On projective stochastic-gradient type methods for solving large scale systems of nonlinear ill-posed equations: Applications to machine learning

J.C. Rabelo[†] Y. Saporito[‡] A. Leitão [‡§]

August 26, 2023

Abstract

We propose and analyze a stochastic-gradient type method for solving systems of nonlinear ill-posed equations. The method considered here extends the SGD type iteration introduced in [2022, Rabelo et al., Inv. Probl. **38** 025003] for solving linear ill-posed systems.

A distinctive feature of our method resides in the *a posteriori* choice of the stepsize, which promotes a relaxed orthogonal projection of the current iterate onto a conveniently chosen convex set. This characteristic distinguish our method from other SGD type methods in the literature (where the stepsize is typically chosen *a priori*) and accounts for the faster convergence observed in the numerical experiments conducted in this manuscript.

The convergence analysis discussed here includes: monotonicity and mean square convergence of the iteration error (exact data case), stability and semi-convergence (noisy data case). In the later case, our method is coupled with an *a priori* stopping rule.

Numerical experiments are presented for two large scale nonlinear inverse problems in machine learning (both with real data): (i) we address, using neural networks, the big data problem of CO-concentration prediction considered in the above cited article; (ii) we tackle the classification problem for the MNIST database (http://yann.lecun.com/exdb/mnist/). The obtained numerical results demonstrate the efficiency of the proposed method.

Keywords. Ill-posed problems; Nonlinear equations; SGD method; Landweber method; Projective method.

AMS Classification: 65J20, 47J06.

1 Introduction

In these notes we extend to nonlinear ill-posed problems the projective stochastic-gradient type method proposed in [25, 24] for solving large scale systems of linear equations.

Problems under consideration

The *inverse problem* under consideration consist of determining an unknown quantity $x \in X$ from a set of data $(y_0, \ldots, y_{N-1}) = (y_i) \in Y^N$, where X and Y are (infinite dimensional) Hilbert spaces. The data (y_i) correspond to indirect observations of the parameter x, this process being

[†]Department of Mathematics, Federal Univ. of Piaui, 64049-550 Teresina, Brazil
[‡]EMAp, Getulio Vargas Fundation, Praia de Botafogo 190, 22250-900 Rio de Janeiro, Brazil
[§]On leave from Department of Mathematics, Federal Univ. of St. Catarina, 88040-900 Floripa, Brazil
Emails: joelrabelo@ufpi.edu.br, yuri.saporito@gmail.com, acgleitao@gmail.com.

described by the model $y_i = F_i(x)$, for $i = 0, \dots, N-1$; where $F_i : D(F_i) \subset X \to Y$ are ill-posed nonlinear operators [8, 10].

We are particularly interested in the situation where $N >> 1$ is large, and the exact data (y_i) are not available; instead, only approximate data (y_i^δ) satisfying

$$\|y_i^\delta - y_i\|_Y \;\leq\; \delta_i\,, \quad i = 0, \dots, N-1\,, \tag{1}$$

are available. Here $\delta_i > 0$ are known noise levels; we write $\delta := (\delta_0, \dots, \delta_{N-1}) \in \mathbb{R}^N$.

The abstract formulation of the problems under consideration can be summarized as follows: Given inexact data (y_i^δ) and the levels of noise (δ_i) as in (1), find an approximate solution to the large scale system of nonlinear operator equations

$$F_i(x) \;=\; y_i^\delta\,, \quad i = 0, \dots, N-1\,. \tag{2}$$

A straightforward approach for solving the inverse problem (1), (2) consists in rewriting (2) as a single operator equation $\mathbf{F}(x) = \mathbf{y}^\delta$, with $\mathbf{F} := (F_0, \dots, F_{N-1}) : X \to Y^N$ and $\mathbf{y}^\delta := \big(y_0^\delta, \dots, y_{N-1}^\delta\big)$, and using standard regularization methods; e.g., *Iterative regularization* [1, 8, 14, 18, 19] or *Tikhonov regularization* [8, 23, 27, 28, 29, 26]. When using this functional analytical formulation, dealing with the numerical challenges of solving a high-dimensional ill-posed equation becomes inevitable. Specifically, when applied to $\mathbf{F}(x) = \mathbf{y}^\delta$, the above mentioned regularization methods often become numerically inefficient when $N >> 1$.

In what follows we briefly discuss alternative approaches for solving the nonlinear inverse problem (1), (2) in a stable manner:

- **Kaczmarz type methods** (cyclic iterations): this technique is considered in [13, 11], [7], [12], [2], [22] and [3] for the Landweber iteration, the Steepest-Descent iteration, the Expectation-Maximization iteration, the Levenberg-Marquardt iteration, the REGINN-Landweber iteration, and the Iteratively Regularized Gauss-Newton iteration respectively;
- **SGD type methods** (non-cyclic iterations): this stochastic technique is considered in [17] with *a priori* chosen stepsize and *a priori* stopping rule (see [16] for the linear case); in [15] with *a priori* chosen stepsize and *a posteriori* stopping rule; in **this manuscript** *a posteriori* chosen stepsize and *a priori* stopping rule (see [25] for the linear case).

The rationale behind our method

The stochastic-gradient (SGD) type method considered in this manuscript aims to compute, in a stable way, approximate solutions to (1), (2). Our method stands out due to the stepsize selection, which is inspired by the *projective Landweber (PLW) method* [21] and the *projective Landweber-Kaczmarz (PLWK) method* [20]. In the sequel, we briefly address these two methods:

- **The PLW method** was proposed in [21] for solving (1), (2) with $N = 1$, i.e. $F_0(x) = y_0^\delta$ with $\|y_0 - y_0^\delta\| \leq \delta$. A sequence (x_k^δ) is generated as follows: at each iteration k, a half space

$$H_{x_k^\delta} := \big\{ z \in X,\, \big\langle z - x_k^\delta, F_0'(x_k^\delta)^* F_\delta(x_k^\delta) \big\rangle \,\leq\, -\|F_\delta(x_k^\delta)\| \big((1-\eta)\|F_\delta(x_k^\delta)\| - (1+\eta)\delta \big) \big\}$$

is defined, where $F_\delta(x) := F_0(x) - y_0^\delta$ (see (A2) in Section 2.1 for the definition of the constant η). Under appropriate assumptions, it is proven that $H_{x_k^\delta}$ contains all solutions of $F_0(x) = y_0$; moreover, if the norm of the residual $\|F_\delta(x_k^\delta)\|$ is above the trashold $(1+\eta)(1-\eta)^{-1}\delta$ then $H_{x_k^\delta}$ does not contain the iterate x_k^δ.[1] The next iterate x_{k+1}^δ is defined as a (relaxed) orthogonal projection of x_k^δ onto $H_{x_k^\delta}$ (see [21, Eq. (8)] for details). Summarizing, PLW corresponds to a Landweber type iteration [19, 8] with stepsize defined by (relaxed) orthogonal projections onto

[1] In this situation we say that **the set $\mathbf{H_{x_k^\delta}}$ separates the iterate $\mathbf{x_k^\delta}$ from the solution set $\mathbf{F_0^{-1}(y_0)}$**; in other words $F_0^{-1}(y_0) \subset H_{x_k^\delta}$, while $x_k^\delta \notin H_{x_k^\delta}$.

the separating sets $H_{x_k^\delta}$.

• **The PLWK method** was proposed in [20] for solving systems of nonlinear ill-posed equations as in (1), (2) with $N > 1$. It consists in coupling the PLW method with the Kaczmarz strategy and incorporating a bang-bang parameter. The corresponding iteration formula reads

$$x_{k+1}^\delta \; := \; x_k^\delta - \theta_k \, \lambda_k \, \omega_k \, F_{[k]}^*(x_k^\delta)\big(F_{[k]}(x_k^\delta) - y_{[k]}^\delta\big),$$

where $[k] := (k \mod N) \in \{0, \ldots, N-1\}$, $\theta_k \in (0, 2)$ is a relaxation parameter and $\omega_k \in \{0, 1\}$ is a bang-bang parameter (see [20, Eq. (6)]). Moreover, $\lambda_k \geq 0$ (see [20, Eq. (12)]) gives the exact orthogonal projection of x_k^δ onto the half space $H_{[k], x_k^\delta}$, where

$$H_{i, x_k^\delta} \; := \; \Big\{ z \in X, \, \langle z - x_k^\delta, F_i'(x_k^\delta)^* F_{i,\delta}(x_k^\delta) \rangle \; \leq \; -\|F_{i,\delta}(x_k^\delta)\| \Big((1-\eta)\|F_{i,\delta}(x_k^\delta)\| - (1+\eta)\delta \Big) \Big\},$$

for $i = 0, \ldots, N-1$; here $F_{i,\delta}(x) := F_i(x) - y_i^\delta$ (see [20, Eq. (11)]). Summarizing, PLWK corresponds to a Landweber-Kaczmarz (cyclic) type iteration [13] with stepsize defined by (relaxed) orthogonal projections onto the separating sets $H_{[k], x_k^\delta}$.

In [25, 24] the projective step of the PLW and PLWK iterations was used as starting point to derive a SGD type method for solving large scale systems of linear ill-posed equations.

The projective stochastic-gradient (pSGD) method

In this manuscript we build upon a well-established nonlinear assumption, namely the *weak tangential cone condition* (wTCC) [14, 8], to expand the method in [25, 24]. As a result, we create a new approach capable of efficiently solving large-scale systems of nonlinear equations of the form (1), (2). For obvious reasons, the method considered in these notes is named *projective stochastic-gradient* (pSGD) method.

Unlike the majority of SGD type methods found in the literature, our approach employs *a posteriori* stepsize selection (see (6)). Additionally, in the noisy data case, our iterative method is combined with an *a priori* stopping rule (see (A6)), classifying pSGD as a regularization method as defined in [8].

Outline of the manuscript

In Section 2 we introduce the pSGD method; the main assumptions used in our analysis are presented. Section 3 is dedicated to convergence analysis of pSGD. In Section 3.1 the exact data case is considered: We estimate the *average gain* (Proposition 3.2), and prove monotonicity of the *average iteration error* (Corollary 3.3) as well as square summability of the *average residuals* (Corollary 3.4). Additionally, a convergence result is proven (Theorem 3.5). Section 3.2 is devoted to analyzing the noisy data case and regularization properties of pSGD. The key findings include a stability result (Theorem 3.9) and a semi-convergence result (Theorem 3.11).

In Section 4 numerical experiments are presented for solving two large scale systems of nonlinear equations with real data. Both inverse problems relate to parameter identification in neural network training, in detail:

• In Section 4.1 we tackle the big data problem of CO-concentration prediction in a gas sensor array [9, 25]. A special neural-network (NN) is used to model the related inverse problem (a variation of the saturated linear activation function is used). We prove in Lemma 4.2 that the nonlinear function modeling this NN satisfies the wTCC.

• In Section 4.2 we address the well-known classification problem for the MNIST database, consisting of images of handwritten digits (see https://en.wikipedia.org/wiki/MNIST_database). Section 5 is devoted to final remarks and conclusions.

2 The method under investigation

This section presents the nonlinear pSGD method under consideration in this notes. We begin by addressing the main assumptions necessary for the analysis derived in the forthcoming sections.

2.1 Main assumptions

Throughout this work we assume that $\bigcap_i D(F_i)$ has nonempty interior, where $D(F_i) \subset X$ is the domain of definition of F_i. Additionally, the initial guess $x_0 \in X$ satisfies $B_\rho(x_0) \subset \bigcap_{i=0}^{N-1} D(F_i)$ for some $\rho > 0$. Moreover, the following assumptions are used:

(A1) Each operator F_i is Fréchet differentiable with continuous derivative F_i'. Moreover, there exists a constant $C > 0$ such that

$$\|F_i'(x)\| \leq C, \quad i = 0, \ldots, N-1, \quad \forall\, x \in B_\rho(x_0); \tag{3}$$

(A2) The *weak Tangential Cone Condition* (wTCC) holds at $B_\rho(x_0)$, with $0 < \eta < 1$, i.e.

$$\|F_i(\bar{x}) - F_i(x) - F_i'(x)(\bar{x} - x)\|_Y \leq \eta \, \|F_i(\bar{x}) - F_i(x)\|_Y, \tag{4}$$

for $i = 0, \ldots, N-1$ and $\forall\, x, \bar{x} \in B_\rho(x_0)$;

(A3) There exists $x^\star \in B_{\rho/2}(x_0)$ such that $F_i(x^\star) = y_i$ for $i = 0, \ldots, N-1$, i.e. $x^\star$ is a (non necessarily unique) solution of (2) with exact data;

(A4) $(\theta_k) \in \mathbb{R}^+$ is a sequence satisfying $0 < \inf_k \theta_k$ and $\sup_k \theta_k < 2$;

(A5) $\gamma > 0$ is a constant satisfying $\gamma > \dfrac{1+\eta}{1-\eta} C$, whith C as in (A1) and η as in (A2);

(A6) The stopping index $k_\delta^* = k^*(\delta) \in \mathbb{N}$, satisfies $\lim\limits_{j\to\infty} k_{\delta^j}^* = \infty$ and $\lim\limits_{j\to\infty} \|\delta^j\|^2 k_{\delta^j}^* = 0$.

