tested every 100 SGD iterations (see remarks 2.2 and 3.1). SGD is always initialized with $x_1=0$, and the maximum number of epochs is fixed at 5000, where one epoch refers to n SGD iterations. The parameter τ in the discrepancy principle (1.3) is fixed at $\tau=1.2$. All the statistical quantities presented below are computed from 100 independent runs.

5.1. Optimality

First, we verify the optimality of the discrepancy principle (1.3), against an order optimal regularization method. There are many possible choices, e.g., Landweber method and conjugate gradient method [4, chapters 6 and 7]. In this work, we employ the Landweber method as the benchmark. The Landweber method generally converges steadily although often slowly. However, it is known to be an order optimal regularization method with infinite qualification [4, theorem 6.5, p 159], when terminated by the discrepancy principle (2.2), and further, it is the population version of SGD [the expected iterates $(\mathbb{E}[x_k^\delta])_{k\in\mathbb{N}}$ are exactly the Landweber iterates; see (2.1)], and thus it serves a good benchmark for performance comparison in terms of the convergence rate. For the comparison, the Landweber method is initialized with $x_1=0$, with a constant stepsize $1/\|A\|^2$, and it is terminated with the discrepancy principle (2.2) with $\tau^*=1.2$ (i.e., the same as for SGD) with the maximum number of iterations being fixed at 5000. The numerical results for the examples are summarized in tables 1–4. In the tables, $e_{\rm sgd}$ and std($e_{\rm sgd}$) denote the (sample) mean and the (sample) standard deviation of the (squared)

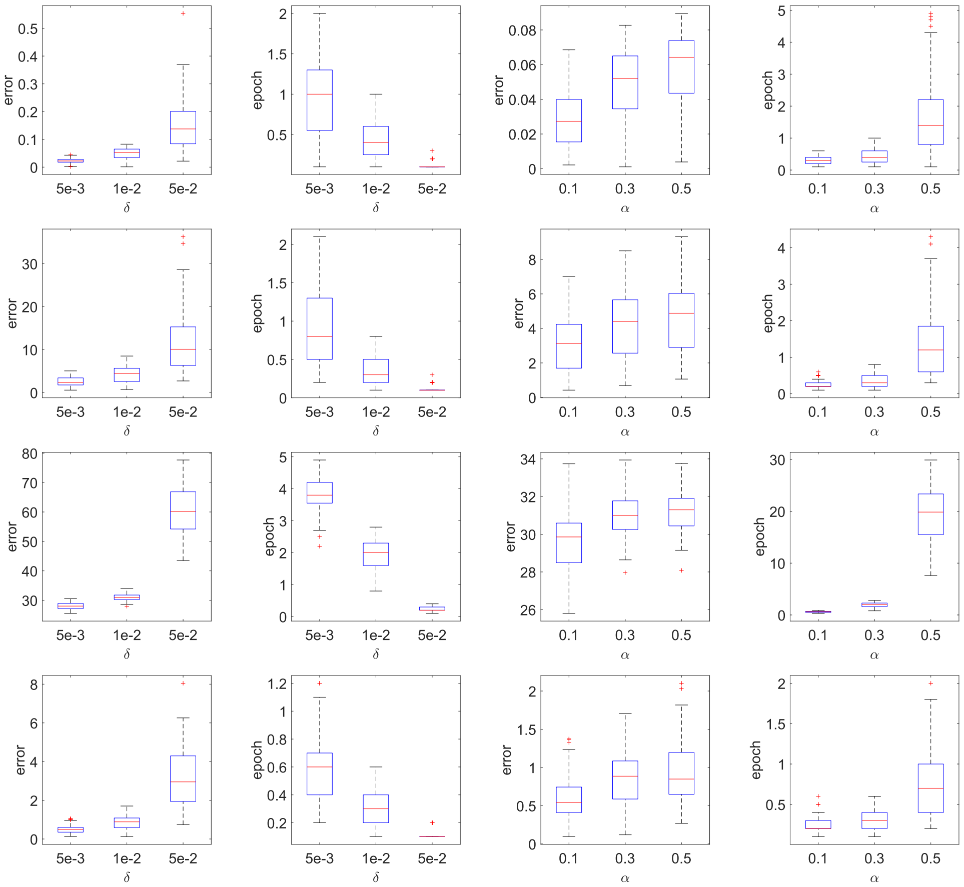

Figure 1. Box plots for the error $\|x_{k_\delta}^\delta - x^\dagger\|^2$ and the stopping index k_δ by SGD. The first two columns are obtained by SGD with $\alpha = 0.3$, whereas the last two columns are for the noise level $\delta = 1 \times 10^{-2}$. The rows from top to bottom refer to `phillips`, `gravity`, `shaw` and `smoothed-phillips`, respectively.

error $\|x_{k_\delta}^\delta - x^\dagger\|^2$, respectively, i.e.,

$$e_{\mathrm{sgd}} = \mathbb{E}[\|x_{k_\delta}^\delta - x^\dagger\|^2] \quad \text{and} \quad \mathrm{std}(e_{\mathrm{sgd}}) = \mathbb{E}[(\|x_{k^\delta}^\delta - x^\dagger\|^2 - e_{\mathrm{sgd}})^2]^{\frac{1}{2}},$$

and $k_{\mathrm{sgd}} = \mathbb{E}[k_\delta]$ is the mean stopping index for SGD, in terms of the number of epochs. Likewise e_{lm} and k_{lm} denote the squared reconstruction error and stopping index, respectively, of the Landweber method, terminated according to the discrepancy principle (2.2).

The numerical results allow drawing a number of interesting observations. First, the exponent α in the stepsize schedule exerts a strong influence on the (expected) stopping index k_{sgd}. At low noise levels (i.e., small δ), k_{sgd} increases dramatically with the value of α. Meanwhile, for any fixed α, the error e_{sgd} increases steadily with the noise level δ, exhibiting the convergence behavior indicated in theorem 1.2. Further, for each fixed δ, the error e_{sgd} is largely comparable for all different α values, although k_{sgd} increases with α. This behavior is qualitatively in good agreement with theorem 1.1: the upper bound scales as $O(\delta^{-\frac{2}{(1-\alpha)(\min(2p,r)+1)}})$. Thus, in practice, in order to obtain relatively efficient SGD, one prefers small α values. Second, in

Table 1. Comparison between SGD and LM for `phillips`.

	$\alpha = 0.1$			$\alpha = 0.3$			$\alpha = 0.5$			LM	
δ	e_{sgd}	$\text{std}(e_{\text{sgd}})$	k_{sgd}	e_{sgd}	$\text{std}(e_{\text{sgd}})$	k_{sgd}	e_{sgd}	$\text{std}(e_{\text{sgd}})$	k_{sgd}	e_{lm}	k_{lm}
1×10^{-3}	8.60×10^{-3}	4.53×10^{-3}	1.424	8.53×10^{-3}	4.42×10^{-3}	4.189	8.34×10^{-3}	4.60×10^{-3}	52.29	5.72×10^{-3}	361
5×10^{-3}	1.70×10^{-2}	8.41×10^{-3}	0.458	2.31×10^{-2}	8.81×10^{-3}	0.975	2.48×10^{-2}	7.38×10^{-3}	6.032	2.26×10^{-2}	128
1×10^{-2}	2.82×10^{-2}	1.62×10^{-2}	0.281	4.72×10^{-2}	2.07×10^{-2}	0.433	5.78×10^{-2}	2.04×10^{-2}	1.647	5.76×10^{-2}	51
5×10^{-2}	1.41×10^{-1}	9.70×10^{-2}	0.157	1.49×10^{-1}	9.01×10^{-2}	0.116	2.11×10^{-1}	9.69×10^{-2}	0.173	2.19×10^{-1}	15

Table 2. Comparison between SGD and LM for `gravity`.

	$\alpha = 0.1$			$\alpha = 0.3$			$\alpha = 0.5$			LM	
δ	e_{sgd}	$\mathrm{std}(e_{\mathrm{sgd}})$	k_{sgd}	e_{sgd}	$\mathrm{std}(e_{\mathrm{sgd}})$	k_{sgd}	e_{sgd}	$\mathrm{std}(e_{\mathrm{sgd}})$	k_{sgd}	e_{lm}	k_{lm}
1×10^{-3}	6.71×10^{-1}	2.61×10^{-1}	1.960	7.46×10^{-1}	2.73×10^{-1}	9.316	7.78×10^{-1}	2.49×10^{-1}	198.5	7.25×10^{-1}	640
5×10^{-3}	2.00×10	8.91×10^{-1}	0.451	2.53×10	1.12×10	0.880	2.76×10	1.14×10	6.217	2.44×10	95
1×10^{-2}	3.12×10	1.57×10	0.250	4.33×10	1.92×10	0.361	4.74×10	2.07×10	1.366	4.02×10	50
5×10^{-2}	9.07×10	5.31×10	0.143	1.15×10^{1}	6.61×10	0.107	1.52×10^{1}	7.46×10	0.135	1.66×10^{1}	9

Table 3. Comparison between SGD and LM for `shaw`.

	$\alpha = 0.1$			$\alpha = 0.3$			$\alpha = 0.5$			LM	
δ	e_{sgd}	$\mathrm{std}(e_{\mathrm{sgd}})$	k_{sgd}	e_{sgd}	$\mathrm{std}(e_{\mathrm{sgd}})$	k_{sgd}	e_{sgd}	$\mathrm{std}(e_{\mathrm{sgd}})$	k_{sgd}	e_{lm}	k_{lm}
1×10^{-3}	8.29×10	9.35×10^{-2}	57.73	8.47×10	5.59×10^{-2}	891.3	2.01×10^{1}	5.64×10^{-1}	5000	1.28×10^{1}	5000
5×10^{-3}	2.77×10^{1}	1.24×10	0.948	2.80×10^{1}	1.16×10	3.811	2.82×10^{1}	1.02×10	51.69	2.81×10^{1}	189
1×10^{-2}	2.96×10^{1}	1.65×10	0.597	3.10×10^{1}	1.14×10	1.938	3.12×10^{1}	1.08×10	19.71	3.11×10^{1}	117
5×10^{-2}	5.02×10^{1}	1.08×10^{1}	0.155	6.07×10^{1}	8.08×10	0.250	6.70×10^{1}	7.41×10	0.818	6.85×10^{1}	22

Table 4. Comparison between SGD and LM for `smoothed-phillips`.