The following inequalities are immediate consequences of (A2):

$$(1-\eta)\|F_i(\bar{x}) - F_i(x)\| \leq \|F_i'(x)(\bar{x} - x)\| \leq (1+\eta)\|F_i(\bar{x}) - F_i(x)\|, \tag{5}$$

for $i = 0, \ldots, N-1$ and $x, \bar{x} \in B_\rho(x_0)$ (see [8, Chapter 11] for further discussion).

2.2 Introducing the nonlinear pSGD iteration

Here we present the nonlinear pSGD method for solving (1), (2). In what follows we adopt the simplified notation: $F_i^\delta(x) := F_i(x) - y_i^\delta$, for $i = 0, \ldots, N-1$; and define the polynomial function $p^\delta(t) := t[(1-\eta)t - (1+\eta)\delta]$.

Given x_0, $\gamma > 0$ and (θ_k) as in Section 2.1, the iteration formula of the pSGD method reads

$$x_{k+1}^\delta = x_k^\delta - \theta_k \, \lambda_{I_k}^\delta \, F_{I_k}'(x_k^\delta)^* \, F_{I_k}^\delta(x_k^\delta), \quad k = 0, 1, \ldots, k_\delta^* - 1. \tag{6a}$$

Here the stepsize $\lambda_{I_k}^\delta \geq 0$ is a function of $(x_k^\delta, y_{I_k}^\delta, \delta_{I_k})$ and is defined by

$$\lambda_{I_k}^\delta := \begin{cases} p^{\delta_{I_k}}(\|F_{I_k}^\delta(x_k^\delta)\|) / \|F_{I_k}'(x_k^\delta)^* \, F_{I_k}^\delta(x_k^\delta)\|^2 & , \text{ if } \|F_{I_k}'(x_k^\delta)^* \, F_{I_k}^\delta(x_k^\delta)\| > \gamma \delta_{I_k} \\ 0 & , \text{ otherwise.} \end{cases} \tag{6b}$$

The (I_k) is an independent and identically distributed sequence of random indexes, taking values in $\{0, \ldots, N-1\}$, in a fixed probability space $(\Omega, \mathcal{F}, \mathbb{P})$. For simplicity of the presentation we assume that $\mathbb{P}(I_k = i) = \frac{1}{N}$ for $i = 0, \ldots, N-1$.

Remark 2.1. *In the exact data case, we write $F_i^0(x) := F_i(x) - y_i$ and $p^0(t) := (1-\eta)t^2$. The iteration formula of the pSGD method reads*

$$x_{k+1} \;=\; x_k \;-\; \theta_k\, \lambda_{I_k}\, F'_{I_k}(x_k)^*\, F^0_{I_k}(x_k)\,,\;\; k = 0,1,\dots \tag{7a}$$

where the stepsize $\lambda_{I_k} \geq 0$ is a function of (x_k, y_{I_k}) and is defined by

$$\lambda_{I_k} \;:=\; \begin{cases} p^0(\|F^0_{I_k}(x_k)\|)/\|F'_{I_k}(x_k)^*\, F^0_{I_k}(x_k)\|^2 & ,\ \text{if}\ \|F'_{I_k}(x_k)^*\, F^0_{I_k}(x_k)\| > 0 \\ 0 & ,\ \text{otherwise.} \end{cases} \tag{7b}$$

Remark 2.2 (Exact projections)**.** *If one takes $\theta_k \equiv 1$ in (6a), then x^δ_{k+1} corresponds to the orthogonal projection of x^δ_k onto $H^\delta_{I_k, x^\delta_k}$, where*

$$H^\delta_{i,x} \;:=\; \big\{ z \in X \mid \langle z - x,\; F'_i(x)^* F^\delta_i(x) \rangle \;\geq\; \|F^\delta_i(x)\| \, \big((1-\eta)\|F^\delta_i(x)\| - (1+\eta)\delta_i \big) \big\}.$$

Alternatively, if $\theta_k \in (0,2)$ x^δ_{k+1} can be interpreted as a relaxed projection of x^δ_k onto H_{I_k, x^δ_k}.

Remark 2.3 (Lower bound for the stepsizes)**.**

- *In the exact data case, it holds $\lambda_{I_k} \geq (1-\eta)C^{-2}$ whenever $\|F'_{I_k}(x_k)^*\, F^0_{I_k}(x_k)\| > 0$. I.e. λ_{I_k} in (7b) is bounded by below, whenever x_k is not a solution of $F_{I_k}(x) = y_{I_k}$.*
- *In the noisy data case, Assumption (A5) implies $\lambda^\delta_{I_k} \geq C^{-2}\,(1 - \eta - C(1+\eta)/\gamma) =: \lambda_{\min}$, whenever $\|F'_{I_k}(x^\delta_k)^*\, F^\delta_{I_k}(x^\delta_k)\| > \gamma \delta_{I_k}$.*

3 Convergence analysis

In this section, analytical properties of the nonlinear pSGD method in (6) are investigated. We start the discussion by considering the case of exact data, i.e. $\delta_i = 0$ for $i = 0, \dots, N-1$.

3.1 The exact data case

In this case, the inverse problem (1), (2) can be written in the form

$$F_i(x) \;=\; y_i\,, \quad i = 0, \dots, N-1\,, \tag{8}$$

or simply $\mathbf{F}(x) = \mathbf{y}$. Next we introduce relevant notation used in this manuscript.

Remark 3.1 (Notation)**.**

- *For $x^* \in X$ a solution of (8), the mean square iteration error $\mathbb{E}\big[\|x^* - x_k\|^2\big]$ is defined by the average error over all possible realizations of $I_0, \dots, I_{k-1}$ that define x_k. E.g., for $k = 0$ and $k = 1$, it holds*

  $$\mathbb{E}\big[\|x^* - x_0\|^2\big] = \|x^* - x_0\|^2, \quad \mathbb{E}\big[\|x^* - x_1\|^2\big] = \tfrac{1}{N} \sum_{i=0}^{N-1} \|x^* - [x_0 - \theta_1 \lambda_i F'_i(x_0)^* F^0_i(x_0)]\|^2.$$

- *Let $k \in \mathbb{N}$ be fixed, and denote by $\mathcal{F}_k$ the σ-algebra generated by $I_0, \dots, I_{k-1}$. It holds*

  $$\mathbb{E}\big[\lambda_I \|F^0_I(x_k)\|^2 | \mathcal{F}_k\big] = \tfrac{1}{N} \sum_{i=0}^{N-1} \lambda_i \, \|F^0_i(x_k)\|^2, \quad \mathbb{E}\big[\|x^* - x_k\|^2 | \mathcal{F}_k\big] = \|x^* - x_k\|^2,$$

  $$\mathbb{E}\big[\|x^* - x_{k+1}\|^2 - \|x^* - x_k\|^2 | \mathcal{F}_k\big] = \tfrac{1}{N} \sum_{i=0}^{N-1} \big[\|x^* - [x_k - \theta_k \lambda_i F'_i(x_k)^* F^0_i(x_k)]\|^2 - \|x^* - x_k\|^2 \big].$$

  Moreover, by the law of iterated expectation, we have

  $$\mathbb{E}\big[\lambda_I \|F^0_I(x_k)\|^2\big] \;=\; \mathbb{E}\big[\mathbb{E}\big[\lambda_I \|F^0_I(x_k)\|^2 | \mathcal{F}_k\big]\big] \;=\; \tfrac{1}{N} \sum_{i=0}^{N-1} \mathbb{E}\big[\lambda_i \|F^0_i(x_k)\|^2\big],$$

where the last expectation averages the residual of equation i times λ_i over all possible realizations of x_k. It is worth noticing that λ_i is a random variable (indeed, λ_i depends on the realization of x_k; see (7b)).

In the next proposition we estimate the difference $\mathbb{E}[\|x^* - x_{k+1}\|^2] - \mathbb{E}[\|x^* - x_k\|^2]$, where $x^* \in X$ is a solution of (8). This is a quintessential result in the forthcoming analysis.

Proposition 3.2. *Let assumptions (A1), (A2) and (A3) hold and (x_k) be a sequence generated by the nonlinear pSGD method (7). If $x_k \in B_\rho(x_0)$, then for any x^* solution of (8) it holds*

$$\mathbb{E}\big[\|x^* - x_{k+1}\|^2\big] - \mathbb{E}\big[\|x^* - x_k\|^2\big] \;\le\; \theta_k(\theta_k - 2)(1-\eta)\,\mathbb{E}\big[\lambda_I \|F_I^0(x_k)\|^2\big]. \tag{9}$$

Proof. If $F'_{I_k}(x_k)^* F^0_{I_k}(x_k) \neq 0$, we obtain from (7a) and (7b)

$$\begin{aligned}
\|x^* \,-\, x_{k+1}\|^2 - \|x^* - x_k\|^2 \;&=\; 2\langle x_{k+1} - x_k, x_{k+1} - x^*\rangle - \|x_k - x_{k+1}\|^2 \\
&= \; -2\theta_k\lambda_{I_k}\langle F'_{I_k}(x_k)^* F^0_{I_k}(x_k), x_{k+1} - x^*\rangle - \theta_k^2\lambda_{I_k}^2\|F'_{I_k}(x_k)^* F^0_{I_k}(x_k)\|^2 \\
&= \; -2\theta_k\lambda_{I_k}\langle F^0_{I_k}(x_k), -F^0_{I_k}(x_k) - F'_{I_k}(x_k)(x^* - x_k) + F'_{I_k}(x_k)(x_{k+1} - x_k) + F^0_{I_k}(x_k)\rangle \\
&\quad\; -\theta_k^2\lambda_{I_k}^2\|F'_{I_k}(x_k)^* F^0_{I_k}(x_k)\|^2 \\
&= \; 2\theta_k\lambda_{I_k}\big[\langle -F^0_{I_k}(x_k), -F^0_{I_k}(x_k) - F'_{I_k}(x_k)(x^* - x_k)\rangle + \theta_k p^0(\|F^0_{I_k}(x_k)\|) - \|F^0_{I_k}(x_k)\|^2\big] \\
&\quad\; - \theta_k^2\lambda_{I_k} p^0(\|F^0_{I_k}(x_k)\|)\,.
\end{aligned}$$

Since $-F^0_{I_k}(x_k) = F_{I_k}(x^*) - F_{I_k}(x_k)$ and $x_k \in B_\rho(x_0)$, we argue with (A2) to obtain

$$\begin{aligned}
&\|x^* \,-\, x_{k+1}\|^2 - \|x^* - x_k\|^2 \\
&\quad\le\; 2\theta_k\lambda_{I_k}\Big[(\eta - 1)\|F^0_{I_k}(x_k)\|^2 + \theta_k p^0(\|F^0_{I_k}(x_k)\|)\Big] - \theta_k^2\lambda_{I_k} p^0(\|F^0_{I_k}(x_k)\|) \\
&\quad=\; 2\theta_k\lambda_{I_k}\Big[(\eta - 1)\|F^0_{I_k}(x_k)\|^2 + \frac{1}{2}\theta_k(1-\eta)\|F^0_{I_k}(x_k)\|^2\Big] \\
&\quad=\; \theta_k(\theta_k - 2)(1-\eta)\lambda_{I_k}\|F^0_{I_k}(x_k)\|^2.
\end{aligned} \tag{10}$$

Otherwise, if $F'_{I_k}(x_k)^* F^0_{I_k}(x_k) = 0$ then $x_{k+1} = x_k$ and $F_{I_k}(x_k) = y_{I_k}$. Consequently, (10) holds also in this case.

Denoting by $\mathcal{F}_k$ the σ-algebra generated by $(I_0, \ldots, I_{k-1})$, we conclude that x_k is measurable with respect to $\mathcal{F}_k$ (while I_k is independent of it). Consequently, we derive from (10) the estimate

$$\mathbb{E}\big[\|x^* - x_{k+1}\|^2 - \|x^* - x_k\|^2 | \mathcal{F}_k\big] \;\le\; \theta_k(\theta_k - 2)(1-\eta)\,\mathbb{E}\big[\lambda_I\|F_I^0(x_k)\|^2 | \mathcal{F}_k\big]\,.$$

Taking the full conditional in the last inequality yields (9). □

If, additionally to (A1), ..., (A3), assumption (A4) holds, then (10) implies that any sequence (x_k) generated by the pSGD method (7) satisfies $x_k \in B_{\rho/2}(x^\star) \subset B_\rho(x_0)$, for $k = 0, 1, \ldots$ Thus, under this additional assumption, Proposition 3.2 implies the monotonicity of the mean square iteration error $\mathbb{E}[\|x^* - x_k\|^2]$, where $x^* \in X$ is a solution of (8). This fact is summarized in

Corollary 3.3. *Let assumptions (A1), ..., (A4) hold true and (x_k) be a sequence generated by the nonlinear pSGD method (7). For any solution x^* of (8) it holds*

$$\mathbb{E}[\|x^* - x_{k+1}\|^2] \;\le\; \mathbb{E}[\|x^* - x_k\|^2]\,,\; k = 0, 1, \ldots \tag{11}$$

An important consequence of Proposition 3.2 is discussed in the sequel (this result is used in the proof of the main convergence Theorem 3.5).