	$\alpha = 0.1$			$\alpha = 0.3$			$\alpha = 0.5$			LM	
δ	e_{sgd}	$\mathrm{std}(e_{\mathrm{sgd}})$	k_{sgd}	e_{sgd}	$\mathrm{std}(e_{\mathrm{sgd}})$	k_{sgd}	e_{sgd}	$\mathrm{std}(e_{\mathrm{sgd}})$	k_{sgd}	e_{lm}	k_{lm}
1×10^{-3}	1.63×10^{-1}	6.87×10^{-2}	1.348	1.59×10^{-1}	5.88×10^{-2}	4.030	1.55×10^{-1}	6.09×10^{-2}	48.02	1.51×10^{-3}	29
5×10^{-3}	3.92×10^{-1}	2.08×10^{-1}	0.367	5.06×10^{-1}	2.05×10^{-1}	0.591	4.92×10^{-1}	1.99×10^{-1}	2.683	1.38×10^{-2}	18
1×10^{-2}	5.95×10^{-1}	2.64×10^{-1}	0.242	8.57×10^{-1}	3.73×10^{-1}	0.303	9.46×10^{-1}	3.93×10^{-1}	0.774	4.06×10^{-2}	15
5×10^{-2}	2.98×10	1.44×10	0.163	3.20×10	1.51×10	0.107	4.35×10	2.13×10	0.130	7.19×10^{-1}	9

terms of accuracy (measured by the mean squared error), SGD is competitive with the classical Landweber method for `phillips`, `gravity` and `shaw`: e_{sgd} and e_{lm} are fairly close to each other in most cases, and e_{sgd} can be smaller than e_{lm}, which fully confirms the order-optimality of the discrepancy principle (1.3) for SGD for low regularity solutions, and also confirming the convergence in theorem 1.2. In fact, empirically, the error seems to converge not only in probability, but also in L^2. A close inspection on the stopping index k_{sgd} is very telling: when the noise level δ is medium to large, the stopping index k_{sgd} of SGD, determined by (1.3), is ten-fold smaller than that for the Landweber method in terms of epoch count. In particular, when the noise level δ is relatively high, SGD can actually deliver an accurate solution within less than one epoch, i.e., going through only a fraction of all the available data points. Thus, in this regime, SGD is much more efficient than the Landweber method. These observations are valid for all the examples, despite their dramatic difference in degree of ill-posedness and solution smoothness. However, for `smoothed-phillips`, the achieved accuracy by SGD is far below than that by the Landweber method for all three exponents α. This suboptimality in convergence rate is attributed to the saturation phenomenon for SGD, due to the dominance of the computational variance, when the true solution $x^\dagger$ is very smooth. The effect of the variance component will be examined more closely below in section 5.2.

The example `shaw` is challenging for numerical recovery, since the solution is far less smooth, and at low noise level $\delta = 1 \times 10^{-3}$, the discrepancy principle (2.2) cannot be reached even after 5000 Landweber iterations, see table 3. A similar behavior is also observed for SGD with $\alpha = 0.3$ and $\alpha = 0.5$. Nonetheless, with $\alpha = 0.1$, the discrepancy principle (1.3) can be reached by SGD after a few hundred epochs, clearly showing the surprisingly beneficial effect of SGD noise for low-regularity solutions.

Next we examine more closely the performance of individual samples. The boxplots are shown in figure 1 for the examples at two different scenarios, i.e., fixed α and fixed δ. On each box, the central mark indicates the median, and the bottom and top edges of the box indicate the 25th and 75th percentiles, respectively; The whiskers extend to the most extreme data points not considered outliers, and the outliers are plotted individually using the '+' symbol. It is observed that for a fixed α, on average the error $\|x_{k(\delta)}^\delta - x^\dagger\|^2$ increases with the noise level δ samplewise, and also its distribution broadens. However, the required number of iterations to fulfill the discrepancy principle (1.3) decreases dramatically, as the noise level δ increases, concurring with the preceding observation that SGD is especially efficient for data with high noise levels. Meanwhile, with the noise level δ fixed, the value of α does not change the results much overall. However, a larger α can potentially make the percentile box larger and also more outliers, as shown by the results for `gravity` in figure 1, and thus give less accurate results. This observation is counter-intuitive in that smaller variance does not immediately lead to better accuracy. This might be related to the delicate interplay between the total error and various problem / algorithmic parameters, e.g., α and p. Further, the outliers in the boxplots mostly lie above the box. These observations are typical for all the examples.

5.2. How influential is the variance?

Now we examine more closely the dynamics of the SGD iteration via the bias-variance decomposition of the error $\mathbb{E}[\|x_k^\delta - x^\dagger\|^2]$ and residual $\mathbb{E}[\|Ax_k^\delta - y^\delta\|^2]$:

$$\begin{aligned}\mathbb{E}[\|x_k^\delta - x^\dagger\|^2] &= \|\mathbb{E}[x_k^\delta] - x^\dagger\|^2 + \mathbb{E}[\|x_k^\delta - \mathbb{E}[x_k^\delta]\|^2],\\ \mathbb{E}[\|Ax_k^\delta - y^\delta\|^2] &= \|A\mathbb{E}[x_k^\delta] - y^\delta\|^2 + \mathbb{E}[\|A(x_k^\delta - \mathbb{E}[x_k^\delta])\|^2].\end{aligned}$$

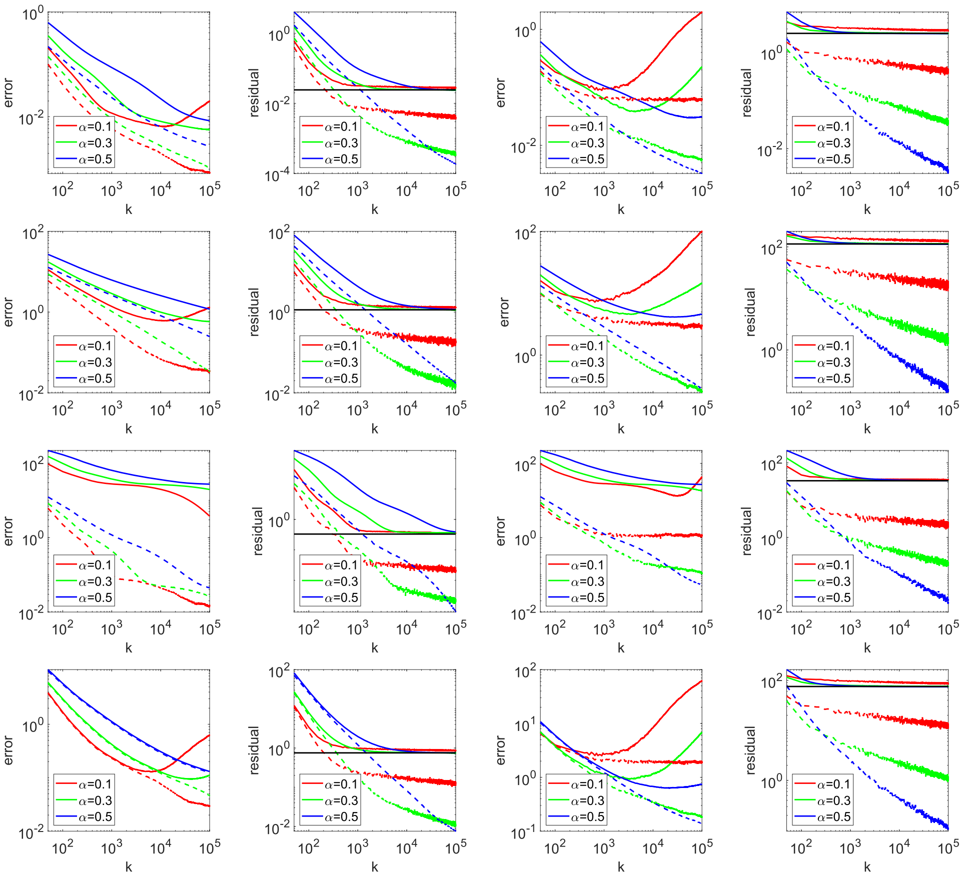

Figure 2. The decay of the mean squared error $\mathbb{E}[\|x_k^\delta - x^\dagger\|^2]$ and residual $\mathbb{E}[\|Ax_k^\delta - y^\delta\|^2]$ and their variance components $\mathbb{E}[\|x_k^\delta - \mathbb{E}[x_k^\delta]\|^2]$ and $\mathbb{E}[\|A(x_k^\delta - \mathbb{E}[x_k^\delta])\|^2]$ versus the SGD iteration number k. The solid and dashed curves denote the mean squared quantity and the variance component, respectively, and the black curve indicates the discrepancy $\delta^2 = \|y^\delta - y^\dagger\|^2$. The first two columns are for the noise level $\delta = 5 \times 10^{-3}$ and the last two columns are for the noise level $\delta = 5 \times 10^{-2}$. The rows from top to bottom refer to `phillips`, `gravity`, `shaw` and `smoothed-phillips`, respectively.

In figure 2, we display the dynamics of mean squared error $\mathbb{E}[\|x_k^\delta - x^\dagger\|^2]$ and the mean squared residual $\mathbb{E}[\|Ax_k^\delta - y^\delta\|^2]$ together with their variance components for the examples at two different relative noise levels, i.e., $\delta = 5 \times 10^{-3}$ and $\delta = 5 \times 10^{-2}$. At each time, SGD is run for 100 epochs (i.e., 1×10^5 SGD iterations), and the results are recorded every 50 SGD iterations, starting from the 50th SGD iterations.