Corollary 3.4. *Let assumptions (A1),...,(A4) hold true and (x_k) be a sequence generated by the nonlinear pSGD method (7). The series*

$$\sum_{k=0}^{\infty} \theta_k(2-\theta_k)(1-\eta)\mathbb{E}[\lambda_I \|F_I^0(x_k)\|^2] \ , \qquad \sum_{k=0}^{\infty} \theta_k \, \mathbb{E}[\lambda_I \|F_I^0(x_k)\|^2] \quad \text{and} \quad \sum_{k=0}^{\infty} \mathbb{E}[\|F_I^0(x_k)\|^2]$$

are summable.

Proof. The summability of the first series follows from Proposition 3.2. The summability of the second and third series follows from (A4) together with the summability of the first series. □

We are now ready to state and prove a convergence result for the pSGD method in (7).

Theorem 3.5 (Convergence for exact data)**.** *Let assumptions (A1),...,(A4) hold. Any sequence (x_k) generated by the nonlinear pSGD method (7) converges in mean square to some element $x^* \in B_\rho(x_0)$, i.e. $\mathbb{E}[\|x_k - x^*\|^2] \to 0$ as $k \to \infty$, which is a solution of (8).*

Proof. We claim that (x_k) is a Cauchy sequence. It is enough to prove that $e_k := x^\star - x_k$ is Cauchy, where $x^\star$ is defined as in (A3). From Corollary 3.3 follows

$$\lim_{k\to\infty} \mathbb{E}[\|e_k\|^2] \ = \ \varepsilon \geq 0 \, , \tag{12}$$

Next we prove that

$$\mathbb{E}[\langle e_n - e_k, e_n\rangle] \to 0 \quad \text{and} \quad \mathbb{E}[\langle e_l - e_n, e_n\rangle] \to 0 \quad \text{as} \quad k,\ l \to \infty, \tag{13}$$

with $k \leq l$ for some $k \leq n \leq l$ (compare with [14, Theorem 2.3]). Notice that $\mathbb{E}[\langle \cdot\, , \cdot \rangle_X]$, $\mathbb{E}[\langle \cdot\, , \cdot \rangle_Y])$ define inner products in $L_2(\Omega; X)$ and $L_2(\Omega; Y)$ respectively.[2]

Notice that, for any fixed $k \leq l$, one can always choose an index n with $k \leq n \leq l$ such that

$$\mathbb{E}[\lambda_I \|F_I^0(x_n)\|^2] \ \leq \ \mathbb{E}[\lambda_I \|F_I^0(x_j)\|^2] \, , \quad \forall\, k \leq j \leq l \, . \tag{14}$$

Next, arguing with (7a) and the Cauchy–Schwartz inequality (for random variables) we estimate

$$\begin{aligned}
\big|\mathbb{E}[\langle e_n - e_k, e_n\rangle]\big| \ &= \ \Big|\sum_{j=k}^{n-1} \mathbb{E}[\langle x_{j+1} - x_j, x^\star - x_n\rangle]\Big| \ = \ \Big|\sum_{j=k}^{n-1} \mathbb{E}[\theta_j \, \lambda_I \langle F_I'(x_j)^* F_I^0(x_j), x_n - x^\star\rangle]\Big| \\
&= \ \Big|\sum_{j=k}^{n-1} \theta_j \, \mathbb{E}[\lambda_I \langle F_I^0(x_j), F_I'(x_j)(x_n - x_j + x_j - x^\star)\rangle]\Big| \\
&= \ \Big|\sum_{j=k}^{n-1} \theta_j \, \mathbb{E}[\langle \lambda_I^{\frac12} F_I^0(x_j), \lambda_I^{\frac12} F_I'(x_j)(x_n - x_j)\rangle] + \theta_j \, \mathbb{E}[\langle \lambda_I^{\frac12} F_I^0(x_j), \lambda_I^{\frac12} F_I'(x_j)(x_j - x^\star)\rangle]\Big| \\
&\leq \ \sum_{j=k}^{n-1} \Big(\theta_j \, \mathbb{E}[\lambda_I \|F_I^0(x_j)\|^2]^{\frac12} \ \mathbb{E}[\lambda_I \|F_I'(x_j)(x_n - x_j)\|^2]^{\frac12} \\
&\qquad\qquad + \theta_j \, \mathbb{E}[\lambda_I \|F_I^0(x_j)\|^2]^{\frac12} \ \mathbb{E}[\lambda_I \|F_I'(x_j)(x_j - x^\star)\|^2]^{\frac12}\rangle]\Big) \, .
\end{aligned}$$

Thus, it follows from (5) that

$$\begin{aligned}
\big|\mathbb{E}[\langle e_n - e_k, e_n\rangle]\big| \ &\leq \ (1+\eta) \sum_{j=k}^{n-1} \theta_j \, \mathbb{E}[\lambda_I \|F_I^0(x_j)\|^2]^{\frac12} \, \mathbb{E}[\lambda_I \|F_I^0(x_n) - F_I^0(x_j)\|^2]^{\frac12} \\
&\qquad + \ (1+\eta) \sum_{j=k}^{n-1} \theta_j \, \mathbb{E}[\lambda_I \|F_I^0(x_j)\|^2] \, .
\end{aligned}$$

[2] $L_2(\Omega; X)$ is the space of square integrable random variables defined on Ω and taking values in X.

The term $\mathbb{E}[\lambda_I \|F_I^0(x_n) - F_I^0(x_j)\|^2]^{\frac{1}{2}}$ on the right hand side of the last estimate can be estimated using (14). Indeed, for each $j = k, \ldots, n-1$ it holds

$$\begin{aligned}\mathbb{E}[\lambda_I \|F_I^0(x_n) - F_I^0(x_j)\|^2]^{\frac{1}{2}} &\leq \Big(2\mathbb{E}[\lambda_I \|F_I^0(x_n)\|^2] + 2\mathbb{E}[\lambda_I \|F_I^0(x_j)\|^2]\Big)^{\frac{1}{2}} \\ &\leq \sqrt{2}\Big(\mathbb{E}[\lambda_I \|F_I^0(x_j)\|^2] + \mathbb{E}[\lambda_I \|F_I^0(x_j)\|^2]\Big)^{\frac{1}{2}} \leq 2\Big[\mathbb{E}[\lambda_I \|F_I^0(x_j)\|^2]\Big]^{\frac{1}{2}}.\end{aligned}$$

Consequently,

$$\big|\mathbb{E}[\langle e_n - e_k, e_n\rangle]\big| \leq 3(1+\eta)\sum_{j=k}^{n-1} \theta_j\, \mathbb{E}[\lambda_I \|F_I^0(x_j)\|^2].$$

Now, Corollary 3.4 allow us to conclude that $\mathbb{E}[\langle e_n - e_k, e_n\rangle] \to 0$ as $k, l \to \infty$. Analogously one proves that $\mathbb{E}[\langle e_l - e_n, e_n\rangle] \to 0$ as $k, l \to \infty$, establishing (13).

Finally, one argues with (13), (12), inequality $\mathbb{E}[\|e_j - e_k\|^2]^{\frac{1}{2}} \leq \mathbb{E}[\|e_j - e_l\|^2]^{\frac{1}{2}} + \mathbb{E}[\|e_l - e_k\|^2]^{\frac{1}{2}}$ and identities

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

to conclude that $\mathbb{E}[\|e_j - e_k\|^2] \to 0$, as $k, l \to \infty$, i.e. (e_k) is a Cauchy sequence in $L_2(\Omega; X)$.

Since (x_k) is Cauchy in $L_2(\Omega; X)$, it has an accumulation point $x^* \in X$. On the other hand, it follows from Corollary 3.4 that the mean square residuals $\mathbb{E}[\|F_I^0(x_k)\|^2]$ converge to zero as $k \to \infty$. Consequently, $\mathbb{E}[\|F_I^0(x^*)\|^2] = 0$, i.e. x^* satisfies $\|F_i(x^*) - y_i\|^2 = 0$ for $i = 0, \ldots, N-1$; in other words, x^* is a solution of (8). □

3.2 The noisy data case

In this section we investigate regularization properties of the nonlinear pSGD method in (6) coupled with the (*a priori*) stopping criterion in (A6).

In the next result, the nonlinear residual norm $\|F_i^\delta(x)\|$ is compared with the norm of the linearization $\|F_i^\delta(x) + F_i'(x)(x^* - x)\|$ for $x \in B_\rho(x_0)$ and $x^* \in B_\rho(x_0)$ a solution of (8).

Lemma 3.6. *Let Assumptions (A1), (A2) and (A3) hold. For all x, $\bar{x} \in B_\rho(x_0)$ we have*

$$\| - F_i^\delta(x) - F_i'(x)(\bar{x} - x)\| \leq \eta \|F_i^\delta(x)\| + (1+\eta)\|F_i^\delta(\bar{x})\|, \ i = 0, \ldots, N-1.$$

In particular, if $\bar{x} = x^* \in B_\rho(x_0)$ is a solution of (2) it holds

$$\| - F_i^\delta(x) - F_i'(x)(x^* - x)\| \leq \eta \|F_i^\delta(x)\| + (1+\eta)\delta_i, \ \forall x \in B_\rho(x_0).$$

Proof. Given x, $\bar{x} \in B_\rho(x_0)$ we conclude from (A2) that

$$\begin{aligned}\| - F_i^\delta(x) - F_i'(x)(\bar{x} - x)\| &= \|y_i^\delta - F_i(x) - F_i'(x)(\bar{x} - x)\| \\ &= \|y_i^\delta - F_i(\bar{x}) + F_i(\bar{x}) - F_i(x) - F_i'(x)(\bar{x} - x)\| \\ &\leq \eta \|F_i(\bar{x}) - F_i(x)\| + \|y_i^\delta - F_i(\bar{x})\| \\ &\leq \eta \|y_i^\delta - F_i(x)\| + (1+\eta)\|y_i^\delta - F_i(\bar{x})\|,\end{aligned}$$

proving the first assertion. The second assertion is a direct consequence of the first one. □

In the sequel we estimate the difference $\mathbb{E}\big[\|x^* - x_{k+1}^\delta\|^2\big] - \mathbb{E}\big[\|x^* - x_k^\delta\|^2\big]$, extending the estimate (9) in Proposition 3.2 to the noisy data case.

Proposition 3.7. *Let Assumptions (A1), (A2) and (A3) hold, (x_k^δ) be a sequence generated by the nonlinear pSGD method (6), and $x^* \in B_\rho(x_0)$ be a solution of (8). If $x_k^\delta \in B_\rho(x_0)$ for some $0 \leq k \leq k_\delta^*$, then*

$$\begin{aligned}\mathbb{E}\big[\|x^* - x_{k+1}^\delta\|^2\big] - \mathbb{E}\big[\|x^* - x_k^\delta\|^2\big] &\leq \\ \theta_k(\theta_k - 2)\Big[(1-\eta)\mathbb{E}\big[\lambda_I^\delta \|F_I^\delta(x_k^\delta)\|^2\big] &- (1+\eta)\mathbb{E}\big[\delta_I \lambda_I^\delta \|F_I^\delta(x_k^\delta)\|\big]\Big]. \quad (15)\end{aligned}$$

Proof. If $\|F'_{I_k}(x_k^\delta)^* F_{I_k}^\delta(x_k^\delta)\| > \gamma\delta_{I_k}$, it follows from (6a) and (6b)

$$\begin{aligned}
\|\, x^* - x_{k+1}^\delta\|^2 - \|x^* - x_k^\delta\|^2 \;&=\; 2\langle x_{k+1}^\delta - x_k^\delta, x_{k+1}^\delta - x^*\rangle - \|x_{k+1}^\delta - x_k^\delta\|^2 \\
&= 2\theta_k\lambda_{I_k}^\delta \Big\langle - F_{I_k}^\delta(x_k^\delta), -F_{I_k}^\delta(x_k^\delta) - F'_{I_k}(x_k^\delta)(x^* - x_k^\delta) + F'_{I_k}(x_k^\delta)(x_{k+1}^\delta - x_k^\delta) + F_{I_k}^\delta(x_k^\delta)\Big\rangle \\
&\qquad - \|x_{k+1}^\delta - x_k^\delta\|^2 \\
&= 2\theta_k\lambda_{I_k}^\delta \Big[\big\langle - F_{I_k}^\delta(x_k^\delta), -F_{I_k}^\delta(x_k^\delta) - F'_{I_k}(x_k^\delta)(x^* - x_k^\delta)\big\rangle + \big\langle - F'_{I_k}(x_k^\delta)^* F_{I_k}^\delta(x_k^\delta), x_{k+1}^\delta - x_k^\delta\big\rangle \\
&\qquad - \|F_{I_k}^\delta(x_k^\delta)\|^2\Big] - \|x_{k+1}^\delta - x_k^\delta\|^2 \\
&= 2\theta_k\lambda_{I_k}^\delta \Big[\big\langle - F_{I_k}^\delta(x_k^\delta), -F_{I_k}^\delta(x_k^\delta) - F'_{I_k}(x_k^\delta)(x^* - x_k^\delta)\big\rangle + \theta_k p^{\delta_{I_k}}(\|F_{I_k}^\delta(x_k^\delta)\|) - \|F_{I_k}^\delta(x_k^\delta)\|^2\Big] \\
&\qquad -\theta_k^2\lambda_{I_k}^\delta p^{\delta_{I_k}}(\|F_{I_k}^\delta(x_k^\delta)\|)\,.
\end{aligned}$$