In the plots, we have indicated the true noise $\|y^\delta - y^\dagger\|^2$, also denoted by δ^2. It is observed that both $\mathbb{E}[\|x_k^\delta - x^\dagger\|^2]$ and $\mathbb{E}[\|Ax_k^\delta - y^\delta\|^2]$ decay steadily at an algebraic rate up to a value comparable to the stopping index $k^*(\delta)$ for the Landweber method (by the discrepancy principle (2.2)). Beyond the critical threshold $k^*(\delta)$, the error $\mathbb{E}[\|x_k^\delta - x^\dagger\|^2]$ exhibits a semiconvergence behavior in that it starts to increase, whereas the residual $\mathbb{E}[\|Ax_k^\delta - y^\delta\|^2]$ nearly levels off at a value comparable with the noise level δ^2 (actually it oscillates slightly, since the SGD iterate

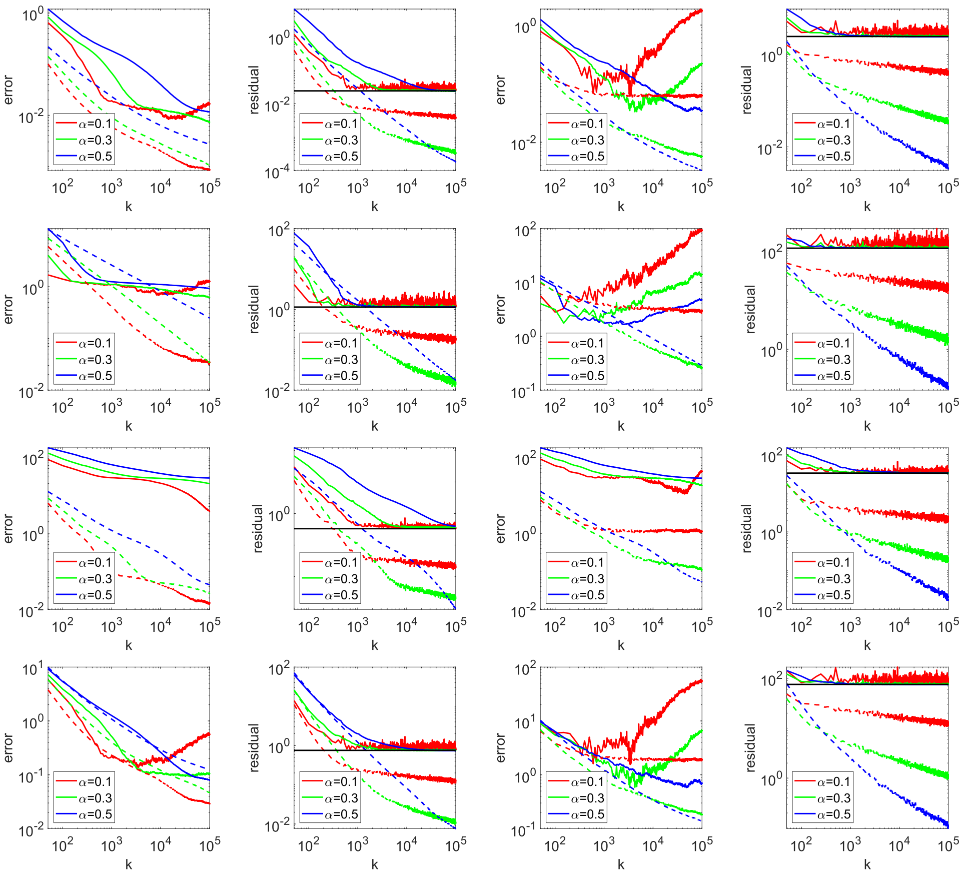

Figure 3. The decay of the squared error $\|x_k^\delta - x^\dagger\|^2$ and residual $\|Ax_k^\delta - y^\delta\|^2$ and their variance components $\mathbb{E}[\|x_k^\delta - \mathbb{E}[x_k^\delta]\|^2]$ and $\mathbb{E}[\|A(x_k^\delta - \mathbb{E}[x_k^\delta])\|^2]$ versus the SGD iteration number k. The solid and dashed curves denote the squared quantity and the variance components, respectively, and the black curve indicates the discrepancy $\delta^2 = \|y^\delta - y^\dagger\|^2$. The first two columns are for the noise level $\delta = 5 \times 10^{-3}$ and the last two columns are for the noise level $\delta = 5 \times 10^{-2}$. The rows from top to bottom refer to `phillips`, `gravity`, `shaw` and `smoothed-phillips`, respectively.