Thus, arguing with the Cauchy-Schwartz inequality, Lemma 3.6 and the definition of $p^\delta(\cdot)$ follows

$$\begin{aligned}
&\|\, x^* - x_{k+1}^\delta\|^2 - \|x^* - x_k^\delta\|^2 \\
&\le 2\theta_k\lambda_{I_k}^\delta \Big[\|F_{I_k}^\delta(x_k^\delta)\|\,\| - F_{I_k}^\delta(x_k^\delta) - F'_{I_k}(x_k^\delta)(x^* - x_k^\delta)\| + \tfrac{1}{2}\theta_k p^{\delta_{I_k}}(\|F_{I_k}^\delta(x_k^\delta)\|) - \|F_{I_k}^\delta(x_{I_k}^\delta)\|^2\Big] \\
&\le 2\theta_k\lambda_{I_k}^\delta \Big[\eta\|F_{I_k}^\delta(x_k^\delta)\|^2 + (1+\eta)\delta_{I_k}\|F_{I_k}^\delta(x_k^\delta)\| + \tfrac{1}{2}\theta_k p^{\delta_{I_k}}(\|F_{I_k}^\delta(x_k^\delta)\|) - \|F_{I_k}^\delta(x_{I_k}^\delta)\|^2\Big] \\
&= 2\theta_k\lambda_{I_k}^\delta \Big[(\eta-1)\|F_{I_k}^\delta(x_k^\delta)\|^2 + (1+\eta)\delta_{I_k}\|F_{I_k}^\delta(x_k^\delta)\| + \tfrac{1}{2}\theta_k\big((1-\eta)\|F_{I_k}^\delta(x_k^\delta)\|^2 \\
&\qquad -(1+\eta)\delta_{I_k}\|F_{I_k}^\delta(x_k^\delta)\|\big)\Big] \\
&= 2\theta_k\lambda_{I_k}^\delta \Big[(1-\eta)(\tfrac{\theta}{2}-1)\|F_{I_k}^\delta(x_k^\delta)\|^2 + (1+\eta)(1-\tfrac{\theta}{2})\delta_{I_k}\|F_{I_k}^\delta(x_k^\delta)\|\Big] \\
&= \theta_k(\theta_k-2)\lambda_{I_k}^\delta \Big[(1-\eta)\|F_{I_k}^\delta(x_k^\delta)\|^2 - (1+\eta)\delta_{I_k}\|F_{I_k}^\delta(x_k^\delta)\|\Big]\,. \qquad (16)
\end{aligned}$$

Otherwise, if $\|F'_{I_k}(x_k^\delta)^* F_{I_k}^\delta(x_k^\delta)\| \le \gamma\delta_{I_k}$, then $\lambda_{I_k}^\delta = 0$ and $x_{k+1}^\delta = x_k^\delta$ (see (6a) and (6b)). Therefore, (16) holds also in this case. By employing a similar reasoning as in the final part of the proof of Proposition 3.2, we establish the validity of (15). ☐

It is worth noticing that the right hand side of (16) can be rewriten in the form

$$\begin{aligned}
\theta_k(\theta_k-2)\lambda_{I_k}^\delta \Big[(1-\eta)\|F_{I_k}^\delta(x_k^\delta)\|^2 - (1+\eta)\delta_{I_k}\|F_{I_k}^\delta(x_k^\delta)\|\Big] \;&=\; \theta_k(\theta_k-2)\lambda_{I_k}^\delta p^{\delta_{I_k}}(\|F_{I_k}^\delta(x_k^\delta)\|) \\
&=\; \theta_k(\theta_k-2)(\lambda_{I_k}^\delta)^2\|F'_{I_k}(x_k^\delta)^* F_{I_k}^\delta(x_k^\delta)\|^2 \ge 0
\end{aligned}$$

(see Assumption (A4)), from where we conclude that $\|x^* - x_{k+1}^\delta\| \le \|x^* - x_k^\delta\|$. This inequality and (A3) allow us to conclude that (x_k^δ) satisfies $\|x^\star - x_{k+1}^\delta\| \le \|x^\star - x_k^\delta\|$ and $x_k^\delta \in B_\rho(x_0)$, for $k = 0, \dots, k_\delta^*$. Summarizing, we have

Corollary 3.8. *Under the assumptions of Proposition 3.7 it holds $\|x^* - x_{k+1}^\delta\| \le \|x^* - x_k^\delta\|$, for $k = 0, \dots, k_\delta^*$. Consequently, $(x_k^\delta) \subset B_\rho(x_0)$. Additionally, for any x^* solution of (8) it holds*

$$\mathbb{E}\big[\|x^* - x_{k+1}^\delta\|\big] \;\le\; \mathbb{E}\big[\|x^* - x_k^\delta\|\big]\,,\ k = 0, \dots, k_\delta^*.$$

We are now ready to state and prove a stability result (Theorem 3.9) and a semi-convergence result (Theorem 3.11) for the pSGD method in (6).

Theorem 3.9 (Stability)**.** *Let Assumptions (A1),(A2) and (A4) hold, $(\delta^j) = (\delta_0^j, \dots, \delta_{N-1}^j) \in (\mathbb{R}^+)^N$ be a sequence with $\|\delta^j\| \to 0$ as $j \to \infty$, and $(y^{\delta^j}) = (y_0^{\delta^j}, \dots, y_{N-1}^{\delta^j}) \in Y^N$ be a corresponding sequence of noisy data satisfying (1). Moreover, let $(x_l)_{l\in\mathbb{N}}$ and $(x_l^{\delta^j})_{l=0}^{k_\delta^*}$ be sequences* *generated by the nonlinear pSGD method in the case of exact and noisy data respectively; all sequences are generated using the same $(I_0, \dots, I_k, \dots)$. For each $k \in \mathbb{N}$ it holds*

$$\lim_{j\to\infty} \mathbb{E}\big[\|x_k^{\delta^j} - x_k\|^2\big] \;=\; 0\,. \tag{17}$$

Prior to establishing this theorem, we examine an auxiliary result:

Lemma 3.10. *Under assumptions of Theorem 3.9, if $\|F'_{I_k}(x_k)^* F^0_{I_k}(x_k)\| > 0$ and $\lim_j x_k^{\delta^j} = x_k$ for some $k \in \mathbb{N}$, then $\lim_j |\lambda^{\delta^j}_{I_k} - \lambda_{I_k}| = 0$.*

Proof. From (A1), (1), $\lim_j \|\delta^j\| = 0$ and $\lim_j x_k^{\delta^j} = x_k$ follow

$$\lim_{j\to\infty} p^{\delta_{I_k}}(\|F^{\delta^j}_{I_k}(x_k^{\delta^j})\|) \;=\; p^0(\|F^0_{I_k}(x_k)\|)\,. \tag{18}$$

On the other hand, from (A1), (1), $\lim_j \|\delta^j\| = 0$ and $\lim_j x_k^{\delta^j} = x_k$ we conclude that

$$\lim_j \|F'_{I_k}(x_k^{\delta^j})^* F^{\delta^j}_{I_k}(x_k^{\delta^j})\| \;=\; \|F'_{I_k}(x_k)^* F^0_{I_k}(x_k)\|.$$

Therefore, the hypothesis $\|F'_{I_k}(x_k)^* F^0_{I_k}(x_k)\| > 0$ (together with the fact $\lim_j \delta^j_{I_k} = 0$) allow us to conclude that $\|F'_{I_k}(x_k^{\delta^j})^* F^{\delta^j}_{I_k}(x_k^{\delta^j})\| > \frac{1}{2}\|F'_{I_k}(x_k)^* F^0_{I_k}(x_k)\| > \gamma\delta^j_{I_k}$ for sufficiently large j. Consequently, the lemma follows from (18) together with the definitions of $\lambda^{\delta^j}_{I_k}$ and λ_{I_k} in (6b) and (7b) respectively. □

Proof. (of Theorem 3.9)
We give an inductive proof in k. Since $x_0 = x_0^{\delta^j}$ for all $j \in \mathbb{N}$, (17) trivially holds for $k = 0$. Assume that $\lim_j \mathbb{E}\big[\|x_l^{\delta^j} - x_l\|^2\big] = 0$ for $l = 0, \dots k$. We aim to prove $\lim_j \mathbb{E}\big[\|x_{k+1}^{\delta^j} - x_{k+1}\|^2\big] = 0$. The argumentation is divided in three steps:

Step 1: We claim that $\lim_j \|x_l^{\delta^j} - x_l\|^2 = 0$ for $l = 1, \dots, k$.
Arguing as in the proof of [25, Theorem 4.3] one proves that, for each realization $(I_0, \dots, I_{l-1})$ the inequality $\|x_l^{\delta_j} - x_l\|^2 \le (\frac{1}{N})^l E\big[\|x_l^{\delta_j} - x_l\|^2\big]$ holds for $l = 1, \dots, k$ and $j \in \mathbb{N}$. Thus, for each fixed $l \in \{1, \dots, k\}$ the inductive hypothesis guarantees that $\lim_j \|x_l^{\delta^j} - x_l\|^2 = 0$.

Step 2: We claim that $\lim_j \|x_{k+1}^{\delta^j} - x_{k+1}\|^2 = 0$. Two distinct cases are considered:
(I) $\|F'_{I_k}(x_k)^* F^0_{I_k}(x_k)\| > 0$. From the iteration formulas (6a) and (7a) follow

$$\begin{aligned}
x_{k+1}^{\delta^j} - x_{k+1} &= \\
&= x_k^{\delta^j} - x_k - \theta_k\big[\lambda^{\delta^j}_{I_k} F'_{I_k}(x_k^{\delta^j})^* F^{\delta^j}_{I_k}(x_k^{\delta^j}) - \lambda_{I_k} F'_{I_k}(x_k)^* F^0_{I_k}(x_k)\big] \\
&= x_k^{\delta^j} - x_k - \theta_k\Big[(\lambda^{\delta^j}_{I_k} - \lambda_{I_k}) F'_{I_k}(x_k)^* F^0_{I_k}(x_k) + \lambda^{\delta^j}_{I_k}\big(F'_{I_k}(x_k^{\delta^j})^* F^{\delta^j}_{I_k}(x_k^{\delta^j}) - F'_{I_k}(x_k)^* F^0_{I_k}(x_k)\big)\Big].
\end{aligned}$$

Therefore, arguing with (A1) we estimate

$$\begin{aligned}
&\|x_{k+1}^{\delta^j} - x_{k+1}\| \;\le \\
&\qquad \|x_k^{\delta^j} - x_k\| + 2C|\lambda^{\delta^j}_{I_k} - \lambda_{I_k}|\,\|F^0_{I_k}(x_k)\| + 2|\lambda^{\delta^j}_{I_k}|\,\|F'_{I_k}(x_k^{\delta^j})^* F^{\delta^j}_{I_k}(x_k^{\delta^j}) - F'_{I_k}(x_k)^* F^0_{I_k}(x_k)\|\,.
\end{aligned}$$

Taking the limit $j \to \infty$ in the above inequality and arguing with Step 1 (for $l = k$), Lemma 3.10 and (A1), we conclude that the three terms on the right hand side do converge to zero;[3] proving our claim in Step 2, in case (I).

[3]Notice that Lemma 3.10 guarantees the boundedness of the sequence $(\lambda^{\delta^j}_{I_k})_j$.

(II) $\|F'_{I_k}(x_k)^*F^0_{I_k}(x_k)\| = 0$. In this case $\lambda_{I_k} = 0$ and $F^0_{I_k}(x_k) = 0$. From (6a) and (7a) follow

$$\|x^{\delta^j}_{k+1} - x_{k+1}\| \;\leq\; \|x^{\delta^j}_k - x_k\| + 2\lambda^{\delta^j}_{I_k}\|F'_{I_k}(x^{\delta^j}_k)^*F^{\delta^j}_{I_k}(x^{\delta^j}_k)\|\,. \tag{19}$$

Given $j \in \mathbb{N}$, if $\|F'_{I_k}(x^{\delta^j}_k)^*F^{\delta^j}_{I_k}(x^{\delta^j}_k)\| \leq \gamma\delta^j_{I_k}$ then $\lambda^{\delta^j}_{I_k} = 0$ and we derive from (19)

$$\|x^{\delta^j}_{k+1} - x_{k+1}\| \;\leq\; \|x^{\delta^j}_k - x_k\|. \tag{†}$$

Otherwise, if $\|F'_{I_k}(x^{\delta^j}_k)^*F^{\delta^j}_{I_k}(x^{\delta^j}_k)\| > \gamma\delta^j_{I_k}$ we argue with Lemma 3.6 to estimate

$$\begin{aligned}
\|F^{\delta^j}_{I_k}(x^{\delta^j}_k)\|^2 \;&=\; \big\langle F_{I_k}(x^{\delta^j}_k) - y^{\delta^j}_{I_k}, -y^{\delta^j}_{I_k} + F_{I_k}(x^{\delta^j}_k) - F'_{I_k}(x^{\delta^j}_k)(x_k - x^{\delta^j}_k) + F'_{I_k}(x^{\delta^j}_k)(x_k - x^{\delta^j}_k)\big\rangle \\
&\leq \|F^{\delta^j}_{I_k}(x^{\delta^j}_k)\|\,\| - F^{\delta^j}_{I_k}(x^{\delta^j}_k) - F'_{I_k}(x^{\delta^j}_k)(x_k - x^{\delta^j}_k)\| + \|F'_{I_k}(x^{\delta^j}_k)^*F^{\delta^j}_{I_k}(x^{\delta^j}_k)\|\,\|x_k - x^{\delta^j}_k\| \\
&\leq \|F^{\delta^j}_{I_k}(x^{\delta^j}_k)\|\Big[\eta\|F^{\delta^j}_{I_k}(x^{\delta^j}_k)\| + (1+\eta)\|F^{\delta^j}_{I_k}(x_k)\|\Big] + \|F'_{I_k}(x^{\delta^j}_k)^*F^{\delta^j}_{I_k}(x^{\delta^j}_k)\|\,\|x_k - x^{\delta^j}_k\| \\
&= \|F^{\delta^j}_{I_k}(x^{\delta^j}_k)\|\Big[\eta\|F^{\delta^j}_{I_k}(x^{\delta^j}_k)\| + (1+\eta)\|y^{\delta^j}_{I_k} \pm y_{I_k} - F_{I_k}(x_k)\|\Big] + \|F'_{I_k}(x^{\delta^j}_k)^*F^{\delta^j}_{I_k}(x^{\delta^j}_k)\|\,\|x_k - x^{\delta^j}_k\| \\
&\leq \|F^{\delta^j}_{I_k}(x^{\delta^j}_k)\|\Big[\eta\|F^{\delta^j}_{I_k}(x^{\delta^j}_k)\| + (1+\eta)(\delta^j_{I_k} + \|F^0_{I_k}(x_k)\|)\Big] + \|F'_{I_k}(x^{\delta^j}_k)^*F^{\delta^j}_{I_k}(x^{\delta^j}_k)\|\,\|x_k - x^{\delta^j}_k\|\,.
\end{aligned}$$

Since $F^0_{I_k}(x_k) = 0$ holds in case (II), the last inequality and the definition of $p^\delta(\cdot)$ allow us to conclude that $p^{\delta^j_{I_k}}(\|F^{\delta^j}_{I_k}(x^{\delta^j}_k)\|) \leq \|F'_{I_k}(x^{\delta^j}_k)^*F^{\delta^j}_{I_k}(x^{\delta^j}_k)\|\,\|x_k - x^{\delta^j}_k\|$. Thus, it follows from (6b)

$$\lambda^{\delta^j}_{I_k}\|F'_{I_k}(x^{\delta^j}_k)^*F^{\delta^j}_{I_k}(x^{\delta^j}_k)\| \;=\; p^{\delta^j_{I_k}}(\|F^{\delta^j}_{I_k}(x^{\delta^j}_k)\|)\,\|F'_{I_k}(x^{\delta^j}_k)^*F^{\delta^j}_{I_k}(x^{\delta^j}_k)\|^{-1} \;\leq\; \|x^{\delta^j}_k - x_k\|\,.$$

From this inequality and (19) follow

$$\|x^{\delta^j}_{k+1} - x_{k+1}\| \;\leq\; 3\|x^{\delta^j}_k - x_k\|\,. \tag{‡}$$

From inequalities (†) and (‡) together with Step 1 (for $l = k$) follow $\lim_j \|x^{\delta^j}_{k+1} - x_{k+1}\| \leq 3\lim_j \|x^{\delta^j}_k - x_k\| = 0$, proving our claim in Step 2, in case (II).

Step 3: We claim that $\lim_j \mathbb{E}\big[\|x^{\delta^j}_{k+1} - x_{k+1}\|^2\big] = 0$, concluding the inductive proof. Indeed, notice that

$$\mathbb{E}\big[\|x^\delta_{k+1} - x_{k+1}\|^2\big] \;=\big(\tfrac{1}{N}\big)^{k+1}\sum_{\substack{i_0=0\\ \cdots\\ i_k=0}}^{N}\|x^\delta_{k;i_0,\ldots,i_k} - x_{k;i_0,\ldots,i_k}\|^2,$$

where $x^\delta_{k;i_0,\ldots,i_k}$ is defined by (6a) taking $(I_0,\ldots,I_{k-1}) = (i_0,\ldots,i_{k-1})$, and $x_{k;i_0,\ldots,i_{k-1}}$ is defined by (7a) analogously. Taking the average in Step 2 over all possible realizations $(I_0,\ldots,I_k)$, one concludes that $\lim_j \mathbb{E}\big[\|x^{\delta^j}_{k+1} - x_{k+1}\|^2\big] = 0$. □

Theorem 3.11 (Semi-convergence). *Let Assumptions (A1), ..., (A6) hold true; $(\delta^j) = (\delta^j_0,\ldots,\delta^j_{N-1}) \in \mathbb{R}^N$ be a sequence with $\lim_j \|\delta_j\| = 0$ and $(\mathbf{y}^{\delta^j}) = (y^{\delta^j}_0,\ldots,y^{\delta^j}_{N-1}) \in Y^N$ be a corresponding sequence of noisy data satisfying (1). Additionally, for each $j \in \mathbb{N}$, let $(x^{\delta^j}_k)_{k=0}^{k^*(\delta^j)}$ be the corresponding sequence generated by the nonlinear pSGD method (6). There exists $x^* \in B_\rho(x_0)$, solution of (8), such that*

$$\lim_{j\to\infty}\mathbb{E}\big[\|x^{\delta^j}_{k^*(\delta^j)} - x^*\|^2\big] \;=\; 0. \tag{20}$$

Proof. Let $\hat{x} \in B_\rho(x_0)$ be a solution of (8) (the existence of $\hat{x}$ follows from (A3)). We claim that, for every fixed $k \in \mathbb{N}$ it holds

$$\|\hat{x} - x^{\delta^j}_{k+1}\|^2 - \|\hat{x} - x^{\delta^j}_k\|^2 \;\leq\; \lambda^2_{\min}\gamma^2(\delta^j_{\min})^2, \tag{21}$$

for sufficiently large j. Here $\delta^j_{\min} := \min_{i\in\{0,\ldots,N-1\}} \delta^j_i > 0$.

From (A6) we conclude that $k^*_{\delta_j} \ge k$ for sufficiently large j. Consequently, $x_k^{\delta j}$ and $x_{k+1}^{\delta j}$ are defined for sufficiently large j.

If $\|F'_{I_k}(x_k^{\delta_j})^* F^\delta_{I_k}(x_k^{\delta j})\| \le \gamma \delta^j_{I_k}$, then $\lambda^{\delta j}_{I_k} = 0$, $x_{k+1}^{\delta j} = x_k^{\delta j}$ and (21) holds trivially. Otherwise, it follows from Remark 2.3, (16) and (6b)

$$\begin{aligned}
\|\hat{x} - x_k^{\delta j}\|^2 - \|\hat{x} - x_{k+1}^{\delta j}\|^2 \;&\ge\; \theta_k(2-\theta_k)\,\lambda_{I_k} p^{\delta j}(\|F^\delta_{I_k}(x_k^{\delta j})\|) \\
&=\; \theta_k(2-\theta_k)\,(\lambda_{I_k})^2 \|F'_{I_k}(x_k^{\delta j})^* F^\delta_{I_k}(x_k^{\delta j})\|^2 \\
&>\; \theta_k(2-\theta_k)\,\lambda^2_{\min}(\gamma\,\delta^j_{I_k})^2 \\
&\ge\; \theta_k(2-\theta_k)\,\lambda^2_{\min}\gamma^2\,(\delta^j_{\min})^2
\end{aligned}$$

(notice that Assumption (A4) guarantees $\theta_k(2-\theta_k) > 0$). Consequently,

$$\|\hat{x} - x_{k+1}^{\delta j}\|^2 - \|\hat{x} - x_k^{\delta j}\|^2 \;<\; |\theta_k(2-\theta_k)|\,\lambda^2_{\min}\gamma^2\,(\delta^j_{\min})^2 \;<\; \lambda^2_{\min}\gamma^2(\delta^j_{\min})^2$$

(Assumption (A3) implies $|\theta_k(2-\theta_k)| < 1$), establishing our claim (21).

Now, taking the average in (21) over all possible realizations $(I_0, \dots, I_k)$ and arguing with Remark 3.1 we obtain

$$\mathbb{E}\big[\|\hat{x} - x_{k+1}^{\delta j}\|^2\big] - \mathbb{E}\big[\|\hat{x} - x_k^{\delta j}\|^2\big] \;\le\; \lambda^2_{\min}\gamma^2(\delta^j_{\min})^2.$$

From (A6) we may assume that $k^*_{\delta j}$ increases strict monotonically with j. Given $m < n$, we add the last inequality, with $j = n$, from $k = k^*_{\delta m}$ to $k^*_{\delta n} - 1$, and adopt the simplified notation $k^*_j = k^*_{\delta j}$, to obtain

$$\begin{aligned}
\mathbb{E}\big[\|\hat{x} - x^{\delta n}_{k^*_n}\|^2\big] \;&\le\; \mathbb{E}\big[\|\hat{x} - x^{\delta n}_{k^*_m}\|^2\big] + \sum_{k=k^*_m}^{k^*_n - 1} \lambda^2_{\min}\gamma^2(\delta^n_{\min})^2 \\
&\le\; 2\mathbb{E}\big[\|\hat{x} - x_{k^*_m}\|^2\big] + 2\mathbb{E}\big[\|x_{k^*_m} - x^{\delta n}_{k^*_m}\|^2\big] + \lambda^2_{\min}\gamma^2(\delta^n_{\min})^2 \sum_{k=k^*_m}^{k^*_n - 1} 1 \\
&\le\; 2\mathbb{E}\big[\|\hat{x} - x_{k^*_m}\|^2\big] + 2\mathbb{E}\big[\|x_{k^*_m} - x^{\delta n}_{k^*_m}\|^2\big] + \lambda^2_{\min}\gamma^2\,\|\delta^n\|^2\,k^*_n. \qquad (22)
\end{aligned}$$

Here (x_k) is the sequence generated by pSGD (7) using exact data and the same $(I_0, \dots, I_k, \dots)$ as the sequences $(x_k^{\delta j})$.

In (22) take $\hat{x} = x^*$, the solution of (8) satisfying $\lim_k \mathbb{E}[\|x^* - x_k\|^2] = 0$ (the existence of x^* is ensured by Theorem 3.5). Therefore, there is a large enough m, s.t. the first term on the rhs of (22) is smaller then $\varepsilon/3$. Additionally,

— from Theorem 3.9 with $k = k^*_m$ we conclude that the second term on the rhs of (22) becomes smaller than $\varepsilon/3$ for large enough n;

— due to (A6) the last term on the rhs of (22) becomes smaller than $\varepsilon/3$ for large enough n.

This concludes the proof. □

4 Numerical experiments

In this section the pSGD method in (6) is implemented for solving two large scale systems of nonlinear operator equations with real data. The corresponding inverse problems relate to parameter identification in neural network training. Computations are performed using MATLAB® R2017a, running in a Intel® Core™ i9-10900 CPU (10 cores, 20 threads).

In Section 4.1 we revisit, using neural networks, an inverse problem discussed in [25], namely the big data problem of CO-concentration prediction in a gas sensor array.[4] In Section 4.2 we consider a well-known benchmark problem problem in machine learning, namely the classification problem for the MNIST database (see https://en.wikipedia.org/wiki/MNIST_database).

[4] Differently from the approach used here, in [25, Section 5.2] this inverse problem was modeled as a multiple linear regression problem.

4.1 Prediction of CO-concentration in a gas sensor array

For the application considered in this section, we use a data set collected in a gas delivery platform facility at the ChemoSignals Laboratory in the BioCircuits Institute at University of California, San Diego. The data set contains the readings of 16 distinct chemical sensors, which were exposed to the mixture of Ethylene and CO at varying concentrations in air. For this gas mixture, the measurement was constructed by the continuous acquisition of the 16-sensor array signals for a duration of approximately 12 hours without interruption (we refer the reader to [9, 25] for a detailed description of the experiment). The real data used in this section is available at the UC Irvine Machine Learning Repository https://archive.ics.uci.edu/ml/index.php, dataset *Gas sensor array under dynamic gas mixtures*.

In [25] the following experimental setting was considered: readings from the last sensor (sensor #16) are used as the response variable, and readings from sensors $\{\#1, \#3, \#4, \ldots, \#15\}$ are used as covariates (readings from sensor #2 are disregarded due to strong lack of accuracy); each sensor data consists of $4,188,262$ scalar measurements.[5] The inverse problem considered in [25] is a *multiple linear regression* problem. It consists of finding an approximate solution to the linear system $Ax = y^\delta$ (with unknown noise level $\delta > 0$), where $A = (A_i)_{i=0}^{N-1} \in \mathbb{R}^{N\times M}$, with $N = 4\,188\,262$ and $M = 15$. Here $x \in \mathbb{R}^M$ is the unknown vector of regression coefficients, $y^\delta \in \mathbb{R}^N$ contains the readings from sensor #16, and A_i (the i^{th}-row of A) is such that: the first 14 columns of A_i contain the i^{th}-readings from sensors $\{\#1, \#3, \#4, \ldots, \#15\}$ and the last column is equal to one.