is only descent for the residual on average). This is typical for iterative regularization methods for inverse problems, since for the later iterates, the noise becomes the dominating driving force. Proposition 4.1 with $s = \frac{1}{2}$ indicates that a similar behavior holds also for their variance components (up to slightly beyond $k^*(\delta)$). Actually, the residual variance $\mathbb{E}[\|A(x_k^\delta - \mathbb{E}[x_k^\delta])\|^2]$ first decays as $O(k^{-2(1-\alpha)})$ (upon ignoring the δ term), which matches well the empirical rate in the plot. For the later iterates, as suggested by the δ term in proposition 4.1, the decay is roughly $O(k^{-\alpha})$. Likewise, the error variance $\mathbb{E}[\|x_k^\delta - \mathbb{E}[x_k^\delta]\|^2]$ decays slower at a rate $O(k^{-(1-\alpha)})$. Interestingly, the decay rates of $\mathbb{E}[\|A(x_k^\delta - \mathbb{E}[x_k^\delta])\|^2]$ and $\mathbb{E}[\|x_k^\delta - \mathbb{E}[x_k^\delta]\|^2]$ in the first and last columns are largely comparable, despite their drastic difference in the smoothness of the exact solution $x^\dagger$. Thus, the decay estimate in proposition 4.1 is actually quite sharp, partially explaining the saturation phenomenon observed earlier. This behavior is consistently observed for all three α values. It is worth noting that for `smoothed-phillips`, the curves for $\mathbb{E}[\|x_k^\delta - \mathbb{E}[x_k^\delta]\|^2]$ and $\mathbb{E}[\|x_k^\delta - x^\dagger\|^2]$ nearly overlay each other, i.e., the bias component is negligible after the initial 50 iterations, due to high smoothness of the true solution, clearly indicating the saturation. For the other three examples, empirically, the variance components are of smaller order right after the initial 50 iterations. In particular, as stated in proposition 2.2, $\mathbb{E}[\|A(x_k^\delta - \mathbb{E}[x_k^\delta])\|^2]$ contributes very little to the mean squared residual $\mathbb{E}[\|Ax_k^\delta - y^\delta\|^2]$ in the neighborhood of $k^*(\delta)$. This occurs for all three values of the exponent α in the stepsize schedule. The observations hold also for individual realizations; see figure 3 for the corresponding plots. The overall behavior of the curves in figure 3 is fairly similar to that in figure 2, except that the residual and error curves exhibit pronounced oscillations due to the randomness of the row index selection. Nonetheless, in the neighborhood of $k^*(\delta)$, the variance components remain much smaller in magnitude. This observation provides the key insight for the analysis in section 2.

5.3. Independent run

The convergence analysis in theorem 1.2 requires an SGD iterate $x_{k(\delta)}^\delta$ independent of the stopping index $k(\delta)$ determined by the discrepancy principle (1.3). In practice this can be achieved by an independent run of SGD, at the expense of slightly increasing the computational effort. Now we examine the impact of this choice, and we denote by DP and i-DP the SGD iterate used in (1.3) and that by an independent SGD run, respectively. The relevant numerical results are presented in tables 5–8, where the numbers outside and inside the bracket denote e_{sgd} and std(e_{sgd}), respectively. It is observed that DP gives only slightly better results in terms of the mean, but its standard deviation std(e_{sgd}) is generally much smaller than that by i-DP. Nonetheless, both the mean e$_{\mathrm{sgd}}$ and the standard deviation std(e_{sgd}) of i-DP are decreasing steadily as the noise level δ decreases to 0, confirming the convergence result in theorem 1.2.

The difference is more clearly visualised in the boxplots in figure 4 (for `phillips` with two noise levels). A close look shows that the mean and percentile are fairly close to each other, but the i-DP result tends to have far more outliers lying above the box (marked by red cross in the plots). This is attributed to the fact that $k(\delta)$ determined by the discrepancy principle (1.3) is occasionally too small for an independent SGD run, and thus the corresponding residual is far above the target noise level in the discrepancy principle (1.3); see the boxplots in the last column of figure 4. That is, the outliers are due to stopping too early. This agrees with the observation that one iteration step of SGD has only a small effect on the high frequency components (because of the scaling with the corresponding small singular values). Thus, small $\|Ax_k^\delta - y^\dagger\|$ for $k \ll k^*(\delta)$ implies that also $\|x_k^\delta - x^\dagger\|$ is small. Although not presented, we note that this behavior is observed for all the examples at different noise levels. Thus, in practice, using the SGD iterate directly from the path for (1.3) is preferred, taking into account both accuracy and computational efficiency. It is an interesting theoretical question to analyze the convergence (and convergence rates) of the SGD iterate by (1.3).

6. Concluding remarks

In this work, we have presented a preliminary study on the discrepancy principle as an *a posteriori* stopping rule for the popular stochastic gradient descent for solving linear inverse problems. We proved a finite-iteration termination property of the principle, and a consistency result in high probability for an independent version of discrepancy principle. Several numerical experiments indicate the feasibility of the rule as a stopping criterion.

There are several outstanding questions that deserve further research. First, one important question is the convergence of the dependent version of the discrepancy principle, and convergence rates (and also optimality, if possible!). This would put the discrepancy principle on

Table 5. Comparison between DP and i-DP for `phillips`.

	$\alpha = 0.1$		$\alpha = 0.3$		$\alpha = 0.5$	
δ	DP	i-DP	DP	i-DP	DP	i-DP
1×10^{-3}	8.60×10^{-3} (4.53×10^{-3})	1.12×10^{-2} (1.18×10^{-2})	8.53×10^{-3} (4.42×10^{-3})	1.28×10^{-2} (1.88×10^{-2})	8.34×10^{-3} (4.60×10^{-3})	1.28×10^{-2} (1.55×10^{-2})
5×10^{-3}	1.70×10^{-2} (8.41×10^{-3})	2.31×10^{-2} (2.43×10^{-2})	2.31×10^{-2} (8.81×10^{-3})	3.43×10^{-2} (3.54×10^{-2})	2.48×10^{-2} (7.38×10^{-3})	4.17×10^{-2} (3.63×10^{-2})
1×10^{-2}	2.82×10^{-2} (1.62×10^{-2})	4.35×10^{-2} (4.44×10^{-2})	4.72×10^{-2} (2.07×10^{-2})	6.43×10^{-2} (5.67×10^{-2})	5.78×10^{-2} (2.04×10^{-2})	6.85×10^{-2} (5.66×10^{-2})
5×10^{-2}	1.41×10^{-1} (9.70×10^{-2})	1.53×10^{-1} (8.97×10^{-2})	1.49×10^{-1} (9.01×10^{-2})	1.80×10^{-1} (1.25×10^{-1})	2.11×10^{-1} (9.69×10^{-2})	2.47×10^{-1} (1.93×10^{-1})

Table 6. Comparison between DP and i-DP for `gravity`.