An inverse problem in machine learning

In this section we consider the problem of predicting the reading from sensor #16 based on the readings from the previous 14 sensors. However, differently from the multiple linear regression approach in [25], we use here a neural network (NN) that inputs the readings from the first sensors and outputs a scalar value, which predicts the reading of the last sensor.

The structure of proposed NN reads:

— **Input:** $z \in \mathbb{R}^{14}$, readings of the first 14 sensors;

— **Output:** $NN(z; W, b) = \sigma(Wz + b) \in \mathbb{R}$, where $W \in \mathbb{R}^{1\times 14}$ is a matrix of weights, $b \in \mathbb{R}$ is a scalar bias and $\sigma : \mathbb{R} \to \mathbb{R}$ is an activation function.

Notice that this is a very simple NN with only one layer (the output layer); the dimention of the corresponding parameter space is 15, the size of (W, b). If the activation function σ is linear, this approach becomes equivalent to the multiple linear regression used in [25].[6]

The inverse problem under consideration is a NN training problem, i.e. one aims to find an approximate solution (a pair of parameters (W, b)) to the nonlinear system

$$F_i(W, b) \;=\; y_i^\delta\,,\ i = 0, \ldots, N_t - 1\,, \tag{23}$$

where $F_i(W, b) := NN(z_i; W, b) = \sigma(Wz_i + b)$. Here $N_t < N$ is the size of the training set and $z_i = ((A_i)_j)_{j=1}^{14}$, where A_i is the i^{th}-row of A.

Once the parameters (W, b) are chosen, the performance $\mathcal{P}$ of the corresponding neural network $NN(\cdot; W, b)$ is defined by

$$\mathcal{P}(NN(\cdot; W, b)) \;:=\; 1 - \frac{1}{N_T} \sum_{i=N_t}^{N_t+N_T-1} \frac{\|NN(z_i; W, b) - y_i^\delta\|}{\|y_i^\delta\|}\,. \tag{24}$$

The sum in the above definition gives the average (relative) misfit betwen the predicted value $NN(z_i; W, b)$ and y_i^δ, evaluated over the test set $\{z_i,\ N_t \le i < N_t + N_T - 1\}$. Clearly it holds $0 \le \mathcal{P}(NN(\cdot; W, b)) \le 1$ for all (W, b), and $\mathcal{P}(NN(\cdot; W, b)) = 1$ is the best possible performance.

[5] See [25, Figure 3] for scatter plots of sensor #i readings against sensor #16, for $i = 1, \ldots, 15$.

[6] We choose the nonlinear activation function σ s.t. it's range contains all possible readings of sensor #16.

Remark 4.1 (Choosing the training set and test set). *At the beginning of the experiment, the rows A_i are arranged in a random order. Consequently, the 'training set' and 'test set' are comprised of random samples with sizes of N_t and N_T respectively. In our numerical experiments we use $N_t = 4,000,000$ and $N_T = 100,000$ (notice that $Nt + N_T < N$ is satisfied).*

The choice of the activation function

The activation function σ used in the definition of our NN is a variation of the *saturated linear activation function* [6]. Here $\sigma : \mathbb{R} \to \mathbb{R}$ is defined by

$$\sigma(t) = \begin{cases} c + a(t-c) & , \ t \geq c \\ t & , \ -c < t < c \\ -c + a(t+c) & , \ t \leq -c \end{cases}, \tag{25}$$

where $0 < a < 1$ and $c > 0$. The choice of this particular activation function is motivated by the next lemma (a proof is postponed to Appendix A).

Lemma 4.2. *The real function σ in (25) satisfies wTCC (4) in $\mathbb{R}$ for $\eta = (1-a)a^{-1}$, i.e.*

$$\|\sigma(\bar{t}) - \sigma(t) - \sigma'(t)(\bar{t} - t)\| \ \leq \ \eta \, \|\sigma(\bar{t}) - \sigma(t)\| \, , \ \forall \ t, \, \bar{t} \in \mathbb{R}.$$

A direct consequence of Lemma 4.2 is the fact that, choosing $a \in (\frac{1}{2}, 1]$, the function σ in (25) satisfies Assumption (A2) with $\eta \in [0, 1)$. In the sequel we prove that, for every fixed z_i, the above defined neural network $NN(z_i; \cdot, \cdot)$ does satisfy wTCC (4).

Lemma 4.3. *The function $F_i : (W, b) \mapsto NN(z_i; W, b) = \sigma(Wz_i + b)$, with σ as in (25), satisfies the wTCC (4) for $\eta = (1-a)a^{-1} \max\{\|z_i\|, 1\}$.*

Proof. If f and g are functions (with $D(f) \supset Rg(g)$) satisfying wTCC (4) for constants η_f and η_g respectively, then

$$\begin{aligned} &\|f(g(\bar{t})) - f(g(t)) - f'(g(t))g'(t)(\bar{t} - t)\| \ \leq \\ &\quad \leq \ \|f(g(\bar{t})) - f(g(t)) - f'(g(t))[g(\bar{t}) - g(t)] + f'(g(t))[g(\bar{t}) - g(t) - g'(t)(\bar{t} - t)]\| \\ &\qquad\qquad \leq \ \eta_f \, \|g(\bar{t}) - g(t)\| + \eta_g \, \|f'(g(t))\| \, \|\bar{t} - t\|. \end{aligned} \tag{26}$$

Notice that $F_i = f \circ g_i$ with $f(t) = \sigma(t)$ and $g_i(W, b) = Wz_i + b$. Since g_i is linear, it satisfies wTCC (4) for $\eta_g = 0$. Consequently, it follows from (26) and Lemma 4.2

$$\begin{aligned} \|F_i(\bar{W}, \bar{b}) - F_i(W, b) - F_i'(W, b)[(\bar{W}, \bar{b}) - (W, b)]\| \ &\leq \ (1-a)a^{-1}\|(\bar{W}z_i - \bar{b}) - (Wz_i - b)\| \\ = \ (1-a)a^{-1}\|(\bar{W} - W)z_i - (\bar{b} - b)\| \ &\leq \ (1-a)a^{-1}\big[\|z_i\| \, \|\bar{W} - W\| + \|\bar{b} - b\|\big], \end{aligned}$$

for $(\bar{W}, \bar{b})$, $(W, b) \in \mathbb{R}^{14} \times \mathbb{R}$. The assertion follows from the last inequality. □

Numerical implementations

In what follows the pSGD method in (6) is implemented for solving the NN training problem (23). In view of Lemma 4.2 we choose distinct values for a in (25), namely $\{0.9091, 0.6666, 0.5253\}$ and $c = 8$. Consequently, we generate three distinct activation functions σ of the form (25), satisfying wTCC (4) for $\eta = 0.1$, $\eta = 0.5$ and $\eta = 0.9$ respectively.

The sensor readings $(z_i, y_i^\delta) \in \mathbb{R}^{14} \times \mathbb{R}$ on the training set are scaled by the factor $\max_{i \leq N_t} \|z_i\|$. An analogous procedure is performed on the test set. Consequently, after scaling, it holds

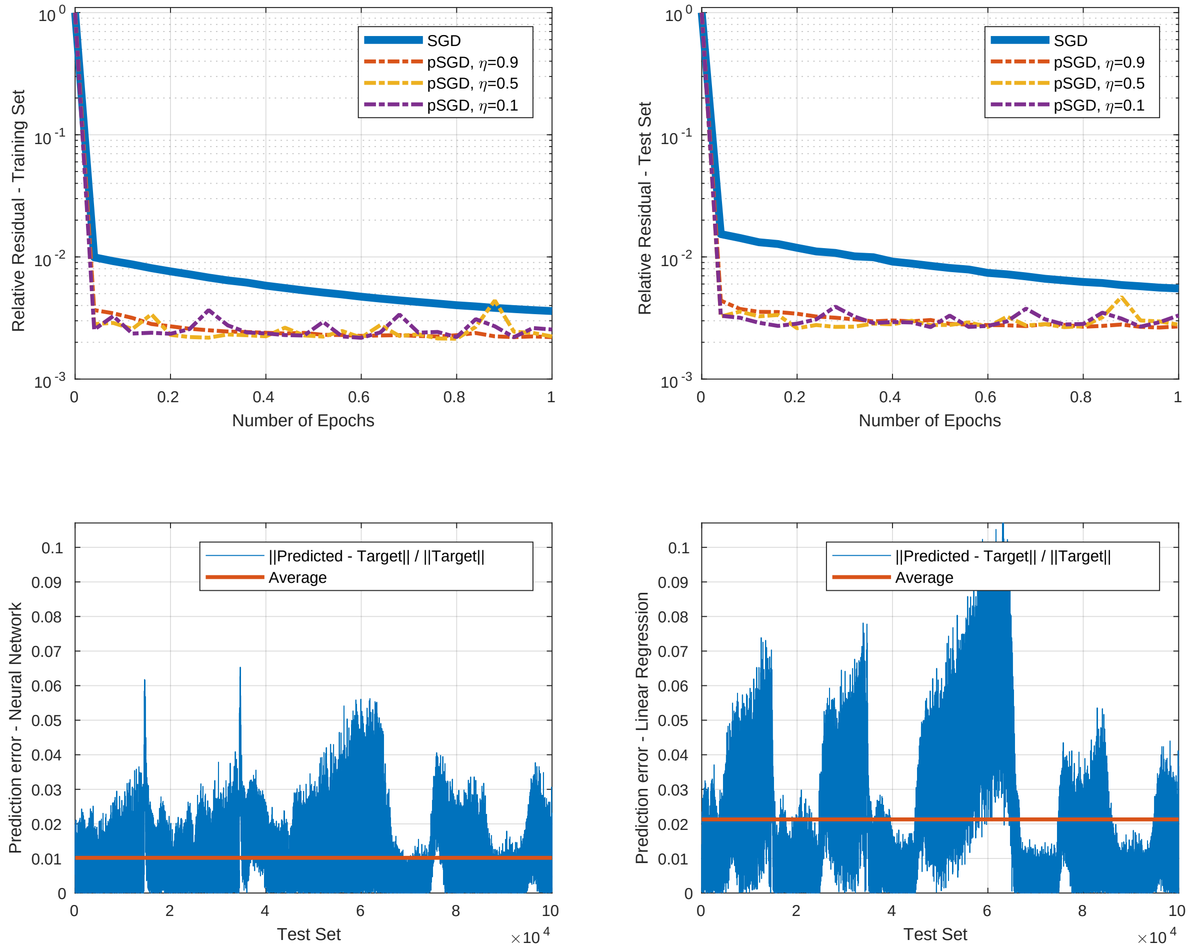

Figure 1: Prediction of CO-concentration in a gas sensor array. (TOP) Evolution of the relative residual: Training set (Left) and Test set (Right); (BOTTOM) Accuracy of the prediction: Neural network (Left) and Linear regression (Right).

$\|z_i\| \le 1$, for $i = 0, \ldots, N_t + N_T$. From Lemma 4.3 it follows that, for each one of activation functions σ above, all operators $F_i(W, b)$ satisfy Assumption (A2) for the same constant η, namely $\{0.1, 0.5, 0.9\}$.

In our experiments the initial guess (W_0, b_0) is a random vector with coordinate values ranging in $(-1, 1)$. Moreover, we choose the sequence $\theta_k \equiv 1$ in (A4). The iteration (W_k, b_k) is computed for one epoch, i.e. for $k = 1, \ldots, N_t$. Three different runs of the pSGD method are presented (one for each choice of activation function σ); in each run the wTCC is satisfied for a different constant η, namely $\{0.1, 0.5, 0.9\}$. Since the noise level δ is not known, we set $p^\delta(t) = (1-\eta)t^2$ in (6b). The computed numerical results are summarized in Figure 1:

(TOP-LEFT) Evolution of relative residual on the training set: $\sum_{i=0}^{N_t-1} \frac{\|NN(z_i;W_k,b_k)-y_i^\delta\|}{\|NN(z_i;W_0,b_0)-y_i^\delta\|}$;

(TOP-RIGHT) Relative residual evolution on the test set: $\sum_{i=N_t}^{N_t+N_T-1} \frac{\|NN(z_i;W_k,b_k)-y_i^\delta\|}{\|NN(z_i;W_0,b_0)-y_i^\delta\|}$;

For comparison purposes, the SGD method was also implemented (this method corresponds to the choice $\theta_k = 1$ and $\lambda_k = 1$ in (6a)).

At regular intervals of $\frac{1}{100}N_t$ steps, $\mathcal{P}(NN(\cdot; W_k, b_k))$ is computed. The index $0 \le k^* \le N_t$ is chosen such that (W_{k^*}, b_{k^*}) exhibits the highest performance among the evaluated ones. The additional task of evaluating $\mathcal{P}(NN(\cdot; W_k, b_k))$ a hundred times (per epoch) leads to a 10% increase in the overall computation time of the pSGD method.[7]

[7]Since $N_T << N_t$, the computational cost associated with calculating $\mathcal{P}(NN(\cdot; W_k, b_k))$ is significantly lower

The prediction accuracy of $NN(\cdot; W_{k^*}, b_{k^*})$ is investigated in Figure 1 (BOTTOM-LEFT): the relative prediction error $\|NN(z_i; W_{k^*}, b_{k^*}) - y_i^\delta\| / \|y_i^\delta\|$ is plotted for the test set $\{z_i,\ N_t \le i < N_t + N_T - 1\}$ (BLUE), the average value (RED) is 0.010. **The performance $\mathcal{P}(NN(\cdot; W_{k^*}, b_{k^*}))$ amounts to 99%**.