	$\alpha = 0.1$		$\alpha = 0.3$		$\alpha = 0.5$	
δ	DP	i-DP	DP	i-DP	DP	i-DP
1×10^{-3}	6.71×10^{-1} (2.61×10^{-1})	9.30×10^{-1} (7.45×10^{-1})	7.46×10^{-1} (2.73×10^{-1})	1.03×10 (8.04×10^{-1})	7.78×10^{-1} (2.49×10^{-1})	1.00×10 (7.23×10^{-1})
5×10^{-3}	2.00×10 (8.91×10^{-1})	2.43×10 (1.39×10)	2.53×10 (1.12×10)	3.74×10 (2.62×10)	2.76×10 (1.14×10)	3.44×10 (2.36×10)
1×10^{-2}	3.12×10 (1.57×10)	4.03×10 (2.54×10)	4.33×10 (1.92×10)	5.24×10 (3.13×10)	4.74×10 (2.07×10)	6.98×10 (4.17×10)
5×10^{-2}	9.07×10 (5.31×10)	1.01×10^{1} (5.49×10)	1.15×10^{1} (6.61×10)	1.19×10^{1} (8.16×10)	1.52×10^{1} (7.46×10)	1.72×10^{1} (1.10×10^{1})

Table 7. Comparison between DP and i-DP for `shaw`.

	$\alpha = 0.1$		$\alpha = 0.3$		$\alpha = 0.5$	
δ	DP	i-DP	DP	i-DP	DP	i-DP
1×10^{-3}	8.29×10 (9.35×10^{-2})	8.30×10 (3.29×10^{-1})	8.47×10 (5.59×10^{-2})	8.50×10 (2.67×10^{-1})	2.01×10^1 (5.64×10^{-1})	2.00×10^1 (5.25×10^{-1})
5×10^{-3}	2.77×10^1 (1.24×10)	2.77×10^1 (1.27×10)	2.80×10^1 (1.16×10)	2.81×10^1 (1.31×10)	2.82×10^1 (1.02×10)	2.80×10^1 (1.22×10)
1×10^{-2}	2.96×10^1 (1.65×10)	3.03×10^1 (2.58×10)	3.10×10^1 (1.14×10)	3.13×10^1 (2.74×10)	3.12×10^1 (1.08×10)	3.16×10^1 (2.44×10)
5×10^{-2}	5.02×10^1 (1.08×10^1)	5.34×10^1 (1.53×10^1)	6.07×10^1 (8.08×10)	6.19×10^1 (1.23×10^1)	6.70×10^1 (7.41×10)	7.04×10^1 (1.35×10^1)

Table 8. Comparison between DP and i-DP for `smoothed-phillips`.

	$\alpha = 0.1$		$\alpha = 0.3$		$\alpha = 0.5$	
δ	DP	i-DP	DP	i-DP	DP	i-DP
1×10^{-3}	1.63×10^{-1} (6.87×10^{-2})	1.92×10^{-1} (1.27×10^{-1})	1.59×10^{-1} (5.88×10^{-2})	2.00×10^{-1} (1.30×10^{-1})	1.55×10^{-1} (6.09×10^{-2})	1.93×10^{-1} (1.88×10^{-1})
5×10^{-3}	3.92×10^{-1} (2.08×10^{-1})	4.68×10^{-1} (3.47×10^{-1})	5.06×10^{-1} (2.05×10^{-1})	6.54×10^{-1} (4.79×10^{-1})	4.92×10^{-1} (1.99×10^{-1})	7.51×10^{-1} (5.73×10^{-1})
1×10^{-2}	5.95×10^{-1} (2.64×10^{-1})	8.12×10^{-1} (5.04×10^{-1})	8.57×10^{-1} (3.73×10^{-1})	1.22×10 (1.03×10)	9.46×10^{-1} (3.93×10^{-1})	1.46×10 (1.13×10)
5×10^{-2}	2.98×10 (1.44×10)	3.25×10 (1.52×10)	3.20×10 (1.51×10)	3.25×10 (1.94×10)	4.35×10 (2.13×10)	4.59×10 (3.29×10)