For comparison purpose we plot in Figure 1 (BOTTOM-RIGHT) the prediction accuracy of the linear regression approach [25] for the same test set (BLUE), the average value is 0.021 (RED). **The performance of this approach amounts to 97.7%**.

4.2 Classification problem for the MNIST database

The *Modified National Institute of Standards and Technology* (MNIST) database consists of images of handwritten digits (Figure 2). Each image is accompanied by a corresponding label indicating the digit it represents. This dataset is commonly used in the field of machine learning for developing neural network architechtures, and for testing training algorithms for neural networks.

The MNIST database contains 60,000 training images (along with 10,000 testing images) of the ten digits. Each image consists of a 28×28 pixel array of grayscale levels.

Figure 2: Sample images from the MNIST database (source Wikipedia).

The corresponding data-files are accessible from many different sources. For the experiments conducted here, the files were downloaded from http://yann.lecun.com/exdb/mnist/.

An inverse problem in machine learning

In this section we consider the well-known classification problem for the MNIST database. In order to model this problem, we use here a NN that inputs a 28×28 pixel array of grayscale levels (i.e. a vector in $\mathbb{R}^{784}$ with coordinates ranging from zero (black) to 255 (white)) and outputs a vector in $\mathbb{R}^{10}$. The classification of the handwritten digit depicted in the 28×28 image is given by the coordinate of this output vector with maximal absolute value (for alternative NN architechtures with outputs in $\mathbb{R}^{10}$ we refer the reader to [5] for a Deep NN, or [4] for a Convolutional NN).

The architechture of the NN used in our experiments is as follows:

— **Input:** $z \in \mathbb{R}^{784}$, pixel array from the MNIST database;

— **Hidden layer:** $\tilde{z} := \sigma_1(W_1 z + b_1) \in \mathbb{R}^{64}$, where $W_1 \in \mathbb{R}^{64,784}$ and $b_1 \in \mathbb{R}^{64}$;

— **Output:** $NN(z; W_1, b_1, W_2, b_2) := \sigma_2(W_2 \tilde{z} + b_2) \in \mathbb{R}^{10}$, where $W_2 \in \mathbb{R}^{10,64}$ and $b_2 \in \mathbb{R}^{10}$.

Here W_1, W_2 are weight matrices and b_1, b_2 are biases vectors. Moreover, $\sigma_1 : \mathbb{R}^{64} \to \mathbb{R}^{64}$ and $\sigma_2 : \mathbb{R}^{10} \to \mathbb{R}^{10}$ are nonlinear activation functions.

The classification of the input image z is given by the scalar value $j \in \{0,1,2,3,4,5,6,7,8,9\}$ defined by $j := \arg\max_{0 \le i \le 9} \big|NN_i(z; W_1, b_1, W_2, b_2)\big|$.[8]

This simple NN has only 2–layers 784–64–10 (one hidden layer and the output layer); the dimension of the corresponding parameter space is $50,890 = 64(784+1) + 10(64+1)$, i.e. the dimension of the set of parameters (W_1, b_1, W_2, b_2).

Typically, much larger NN's are used for solving the MNIST classification problem (e.g., the Deep NN in [5] has 6–layers 784-2500-2000-1500-1000-500-10 and achieves an accuracy rate of

compared to the cost of calculating the average relative residual on the training set.

[8] We adopt here the notation $NN(\cdot) = [NN_i(\cdot)]_{i=0}^{9} \in \mathbb{R}^{10}$.

99.65%). Our goal with this experiment is not to investigate state-of-the-art NN-architechtures. Instead, we aim to test the efficiency of the pSGD method in (6) as a training algorithm. For this purpose, the above described NN-architechture is rich enough to define a challenging inverse problem as we shall see next.

The inverse problem under consideration is a NN training problem, i.e. one aims to find an approximate solution (W_1, b_1, W_2, b_2) to the nonlinear system

$$F_i(W_1, b_1, W_2, b_2) \;=\; y_i\,,\;\; i = 0, \dots, N_t - 1\,. \tag{27}$$

Here $N_t = 60,000$ is the size of the training set and $F_i(W_1, b_1, W_2, b_2) := NN(z_i; W_1, b_1, W_2, b_2) = \sigma_2\big(W_2\sigma_1(W_1 z_i + b_1) + b_2\big)$, where z_i is the i^{th}-image of the MNIST database for $i = 0, \dots, N_t - 1$. The right hand side in (27) is a vector of the type $y_i = (0, \dots, 0, 1, 0, \dots, 0) \in \mathbb{R}^{10}$, where the index of the coordinate with value "1" indicates the digit depicted in the image z_i (e.g., if z_i is an image of the digit 2, then $y_i = (0, 1, 0, \dots, 0)$). Note that the data in (27) is exact, i.e. the noise level is $\delta = 0$.

The activation functions σ_1, $\sigma_2 : \mathbb{R} \to \mathbb{R}$ used in the above NN are variations of the *sigmoid function* [6], namely: $\sigma_1(t) = \frac{1}{2}\tanh(t/10)$ and $\sigma_2(t) = 2\tanh(t/10)$.

Numerical implementations

The pSGD method in (6) is implemented for solving the NN training problem (27). It is worth mentioning that the operators F_i in (27) do not satisfy the wTCC (4).

In our numerical experiments, the initial guess $(W_1^0, b_1^0, W_2^0, b_2^0)$ consists of random matrices/vectors with coordinate values ranging in $(-1, 1)$. Since the noise level is $\delta = 0$, we set $p^\delta(t) = t^2$ in (6b).

Three different runs of the pSGD method are presented in Figure 3. In each one of them the iteration $(W_1^k, b_1^k, W_2^k, b_2^k)$ is computed for 20 epochs, i.e. for $k = 1, \dots, 20N_t$. In the first 2 runs we choose the sequence $\theta_k \equiv 1$ in (A4), while in the last run a random sequence $\theta_k \in (0, 2)$ is choosen. The numerical results plotted in Figure 3 show:

(TOP) Evolution of relative residual on the training set: $\sum_{i=0}^{N_t-1} \frac{\|NN(z_i;W_1^k,b_1^k,W_2^k,b_2^k)-y_i\|}{\|NN(z_i;W_1^0,b_1^0,W_2^0,b_2^0)-y_i\|}$;

(BOTTOM) Evolution of relative residual on the test set: $\sum_{i=N_t}^{N_t+N_T-1} \frac{\|NN(z_i;W_1^k,b_1^k,W_2^k,b_2^k)-y_i\|}{\|NN(z_i;W_1^0,b_1^0,W_2^0,b_2^0)-y_i\|}$.

Here $N_T = 10,000$ is the number of images in the MNIST database test set. For comparison purposes, the SGD method was also implemented for solving (27) and the evolution of the corresponding residuals is also plotted.

Following every $\frac{1}{10}N_t$ steps, the average relative residual is computed on the test set; the index $0 \le k^* \le 20N_t$ is chosen such that $(W_1^{k^*}, b_1^{k^*}, W_2^{k^*}, b_2^{k^*})$ exhibits the smallest relative residual. After selecting the set of parameters $(W_1^{k^*}, b_1^{k^*}, W_2^{k^*}, b_2^{k^*})$, the accuracy rate of the corresponding neural network $NN(\cdot; W_1^{k^*}, b_1^{k^*}, W_2^{k^*}, b_2^{k^*})$ is calculated using the test set. In the experiment above, $k^* = 19.4N_t$ is obtained from the first run (run 1 in Figure 3) of the pSGD method. The accuracy rate of $NN(\cdot; W_1^{k^*}, b_1^{k^*}, W_2^{k^*}, b_2^{k^*})$ is 95.96%.

Some remarks regarding the numerical experiments:

• **Computation of $\mathbf{k}^*$:** Since $N_T << N_t$, the computational cost associated with calculating the (average) residual on the test set is insignificant compared to the cost of computing the (average) residual on the training set. The computation required in order to determine k^* is performed 10 times per epoch; this additional task does not impose a significant numerical burden on the implementation of the pSGD method.

• **Choice of $(\theta_\mathbf{k})$:** The experiments presented above suggest that the choice of the relaxation parameters (θ_k) does not significantly impact the decay rate of the residual. Different runs of

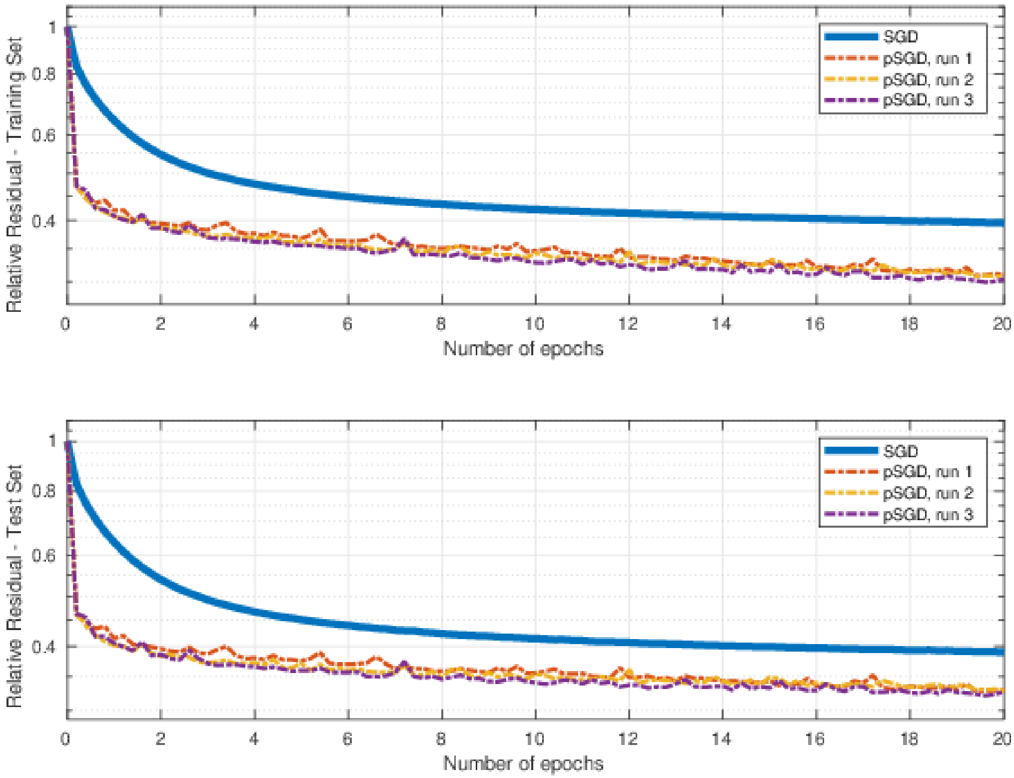

Figure 3: MNIST classification problem. (TOP) Evolution of the relative residual for training set; (BOTTOM) Evolution of the relative residual for test set.

the pSGD method using sequences $\theta_k \in (0,1)$ (under relaxation), or $\theta_k \in (1,2)$ (over relaxation), or random $\theta_k \in (0,2)$ produce similar numerical results.

• **Accuracy rate:** The accuracy rate of $NN(\cdot; W_1^{k^*}, b_1^{k^*}, W_2^{k^*}, b_2^{k^*})$ is given by the trace of the confusion matrix divided by N_T. The confusion matrix is a table that is used to evaluate the performance of a classification model. It provides a summary of how well the model has classified the different classes in a dataset. It is typically used for problems like the MNIST classification, where the output of the model can belong to multiple classes. It displays the actual class labels of the data against the predicted class labels generated by the model. The main diagonal in Figure 4 represents the correctly classified instances, while the off-diagonal elements represent misclassifications. The final entry in a row/column represents the cumulative sum of all preceding elements in that particular row/column.

5 Conclusions

In this manuscript we investigate a nonlinear *projective stochastic-gradient* (pSGD) method for computing stable approximate solutions to large scale systems of nonlinear ill-posed equations.

We build upon a well-established nonlinear assumption, namely the weak tangential cone condition (wTCC), see (A2), to expand the method in [25, 24]. As a result, we create a new approach capable of efficiently solving large-scale systems of nonlinear equations.

Our method stands out due to the stepsize selection, which is inspired by the *projective Landweber (PLW) method* [21] and the *projective Landweber-Kaczmarz (PLWK) method* [20].

Highlighted among the key findings established in this manuscript are: (i) Estimates for the *average gain* and monotonicity results for the *average iteration error*; (ii) A convergence result

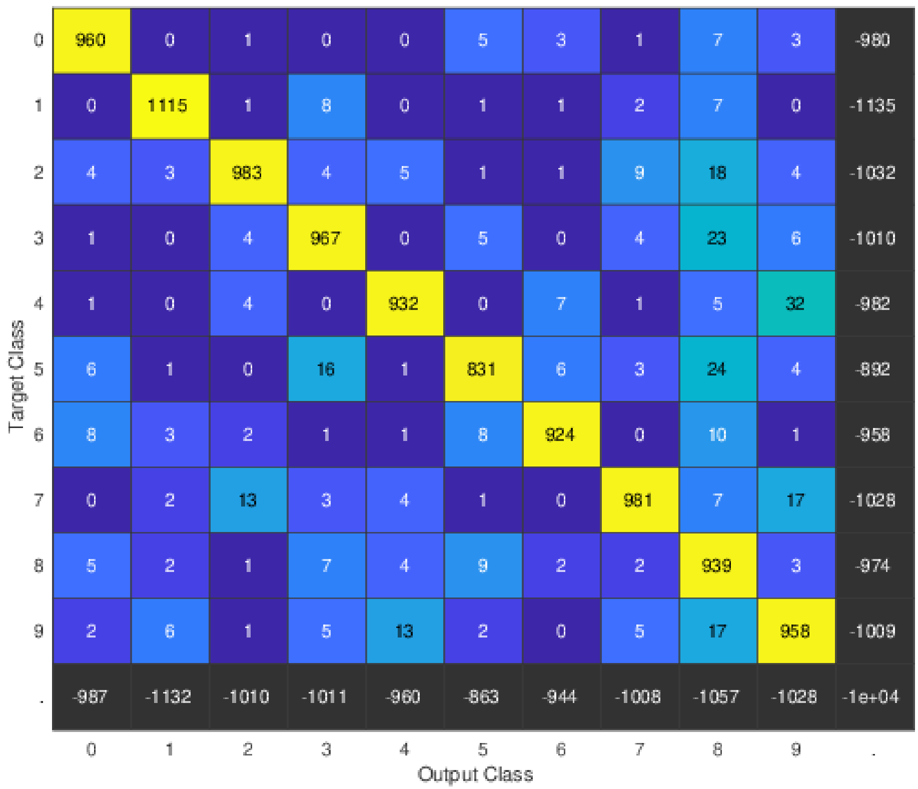

Figure 4: MNIST classification problem. Confusion matrix for $NN(\cdot; W_1^{k^*}, b_1^{k^*}, W_2^{k^*}, b_2^{k^*})$.

for the pSGD method in the exact data case (Theorem 3.5); (iii) Regularization properties of the pSGD method: a stability result (Theorem 3.9) and a semi-convergence result (Theorem 3.11); (iv) In Lemma 4.3 we prove that the neural-network used to model the inverse problem in Section 4.1 satisfies the wTCC.

Numerical experiments are presented for two large scale nonlinear problems in machine learning: (i) the big data problem of CO-concentration prediction in a gas sensor array considered in [9, 25]; (ii) the classification problem for the MNIST database. The obtained numerical results demonstrate the efficiency of the pSGD method.

A Appendix: Proof of Lemma 4.2

In what follows we prove that the real activation function σ defined in (25) (see Figure A) satisfies wTCC (4) in $\mathbb{R}$ for the constant $\eta = (1-a)a^{-1}$.

The first step is to verify that the piecewise-linear function $h : \mathbb{R} \to \mathbb{R}$ defined by

$$h(t) := \begin{cases} \alpha t\,, & t \geq 0 \\ \beta t\,, & t < 0 \end{cases} \tag{28}$$

with $\alpha \geq \beta > 0$ does satisfy wTCC in $\mathbb{R}$ for the constant $\widetilde{\eta} = (\alpha - \beta)\beta^{-1}$.

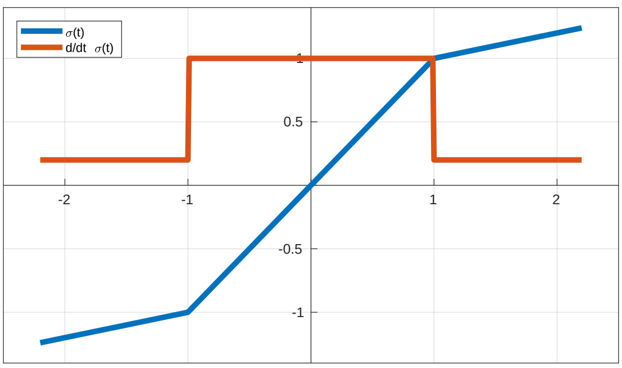

Figure A: Activation function $\sigma(t)$ in (25) and it's derivative $\frac{d}{dt}\sigma(t)$ for $c = 1$ and $a = 1/4$.

Since $h|_{[0,\infty)}$ and $h|_{[-\infty,0)}$ are linear, then h satisfies wTCC for $\widetilde{\eta} = 0$ whenever t, $\bar{t} \in [0, \infty)$ or t, $\bar{t} \in (-\infty, 0]$. Now, assume that $\bar{t} \geq 0$ and $t < 0$. It holds

$$\begin{aligned}\|h(\bar{t}) - h(t) - h'(t)(\bar{t} - t)\| &= |\alpha\bar{t} - \beta t - \beta(\bar{t} - t)| = |(\alpha - \beta)\bar{t}| = \frac{\alpha - \beta}{\alpha}\alpha\bar{t} \\ &\leq \frac{\alpha - \beta}{\alpha}\alpha\bar{t} - \frac{\alpha - \beta}{\alpha}\beta t = \frac{\alpha - \beta}{\alpha}|\alpha\bar{t} - \beta t| = \frac{\alpha - \beta}{\alpha}\|h(\bar{t}) - h(t)\| .\end{aligned}$$

On the other hand,

$$\begin{aligned}\|h(t) - h(\bar{t}) - h'(\bar{t})(t - \bar{t})\| &= |\beta t - \alpha\bar{t} - \alpha(t - \bar{t})| = |(\beta - \alpha)t| = (\beta - \alpha)t = \frac{\beta - \alpha}{\beta}\beta t \\ &\leq \frac{\beta - \alpha}{\beta}\beta t - \frac{\beta - \alpha}{\beta}\alpha\bar{t} = \left|\frac{\beta - \alpha}{\beta}\right||\beta t - \alpha\bar{t}| = \frac{\alpha - \beta}{\beta}\|h(t) - h(\bar{t})\| .\end{aligned}$$

Due to the hipothesis $\alpha \geq \beta$, we conclude from the last inequalities that the function h in (28) satisfies wTCC in $\mathbb{R}$ for $\widetilde{\eta} = (\alpha - \beta)\beta^{-1}$.

Proof. (of Lemma 4.2)
Due to the definition of σ in (25) we observe that σ does satisfy wTCC for $\eta = 0$ whenever $t, \bar{t} \in (-\infty, -c]$ or $t, \bar{t} \in [c, \infty)$ or $t, \bar{t} \in [-c, c]$.

Chosing $\alpha = 1$ and $\beta = a$ in (28), it follows from (25) that $\sigma(t) = h(t+c)-c$, for $t \in (-\infty, c]$. Additionally, it holds $\sigma(t) = c - h(c - t)$, for $t \in [-c, \infty)$. Consequently, σ satisfies wTCC for $\eta = (1 - a)a^{-1}$ whenever $t, \bar{t} \in (-\infty, c]$ or $t, \bar{t} \in [-c, \infty)$.

It remains to consider the case where $\bar{t} \geq c$ and $t \leq -c$. Notice that, in this case, it holds $\sigma(\bar{t}) - \sigma(t) \geq 2c$ (see (25) and Figure A). Thus, $\|\sigma(\bar{t}) - \sigma(t) - \sigma'(t)(\bar{t} - t)\| = |2c - 2ac| = (1-a)2c \leq (1-a)\,\|\sigma(\bar{t}) - \sigma(t)\|$. Analogously, one proves that $\|\sigma(t) - \sigma(\bar{t}) - \sigma'(\bar{t})(t - \bar{t})\| \leq (1-a)\,\|\sigma(t) - \sigma(\bar{t})\|$. Consequently, σ satisfies wTCC for $\eta = (1 - a)$ whenever $\bar{t} \in [c, \infty)$ and $t \in (-\infty, -c]$.

Since $0 < a < 1$ in (25), we come to the concluson that σ satisfies wTCC (4) in $\mathbb{R}$ for $\eta = (1 - a)a^{-1}$, concluding the proof. □

References

[1] A.B. Bakushinsky and M.Y. Kokurin. *Iterative Methods for Approximate Solution of Inverse Problems*, volume 577 of *Mathematics and Its Applications*. Springer, Dordrecht, 2004.

[2] J. Baumeister, B. Kaltenbacher, and A. Leitão. On levenberg-marquardt-kaczmarz iterative methods for solving systems of nonlinear ill-posed equations. *Inverse Probl. Imaging*, 4(3):335–350, 2010.

[3] M. Burger and B. Kaltenbacher. Regularizing Newton-Kaczmarz methods for nonlinear ill-posed problems. *SIAM J. Numer. Anal.*, 44:153–182, 2006.

[4] D. Cireşan, U. Meier, and J. Schmidhuber. Multi-column deep neural networks for image classification. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3642–3649, 2012.

[5] D.C. Cireşan, U. Meier, L.M. Gambardella, and J. Schmidhuber. Deep, big, simple neural nets for handwritten digit recognition. *Neural Computation*, 22(12):3207–3220, 2010.

[6] P.L. Combettes and J.-C. Pesquet. Deep neural network structures solving variational inequalities. *Set-Valued Var. Anal.*, 28(3):491–518, 2020.

[7] A. De Cezaro, M. Haltmeier, A. Leitão, and O. Scherzer. On steepest-descent-Kaczmarz methods for regularizing systems of nonlinear ill-posed equations. *Appl. Math. Comput.*, 202(2):596–607, 2008.

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

[9] J. Fonollosa, S. Sheik, R. Huerta, and S. Marco. Reservoir computing compensates slow response of chemosensor arrays exposed to fast varying gas concentrations in continuous monitoring. *Sensors and Actuators B: Chemical*, 215:618–629, 2015.

[10] C. W. Groetsch. *Stable Approximate Evaluation of Unbounded Operators*, volume 1894 of *Lecture Notes in Mathematics.* Springer-Verlag, Berlin, 2007.

[11] M. Haltmeier, R. Kowar, A. Leitão, and O. Scherzer. Kaczmarz methods for regularizing nonlinear ill-posed equations. II. Applications. *Inverse Probl. Imaging*, 1(3):507–523, 2007.

[12] M. Haltmeier, A. Leitão, and E. Resmerita. On regularization methods of EM-Kaczmarz type. *Inverse Problems*, 25:075008, 2009.

[13] M. Haltmeier, A. Leitão, and O. Scherzer. Kaczmarz methods for regularizing nonlinear ill-posed equations. I. convergence analysis. *Inverse Probl. Imaging*, 1(2):289–298, 2007.

[14] M. Hanke, A. Neubauer, and O. Scherzer. A convergence analysis of Landweber iteration for nonlinear ill-posed problems. *Numerische Mathematik*, 72:21–37, 1995.

[15] Tim Jahn and Bangti Jin. On the discrepancy principle for stochastic gradient descent. *Inverse Problems*, 36(9):095009, 2020.

[16] Bangti Jin and Xiliang Lu. On the regularizing property of stochastic gradient descent. *Inverse Problems*, 35(1):015004, 2018.

[17] Bangti Jin, Zehui Zhou, and Jun Zou. On the convergence of stochastic gradient descent for nonlinear ill-posed problems. *SIAM Journal on Optimization*, 30(2):1421–1450, 2020.

[18] B. Kaltenbacher, A. Neubauer, and O. Scherzer. *Iterative regularization methods for nonlinear ill-posed problems*, volume 6 of *Radon Series on Computational and Applied Mathematics.* Walter de Gruyter GmbH & Co. KG, Berlin, 2008.

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

[20] A. Leitão and B.F. Svaiter. On projective Landweber-Kaczmarz methods for solving systems of nonlinear ill-posed equations. *Inverse Problems*, 32(1):025004, 2016.

[21] A. Leitão and B.F. Svaiter. On a family of gradient type projection methods for nonlinear ill-posed problems. *Numerical Functional Analysis and Optimization*, 39(2-3):1153–1180, 2018.

[22] F. Margotti, A. Rieder, and A. Leitão. A Kaczmarz version of the `reginn`-Landweber iteration for ill-posed problems in Banach spaces. *SIAM J. Numer. Anal.*, 52(3):1439–1465, 2014.

[23] V.A. Morozov. *Regularization Methods for Ill–Posed Problems.* CRC Press, Boca Raton, 1993.

[24] J.C. Rabelo and A. Leitão. Addendum: On stochastic Kaczmarz type methods for solving large scale systems of ill-posed equations (2022 Inverse Problems **38** 025003). *Inverse Problems*, 38(5):Paper No. 059401, 3, 2022.

[25] J.C. Rabelo, Y. Saporito, and A. Leitão. On stochastic kaczmarz type methods for solving large scale systems of ill-posed equations. *Inverse Problems*, 38(2):025003, 2022.

[26] O. Scherzer. Convergence rates of iterated Tikhonov regularized solutions of nonlinear ill-posed problems. *Numerische Mathematik*, 66(2):259–279, 1993.

[27] T.I. Seidman and C.R. Vogel. Well posedness and convergence of some regularisation methods for non–linear ill posed problems. *Inverse Probl.*, 5:227–238, 1989.

[28] A.N. Tikhonov. Regularization of incorrectly posed problems. *Soviet Math. Dokl.*, 4:1624–1627, 1963.

[29] A.N. Tikhonov and V.Y. Arsenin. *Solutions of Ill-Posed Problems*. John Wiley & Sons, Washington, D.C., 1977. Translation editor: Fritz John.