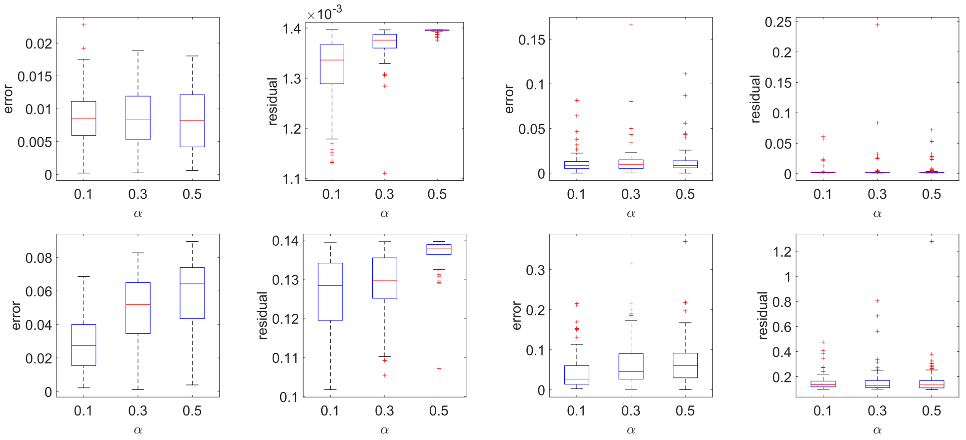

Figure 4. Boxplots for the error $\|x_{k(\delta)}^{\delta} - x^{\dagger}\|^2$ and the residual $\|Ax_{k(\delta)}^{\delta} - y^{\delta}\|^2$ for DP (the first two columns) and i-DP (the last two columns), for `phillips` at two noise levels, i.e., $\delta = 1 \times 10^{-3}$ (top) and $\delta = 1 \times 10^{-2}$ (bottom).

a firm mathematical basis. Second, it is of much interest to study stochastic gradient descent for inverse problems with random noise, with either *a priori* or *a posteriori* stopping rules. In particular, in this context, the discrepancy principle may have to be properly adapted; see the works [1, 8] for interesting discussions with deterministic inversion techniques. Third, the analysis so far does not cover the critical case $\alpha = 1$ in the stepsize schedule. This choice is often adopted in the context of stochastic approximation [15] for optimal asymptotic behavior, but it is unclear whether the discrepancy principle can be applied then.

Acknowledgments

The work of Bangti Jin is supported by UK EPSRC grant EP/T000864/1.

ORCID iDs

Bangti Jin https://orcid.org/0000-0002-3775-9155

References

[1] Blanchard G and Mathé P 2012 Discrepancy principle for statistical inverse problems with application to conjugate gradient iteration *Inverse Problems* **28** 115011
[2] Bottou L 2010 Large-scale machine learning with stochastic gradient descent *Proc. of COMPSTAT'2010* (Berlin: Springer) pp 177–86
[3] Bottou L, Curtis F E and Nocedal J 2018 Optimization methods for large-scale machine learning *SIAM Rev.* **60** 223–311
[4] Engl H W, Hanke M and Neubauer A 1996 *Regularization of Inverse Problems* (Dordrecht: Kluwer)
[5] Feller W 1968 *An Introduction to Probability Theory and its Applications* 3rd edn Vol I (New York: Wiley)
[6] Gordon R, Bender R and Herman G T 1970 Algebraic reconstruction techniques (art) for three-dimensional electron microscopy and x-ray photography *J. Theor. Biol.* **29** 471–81
[7] Hansen P C 2007 Regularization tools version 4.0 for Matlab 7.3 *Numer. Algorithms* **46** 189–94

[8] Harrach B, Jahn T and Potthast R 2020 Beyond the Bakushinskii veto: regularising linear inverse problems without knowing the noise distribution *Numer. Math.* **145** 581–603

[9] Ito K and Jin B 2015 *Inverse Problems: Tikhonov Theory and Algorithms* (Singapore: World Scientific)

[10] Jin B and Lu X 2019 On the regularizing property of stochastic gradient descent *Inverse Problems* **35** 015004

[11] Jin B, Zhou Z and Zou J 2020 On the convergence of stochastic gradient descent for nonlinear ill-posed problems *SIAM J. Optim.* **30** 1421–50

[12] Johnson R and Zhang T 2013 Accelerating stochastic gradient descent using predictive variance reduction *NIPS'13* ed C J C Burges, L Bottou, M Welling, Z Ghahramani and K Q Weinberger (Lake Tahoe, Nevada)pp 315–23

[13] Kaltenbacher B, Neubauer A and Scherzer O 2008 *Iterative Regularization Methods for Nonlinear Ill-Posed Problems* (Berlin: Walter de Gruyter GmbH)

[14] Kluth T and Jin B 2019 Enhanced reconstruction in magnetic particle imaging by whitening and randomized SVD approximation *Phys. Med. Biol.* **64** 125026

[15] Kushner H J and Yin G G 2003 *Stochastic Approximation and Recursive Algorithms and Applications* 2nd edn (New York: Springer)

[16] Landweber L 1951 An iteration formula for Fredholm integral equations of the first kind *Am. J. Math.* **73** 615–24

[17] Morozov V A 1966 On the solution of functional equations by the method of regularization *Sov. Math. Dokl.* **7** 414–7

[18] Natterer F 1986 *The Mathematics of Computerized Tomography* ed Teubner B G (New York: Wiley)

[19] Robbins H and Monro S 1951 A stochastic approximation method *Ann. Math. Stat.* **22** 400–7